Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods

Supplementary Material

Ali Faisal[a], Frank Dondelinger[*,b,c], Dirk Husmeier[b], Colin M. Beale[**,d]

[a]*Helsinki Institute for Information Technology, Adaptive Informatics Research Centre, Department of Information and Computer Science, Aalto University, P.O. Box 15400, FI-02015 Aalto, Finland*

[b]*Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, United Kingdom*

[c]*Institute for Adaptive Neural Computation, School of Informatics, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom*

[d]*Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen, AB15 8QH, United Kingdom*

1. Introduction

This document contains the supplementary material for the paper "Inferring Species Interaction Networks from Species Abundance Data". Please refer to the main paper for a discussion of the background and motivation for the work. We also present and discuss the main findings there. The supplementary material contains extensions to the methods that we used, as well as some

[*]Corresponding author. Tel: +44 (0)131 650 7536, Fax: +44 (0)131 650 4901

[**]Present Address: Department of Biology (Area 18), University of York, PO Box 373, York, YO10 5YW

Email addresses: ali.faisal@tkk.fi (Ali Faisal), frankd@bioss.ac.uk (Frank Dondelinger), dirk@bioss.ac.uk (Dirk Husmeier), cb751@york.ac.uk (Colin M. Beale)

additional findings that had to be omitted from the main paper due to space constraints.

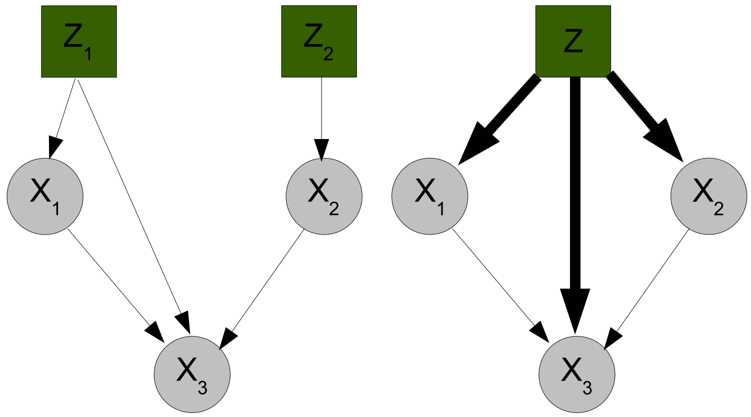
In Section 2.1, we describe an extension to the Bayesian network method (see Section 3.1.4 in the main paper) that allows including unobserved factors in the network inference. Section 2.2 describes how we extended the simulation model that was used to generate the synthetic data with an observation process that discretises the data by deciding for each location whether the presence of a species was observed or not. In Section 3 we investigate the difference between edge strengths and confidence values for edges in regression. Section 4 presents some additional experiments on the synthetic data. Section 5 gives additional information on how the networks inferred from the bird atlas data were evaluated, and presents the recovered networks and their characteristics in more detail.

2. Methods

2.1. Latent variable model allowing for unobserved factors

We want to extend the Bayesian network approach to allow for unobserved factors in the environment, e.g. related to climate change or the availability of natural resources. This can be achieved by including additional so-called latent variables in the model. Inference can be carried out with the allocation sampler described in Nobile and Fearnside (2007) and Grzegorczyk *et al.* (2008), which is based on the following iterative procedure: Given the network structure, new values for the latent variables are inferred (imputation step). Then, given the complete data (real data, and imputed values for the latent variables), the network structure is modified with a standard structure MCMC step (Madigan and York, 1995). This procedure is iterated, and leads to a Markov chain which (on convergence) samples both the network structure and the allocation of the latent variables from the posterior distribution.

Ideally, the interactions between the latent variables and the species are treated as flexible (Fig. 1 a). To reduce the computational complexity, we keep them fixed, i.e. they were enforced to be connected to all species. It is easy to prove that for discrete values, this is equivalent to a model with a single latent variable and a flexible number of discretisation levels (Fig. 1 b); this is the model described in Grzegorczyk *et al.* (2008).



(a) General Latent Variable Model

(b) Restricted Latent Variable Model

Figure 1: (a) Unrestricted latent variable model, here with two latent variables and three observed ones. (b) Alternative model with a single completely connected latent variable; this is effectively a mixture model. Zs are latent variables, Xs are observed variables. Thin edges are learnt, thick edges are fixed.

While the application of this scheme to the simulated data led to encouraging results (Section 4.3), the MCMC simulations did not properly converge for the warbler data. The reason is that a straightforward adaptation of the method proposed in Grzegorczyk *et al.* (2008) introduces a separate latent variable for each spatial location, leading to a model that is significantly more complex than explored in the original application. Our future work therefore aims to simplify the model complexity and explore alternative inference schemes based on variational learning.

2.2. Observation Process for Simulation Data

The simulation described in Section 2.2 of the main paper produces continuous values for the population densities. In order to transform these into presence/absence data similar to the Bird Atlas data, we implement an observation process. We assume that the probability $P(x_g)$ of missing (i.e. not observing) a population of density x_g is modelled by a Gaussian $N(\mu, \sigma^2)$. Then the probability of observing a species with density x_g is $P(X < x_g)$, i.e. the cumulative distribution function:

$$P(X < x_g) = \frac{1}{2}\left(1 + erf\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right) \tag{1}$$

We can then sample a discrete value for x_g from a binomial distribution, using $P(X < x_g)$ as the parameter. Mean and variance of the Gaussian distribution are fitted so that the distribution of ones and zeros over all locations and species is the same as in the real data set.

2.3. Implementations

Table 1 shows which software we used for the different network reconstruction methods described in Section 3.1 in the main paper, as well as where to get the MATLAB code for our own implementations of the extensions in Section 3.2 of the main paper.

Method	**Software**	**Package**	**Description**
GGM	R	GeneNet	The software implementing Graphical Gaussian models is described in Schäfer and Strimmer (2005) and can be found at: `http://strimmerlab.org/software/genenet/`
LASSO (Linear)	MATLAB	Genelab	For LASSO regression with continuous data, we used software from the Genlab package referenced in van Someren *et al.* (2006).
LASSO (Logistic)	C	BBR	For LASSO regression with discrete data, we used the BBR package (for Bayesian Binary Regression), which implements logistic LASSO regression. The package can be found at: `http://www.stat.rutgers.edu/~madigan/BBR/`
SBR	MATLAB	RegNets	We used the sparse Bayesian regression software referenced in Rogers and Girolami (2005) and available here: `http://www.dcs.gla.ac.uk/~srogers/reg_nets.htm`
Structure MCMC	MATLAB	None	The implementations for Structure MCMC and Structure MCMC with latent variables were developed from code by Marco Grzegorczyk and can be found at: `http://www.bioss.ac.uk/students/frankd.html`
Population Simulation	MATLAB	None	The simulation code was developed by Jonathan Yearsley and slightly modified for this project. It can be found at: `http://www.bioss.ac.uk/students/frankd.html`

Table 1: Network reconstruction software used

3. Investigation into LASSO Weights versus Confidence Values for Edges

3.1. Motivation

When using LASSO linear regression to reconstruct an interaction network, we have two options. One is to use the weights found during the regression and interpret them as edge strengths between the target variable and the other variables in the network (we will refer to this as "the weight method"). The other is to obtain confidence values for the presence of an edge ("the confidence value method"). Obtaining the weights is straightforward, and only requires one regression per variable. However, it is potentially biased towards edges that have a strong effect, and may ignore edges with a small (but consistent) effect.

To obtain confidence values, we use a method that is essentially an approximation of a full Bayesian approach to regression. Rather than obtaining the probability that an edge is zero from a posterior distribution of the weights, we follow Friedman *et al.* (2000) and approximate this value by 'sampling' data from the original dataset[1] using bootstrapping and subsampling. In bootstrap sampling, we sample data points with replacement until the sample size is the same as the size of the original dataset. In subsampling, we sample without replacement until we have obtained a dataset that is half the size of the original dataset.

For each dataset sampled in this way, we run a LASSO regression. Then we record the non-zero weights. After we have done this for a large number of samples, we average over the results. This gives the confidence value for the occurrence of each edge, independent of the strength of that edge. The drawback is that it requires many more runs of the regression algorithm than just calculating the weights once.

We wanted to find out if the difference between using confidence values and using the weights was substantial enough to warrant the extra computational cost. For that reason, we used two synthetic datasets: A simple network model without cycles (in other words, a DAG) from which we generated data using a linear regression model, and a more complex ecological simulation based on Lotka-Volterra interactions between species in a food web (see Section 2.2 of

[1] This should not be confused with sampling from a posterior distribution.

the main paper).

3.2. Simple Network Model

To simulate data from the simple network model based on linear regression, we first sample a network from the niche model described in Section 2.2 of the main paper. If the model is not a DAG, we remove edges until acyclicity has been restored. For each remaining edge, we draw an interaction strength from the Gaussian distribution $N(0, 1)$.

Then we identify species without any parents in the network and draw their population numbers from the Gaussian distribution $N(0, 1)$[2]. For each of the remaining species, we do a standard regression:

$$\hat{y_g} \;=\; \sum_r w_{gr} x_r + N(0, 0.1) \tag{2}$$

where r ranges over all species x_r that are parents of species y_g, and w_{gr} is the weight of the edge linking x_r and y_g. The $N(0, 0.1)$ factor adds a small amount of observational noise. We repeat this process, drawing new population numbers each time to generate different data points.

3.3. Results

Simple Network Model. We generated data from 10 random networks using the simple linear regression model, and for each network we generated 100 bootstrap/subset replica. Figure 2 shows the results. We started off by computing the confidence values straightforwardly: For each sampled dataset, every weight that was not set to 0 by the LASSO regression was counted as detecting an edge. The results of this basic approach are shown in the unshaded boxes in Figure 2.

Using a two-sided paired t-test, we determined that while the difference in TPFP5 values was not significant, the difference in AUC values between the two sampling methods and the weight method was significant ($p < 0.01$). It is surprising to see the weight method outperform the confidence value methods,

[2]This allows for negative population numbers, but this is not a problem since LASSO regression does not assume that population numbers have to be positive.

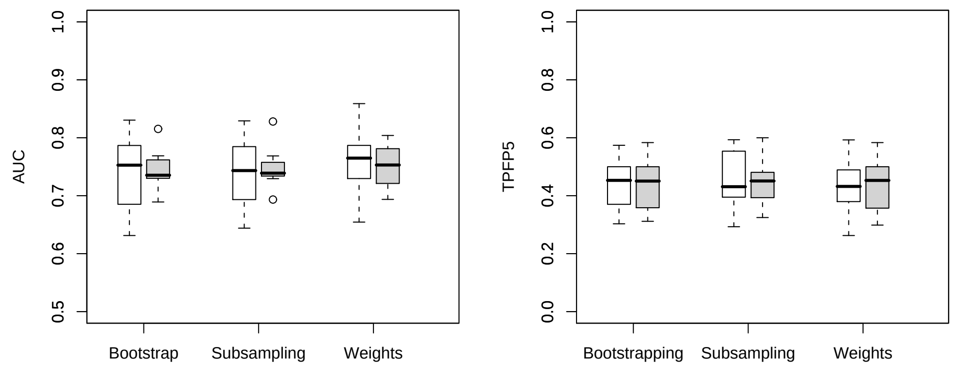


Figure 2: AUC and TPFP5 performance measures for the LASSO reconstruction of the simple network model. Shaded boxes show the result when thresholding is applied.

as we would expect confidence values to produce equally good if not better results.

The reason for this discrepancy becomes apparent once we change the procedure for estimating confidence values slightly. Instead of treating all non-zero weight values in each sampled dataset as evidence of an edge, we only keep those above a certain threshold (arbitrarily set at 0.1). To be fair in our comparison, we also apply the threshold to the weight method. When we do this, we notice that the AUC values between the confidence value methods and the weight method are no longer significantly different ($p > 0.3$).

The problem is that the selection process which sets some weights to zero is not a very conservative process. This means that some weights may never or rarely get set to zero, despite having a very low value. A threshold artificially removes those weights, and thus reduces the variance in the performance. This evens out the difference between the weight method and the confidence value methods.

Ecological Network Simulation Model. We also want to compare the different methods using the simulation model described in Section 2.2 of the main paper. We use the same datasets that were used in the rest of this study.

Since we have already established that thresholding is needed to remove the variance due to small but persistent weights in the confidence value methods, we also use this method here. Figure 3 shows the results on the ecological simulation data. A two-sided paired t-test shows that all differences in AUC

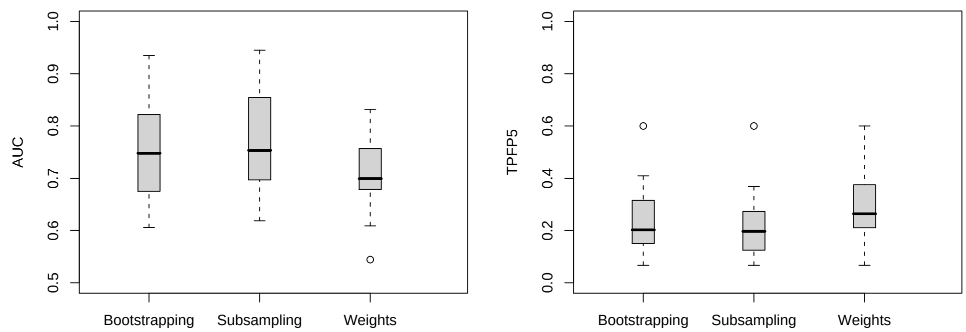


Figure 3: AUC and TPFP5 performance measures for the LASSO reconstruction of the ecological network simulation model.

values are significant ($p < 0.01$), but none of the differences in TPFP5 values are ($p > 0.08$).

Interestingly, the significant difference in AUC now shows an increased performance for the confidence value methods. However, one must remember that the model does not include any spatial autocorrelation (cf. Section 3.2.1 of the main paper), which is by necessity, as sampling destroys the spatial structure. But this also means that sampling reduces the spatial autocorrelation, because we only sample a subset of the total number of nodes, so some of the neighbours of a selected location are left out. This explains why we see a slight increase in performance in AUC. It is reasonable that it would not be mirrored in the TPFP5 score, because this score relies on edges with high edge weights, which will be found in any case.

4. Additional Results on Simulated Data

In this section, we present additional results on the simulated data that could not be included in the main paper due to space restrictions. We show the results for discretised data (Section 4.1), a study of different types of consensus networks (Section 4.2) and the results of a latent variable model for Bayesian networks (Section 4.3). We also list the significance of all results in a separate section (Section 4.4).

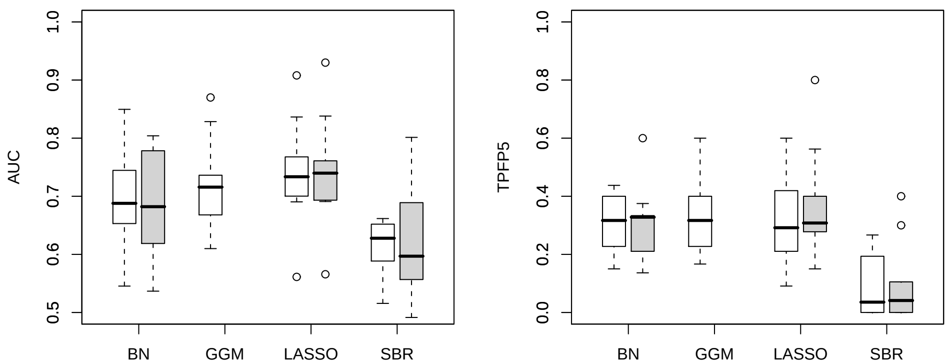


Figure 4: AUC and TPFP5 performance measures for discretised simulation data. Shaded boxes show the result when spatial autocorrelation is included in the model.

4.1. Discrete Data

To better simulate a real dataset, we discretise the continuous data from the simulation model using the observation process described in Section 2.2. The results of applying our network reconstruction methods on the discrete data can be seen in Figure 4.

As expected, the performance decreased when compared to the continuous data (see Section 4.1 in the main paper), due to the information loss inherent in the discretisation process. The AUC scores dropped around 0.1 for all methods, and the TPFP5 scores showed a similar drop, except in the case of SBR, which stayed about the same. This is because discretisation mostly hinders the identification of the more subtle interactions, which SBR had not even detected in the continuous case. Apart from SBR, there is no significant difference in the scores between methods for discrete data.

To finish our investigation, we looked at the effect of including spatial autocorrelation for the discretised data. The results are shown in Figure 4 (shaded boxes).

Unfortunately, none of the scores improved significantly when including spatial autocorrelation in the discrete case. This is likely due to the information loss in the observation process, which makes it harder to estimate spatial autocorrelation effects reliably. Our future work aims to reduce the information loss by applying more complex spatial-temporal models, e.g. along the lines of the Markov random field model proposed in Wei and Li (2007).

4.2. Consensus Networks

As described in the main paper, it is useful to combine outputs of different network reconstruction methods into one single recovered network. We call this a consensus network, because it captures the consensus between the various methods, whilst simultaneously allowing the strengths of the different methods to be combined. There are several different ways in which we can combine these methods:

- Arithmetic Mean: Edge strengths produced by regression methods are scaled to the range $[0, 1]$ (posterior probabilities obtained by Bayesian nets are left unchanged), then we take the arithmetic mean of the scaled strengths and probabilities obtained by all methods and use this as indication of the confidence we have in each edge.
- Harmonic Mean: This is the same as the previous method, but instead of using the arithmetic mean, we calculate the harmonic mean, which is generally more appropriate for rates.
- Thresholded: In this method, we use the posterior probabilities obtained by Bayesian nets as a threshold. All edges with probability less than 0.1 are removed. Then the remaining edges are evaluated based on the interaction strengths found in regression.

Note that some of these methods potentially confuse confidence values (probabilities) with interaction strengths, but for methods where both were available we found a very strong Spearman rank correlation between the two ($\rho = 0.92$), so this is not problematic. As a base line, we used the mean of the AUC or TPFP5 scores obtained from the different network reconstruction methods in isolation. A consensus method works if it produces a better score than the mean score of the individual methods.

Figure 5 shows the results using the discretised dataset with spatial autocorrelation modelled. This most closely mirrors the experiments on the bird data; however, results using continuous data and data without modelling the spatial autocorrelation were similar. As can be seen, the only method performing better than our baseline is the arithmetic mean. For AUC the difference is significant (using a two-sided paired t-test, $p = 0.03$) while the harmonic mean does not perform significantly different (though only barely, $p = 0.05$) and the thresholded approach performs significantly worse ($p = 10^{-3}$). For the TPFP5

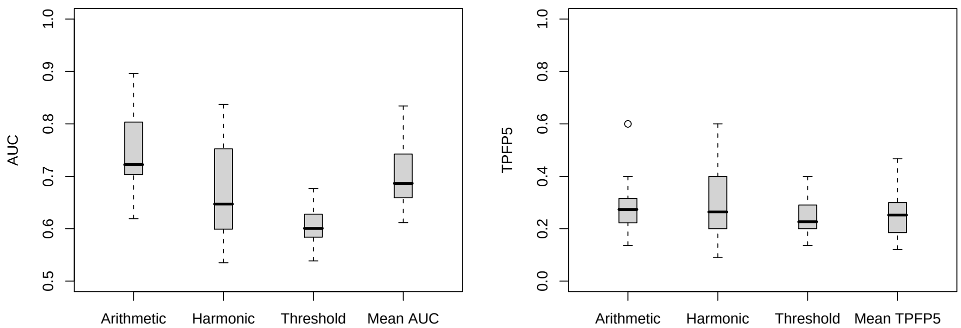


Figure 5: AUC and TPFP5 performance measures for different types of consensus networks. This figure only shows the results for discrete data with an spatial autocorrelation model. Results for other datasets were similar.

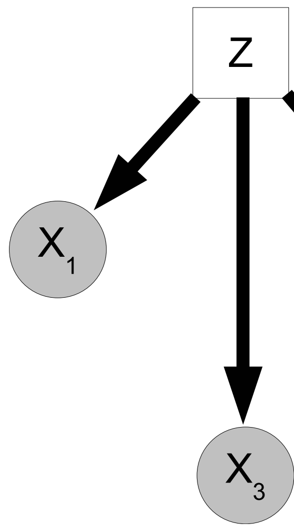
score, none of the three consensus methods performs significantly different from the baseline of taking the mean of the scores, although the arithmetic mean comes closest ($p = 0.06$ versus $p = 0.25$ and $p = 0.58$ for harmonic mean and thresholded approach, respectively).

These results show that the arithmetic mean performs best when it comes to combining different network reconstruction methods. On the basis of this investigation, we have used the arithmetic mean to construct consensus networks for the bird atlas data.

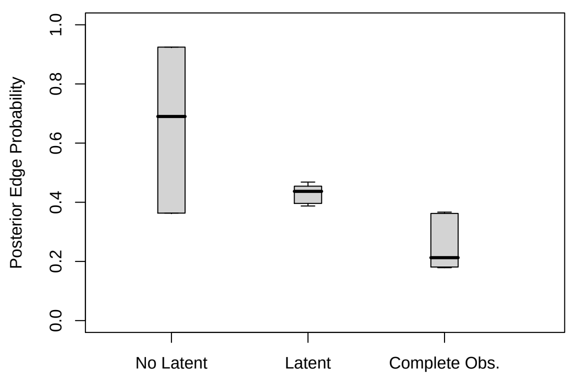
4.3. Allowing for Unobserved Effects

As explained in Section 2.1, we may want to take account of unobserved effects that act on the different species. While there are no explicit environmental factors (other than noise) in the simulation model, it is easy to model an unobserved effect by adding a species that acts directly on all other species, and removing the presence/absence data for that species reconstructing the network. To assess the helpfulness of this approach, we tested it on a small network consisting of three observed nodes and one unobserved node, with no interactions between the observed nodes (Fig. 6a). Under these circumstances, the latent variable model should produce fewer spurious interactions than a model without latent variables. In the Bayesian network model, this means that the posterior probability of edges between observed nodes should be lower when using the latent variable model.

Figure 6b shows the performance of the Latent Variable Model, compared with



(a) Test Network



(b) Results

Figure 6: (a) The network used to test the performance of the latent variable model, consisting of one fully connected species Z, and three unconnected species X_1, X_2, X_3. (b) Boxplot showing the posterior probabilities of spurious edges found using Structure MCMC with one fully-connected missing species, Structure MCMC with a latent variable, and Structure MCMC with a complete dataset (no missing species).

(a) Continuous, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.06	**0.02**	0.06
GGM		1	0.28	**0.00**
LASSO			1	**0.00**
SBR				1

(b) Discrete, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.09	0.06	**0.00**
GGM		1	0.08	**0.00**
LASSO			1	**0.00**
SBR				1

(c) Continuous, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.21	0.52
LASSO		1	**0.00**
SBR			1

(d) Discrete, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.08	0.16
LASSO		1	**0.01**
SBR			1

Table 2: Significance values obtained using a two-sided paired t-test when comparing different methods based on the AUC scores of the reconstructed networks. Significant results (with threshold $p = 0.05$) are marked in bold.

the baseline of using simple Structure MCMC with a missing species and the optimal scenario of having complete data. As can be seen, the Latent Variable Model succeeds in reducing the median probability of spurious edges, although not quite to the level of having complete knowledge of the data.

4.4. Significance Tests

This section gives an overview of the significance of the differences between the network reconstruction methods (Tables 2 and 3), as well as between methods that include spatial autocorrelation and those that do not (Table 4). We have used two-sided paired t-tests everywhere, pairing up results on data simulated from the same network. The threshold for statistical significance is set at $p = 0.05$. For an interpretation of the significant results, see Section 5 in the main paper.

(a) Continuous, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.55	**0.02**	**0.00**
GGM		1	0.38	**0.00**
LASSO			1	**0.00**
SBR				1

(b) Discrete, No Spat. Autocorr. Model

	BN	GGM	LASSO	SBR
BN	1	0.22	0.58	**0.00**
GGM		1	0.71	**0.00**
LASSO			1	**0.01**
SBR				1

(c) Continuous, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.17	**0.00**
LASSO		1	**0.00**
SBR			1

(d) Discrete, With Spat. Autocorr. Model

	BN	LASSO	SBR
BN	1	0.06	**0.01**
LASSO		1	**0.01**
SBR			1

Table 3: Significance values obtained using a two-sided paired t-test when comparing different methods based on the TPFP5 scores of the reconstructed networks. Significant results (with threshold $p = 0.05$) are marked in bold.

(a) Continuous Data

	AUC	TPFP5
BN	**0.01**	**0.00**
LASSO	**0.00**	**0.00**
SBR	**0.00**	**0.04**

(b) Discrete Data

	AUC	TPFP5
BN	0.58	0.80
LASSO	0.51	0.07
SBR	0.65	0.84

Table 4: Significance values obtained using a two-sided paired t-test when network reconstruction methods with spatial autocorrelation model to those without. Significant results (with threshold $p = 0.05$) are marked in bold.

4.5. Number of Neighbouring Locations in the Spatial Autocorrelation Model

As described in Section 3.2.1 of the main paper, our model for spatial autocorrelation calculates the average population at neighbouring locations. One open question is how many neighbouring locations to consider. If we assume that locations are distributed on a grid, then two natural choices are to either consider 4 direct neighbours, or all 8 surrounding neighbours.

We have compared the effect of calculating spatial autocorrelation using 4 direct neighbours versus using 8 neighbours for the LASSO network reconstruction method on simulated data. There was no significant difference between the two approaches ($p > 0.2$ for AUC and TPFP5 scores). Figure 7 shows scatterplots comparing the edge weights, AUC and TPFP5 scores for simulated data. We have also investigated the effect of using 8 neighbours for the Warbler dataset (Figure 8), and found that the edge weights inferred with 4 neighbours correlate very well with the edge weights inferred for 8 neighbours. These findings lead us to conclude that 4 direct neighbours are sufficient to accurately model the spatial autocorrelation.

5. Application to the European bird atlas data

5.1. A priori network construction

To construct the *a priori network*, we used two sources: knowledge from the literature, and expert judgement.

First, we searched the ecological literature using ISI Web of Knowledge [3] (accessed on 10/5/09). For each species, we searched for all articles using the complete scientific name. If more than 100 articles were returned, we refined the search adding the terms 'interaction' or 'competition'. We studied all abstracts and identified papers containing information about interspecific interactions for detailed reading. We identified 30 interactions using this method.

For the remaining 711 pairwise interactions we used our expert judgement to answer the question: In areas where these species occur in close proximity, is it plausible that one of the species would become more abundant or expand

[3] Found at `http://www.isiwebofknowledge.com/`.

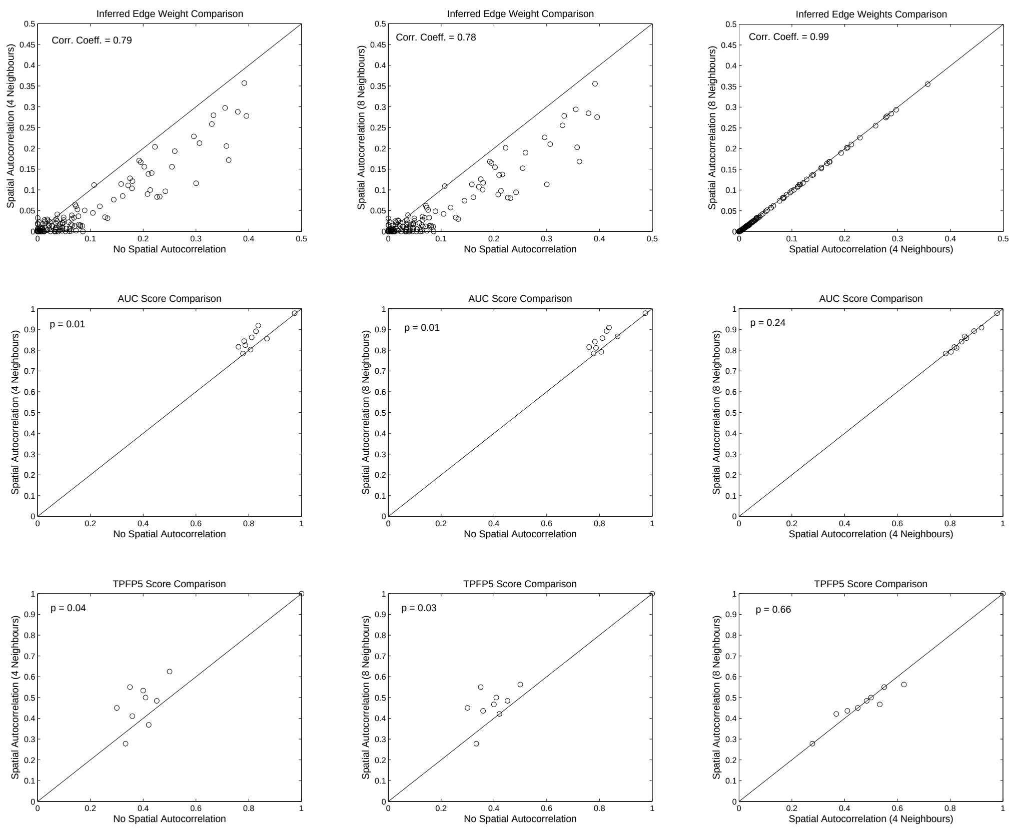


Figure 7: Comparison between no spatial autocorrelation modelled (0 neighbours), spatial autocorrelation with 4 neighbours and spatial autocorrelation with 8 neighbours, for networks reconstructed from simulated data with LASSO. From left to right, we compare 0 neighbours with 4 neighbours, 0 neighbours with 8 neighbours and 4 neighbours with 8 neighbours. The top row compares the inferred edge weights. The middle row compares the AUC network reconstruction scores for each of the 10 simulated networks. The bottom row compares the TPFP5 reconstruction scores for each of the 10 simulated networks. For edge weights, we show the Spearman rank correlation coefficient, while for network reconstruction scores, we show the p value of a two-sided paired t-test.

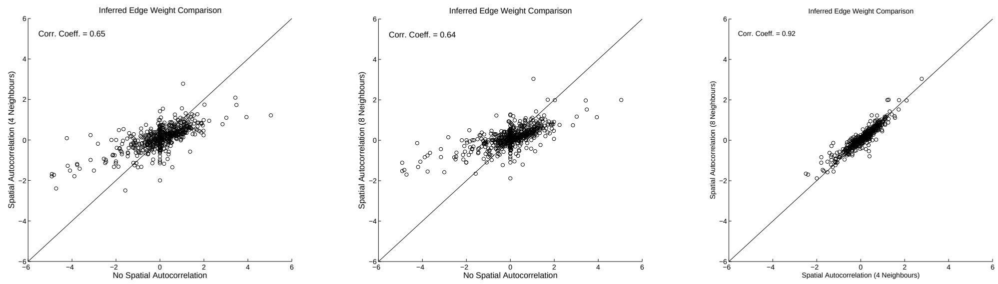


Figure 8: Comparison between no spatial autocorrelation modelled (0 neighbours), spatial autocorrelation with 4 neighbours and spatial autocorrelation with 8 neighbours, for networks reconstructed from the Warbler dataset with LASSO. From left to right, we compare 0 neighbours with 4 neighbours, 0 neighbours with 8 neighbours and 4 neighbours with 8 neighbours. In each plot, the inferred edge weights are compared.

into different habitats if the other species were absent? In cases where we considered this likely we recorded an interaction in the network.

The final network can be found at `http://www.bioss.ac.uk/students/frankd.html`.

5.2. Phylogenetic distance analysis

To calculate the phylogenetic distances between warbler species, we first needed to get general information on warbler phylogeny. To that end, we searched the taxonomic literature (e.g. Alstroem *et al.* (2006)) and 'Tree of Life' servers (such as The Tree of Life Web Project in Maddison *et al.* (2007)). A conservative consensus tree was generated depicting relationships between the 39 warbler species as in Figure 9.

As path lengths were unavailable we computed a range of distances using the method advocated by Grafen (1989) with values of ρ of 1, 0.6 and 0.3. Although correlations between the phylogenetic distance and recovered interaction scores were not qualitatively different when these different distances were assumed, they are arbitrary choices none the less. Consequently, we repeated the correlation analysis using Kendall's τ as a measure of rank correlation that is unaffected by assumed branch lengths. Again, results were qualitatively similar; they can be found in Table 5. For the correlation analyses we used only data from the upper triangle of the distance matrices.

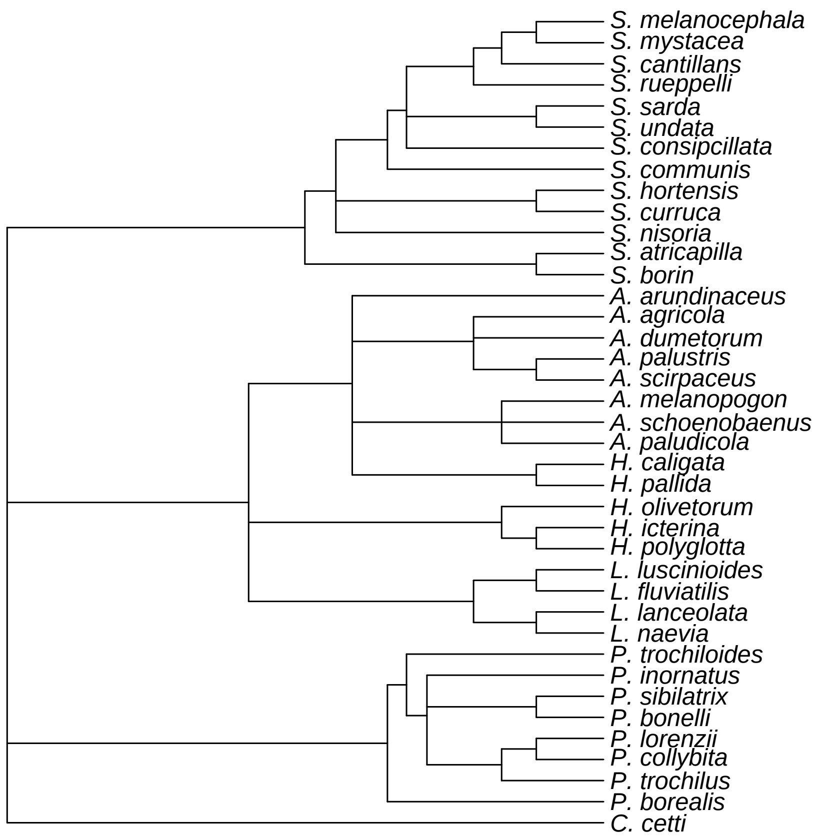


Figure 9: Phylogenetic tree for the warbler species in our study.

Network	ρ	**Correlation**
Basic	1	-0.12 (-0.18, -0.04)
	0.6	-0.11 (-0.18, -0.04)
	0.3	-0.12 (-0.19, -0.05)
	Kendall's τ	-0.08
Spat. Autocorr	1	-0.12 (-0.19, -0.05)
	0.6	-0.12 (-0.19, -0.05)
	0.3	-0.12 (-0.19, -0.05)
	Kendall's τ	-0.08
Spat. Autocorr. and Bio-Climate Covariates	1	-0.14 (-0.21, -0.07)
	0.6	-0.14 (-0.21, -0.07)
	0.3	-0.12 (-0.22, -0.07)
	Kendall's τ	-0.09

Table 5: Correlation coefficients of reconstructed networks with the phylogenetic tree whose branch lengths have been generated with different values of ρ, or with Kendall's τ. Numbers in brackets show the confidence intervals at 95%. None of the confidence intervals includes zero, indicating that the correlations are significant.

5.3. Ecological distance analysis

Ecological trait data for each of the 39 species is presented in Table 6. From the habitat and migration status data we generated indicator variables identifying species with shared habitat and shared migration strategy. We combined these indicator variables with the morphological data and clutch size, centred and scaled each variable and calculated the Euclidian distance. As with the phylogenetic distance analysis, we used only data from the upper triangle of the distance matrix in correlation analyses.

Species	**Length**	**Mass**	**Wingspan**	**Clutch**	**Migrant status**	**Preferred Habitat**
Acrocephalus agricola	13	11	16	4.5	Long Distance	Resident
Acrocephalus arundinaceus	20	33	26	4.5	Long Distance	Resident
Acrocephalus dumetorum	13	12	18	5	Long Distance	Resident
Acrocephalus melanopogon	12	12	16	4.5	Short Distance	Resident
Acrocephalus paludicola	13	12	18	5	Long Distance	Resident
Acrocephalus palustris	13	13	20	4.5	Long Distance	Shrub
Acrocephalus schoenobaenus	13	12	19	5	Long Distance	Shrub
Acrocephalus scirpaceus	13	13	19	4	Long Distance	Resident
Cettia cetti	14	13.5	17	4.5	Resident	Resident
Hippolais icterina	14	13	22	4.5	Long Distance	Broad-leaf Forest
Hippolais olivetorum	15	18	25	3.5	Long Distance	Broad-leaf Forest
Hippolais pallida	13	11	20	2.5	Long Distance	Shrub
Hippolais polyglotta	13	13	18	4	Long Distance	Broad-leaf Forest
Locustella fluviatilis	13	17	20	6	Long Distance	Shrub
Locustella luscinioides	14	18	20	5	Long Distance	Resident
Locustella naevia	13	14	17	5.55	Long Distance	Shrub
Phylloscopus collybita collybita	10	9	18	5.49	Short Distance	Broad-leaf Forest
Phylloscopus bonelli bonelli	12	9	18	5	Long Distance	Broad-leaf Forest
Phylloscopus borealis	11	10	19	5.5	Long Distance	Pine Forest
Phylloscopus trochiloides	10	8	18	5.49	Long Distance	Pine Forest
Phylloscopus sibilatrix	12	10	22	5.77	Long Distance	Broad-leaf Forest
Phylloscopus trochilus	11	10	19	5.93	Long Distance	Broad-leaf Forest
Sylvia atricapilla	13	21	22	4.56	Short Distance	Shrub
Sylvia borin	14	19	22	4.32	Long Distance	Shrub

Sylvia cantillans	12	11	17	4	Long Distance	Garrigue
Sylvia communis	14	16	20	4.64	Long Distance	Shrub
Sylvia conspicillata	12	10	15	4.5	Short Distance	Garrigue
Sylvia curruca	13	12	18	4.67	Long Distance	Shrub
Sylvia hortensis	15	21	22	5	Long Distance	Broad-leaf Forest
Sylvia melanocephala	14	13	16	4.5	Resident	Garrigue
Sylvia nisoria	16	25	25	4.5	Long Distance	Shrub
Sylvia rueppelli	14	14	20	5	Long Distance	Broad-leaf Forest
Sylvia sarda	12	10	16	4	Resident	Garrigue
Sylvia undata	12	10	16	4	Resident	Garrigue
Sylvia mystacea	14	10	17	4.5	Short Distance	Shrub
Hippolais caligata	12	10	20	3.5	Long Distance	Shrub
Phylloscopus inornatus	10	7	17	4	Long Distance	Pine Forest
Phylloscopus lorenzii	10	8	18	5.5	Short Distance	Broad-leaf Forest
Locustella lanceolata	12	12	15	5	Long Distance	Shrub

Table 6: Ecological traits for the warbler birds.

	A priori net	Phylogenetic Dist.	Ecological Dist.
A priori net	1	0.38 (0.73, 0.02)	0.08 (0.21, -0.05)
Phylogenetic Dist.		1	0.28 (0.21, 0.34)
Ecological Dist.			1

Table 7: Results of comparison between the ecological measures represented by the *a priori* interaction network, phylogenetic distance and ecological distance. *A priori* comparisons made with logistic regression are the regression coefficient, other results are Pearson's correlation coefficients, all with 95% confidence intervals

5.4. Comparison of Ecological Measures

We have three different ecological indicators that we can compare our reconstructed networks to: The a priori network, the phylogenetic distance and the ecological distance. The correlation of these indicators with the reconstructed networks that we present in the main paper is always significant, but also far from perfect correlation. This can be explained by the fact that these measures are not a true gold standard. In fact, each measure captures different aspects of the true relationships between species. In Table 7 we present the correlation coefficients between the three ecological measures and show that they are also small but (mostly) significant.

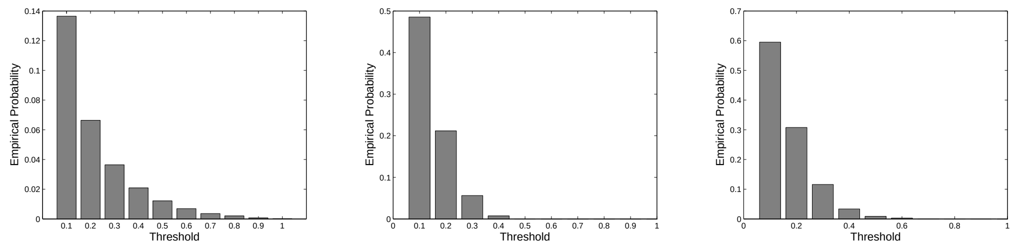
Another way to compare the ecological indicators is by taking the a priori network as a gold standard, and calculating the AUC and TPFP5 values for the phylogenetic and ecological distance measures. In effect, we are treating these distance measures as inverse edge scores. The results are shown in Table 8. Again, The scores are better than random expectation (AUC=0.5, TPFP5=0.05), but far from perfect (AUC=TPFP5=1.0). This indicates that the various measures capture relevant, but only partial aspects of the unknown true interaction network.

5.5. Thresholding on Edge Interactions

To produce a single, interpretable network from the edge interaction strengths, we need to set a threshold to discard edges with low values. Recall that the "interaction strengths" are of different nature: marginal posterior probabilities for Bayesian networks, and regularised regression coefficients for LASSO. We would like to map them to p-values, which are more commonly used in

	AUC	TPFP5
Phylogenetic Distance	0.79	0.37
Ecological Distance	0.67	0.22

Table 8: Comparison between the ecological measures by computing AUC and TPFP5 scores for phylogenetic and ecological distance measures, using the *a priori* interaction network as a gold standard.



(a) BN (b) LASSO (c) Consensus

Figure 10: Distribution of edge strengths/posterior probabilities under the null hypothesis, averaged over 15,210 random species interactions from permuted data.

statistics. To this end, we carried out a randomisation test. The rows and columns of the original warbler data were permuted ten times, and on each of these replications we carried out the same inference as for the original data. Since the permutation destroys all genuine associations among the species, the distribution of "interaction strengths" represents the null hypothesis of no species interaction. From this distribution, the p-value is easily computed as the probability of exceeding a given threshold.

Figure 10 shows the null distributions obtained for Bayesian networks (left panel), LASSO (centre panel), and the consensus network (right panel). Table 9 shows the "interaction strengths" corresponding to p-values of 0.1 and 0.01. Note that the p-values are used as descriptive measures, and no Bonferroni correction (which would be too conservative) was carried out.

p-value	**BN**	**LASSO**	**Consensus**
0.1	0.2	0.3	0.4
0.01	0.5	0.4	0.5

Table 9: Mapping from p-value thresholds to edge strengths/posterior probabilities.

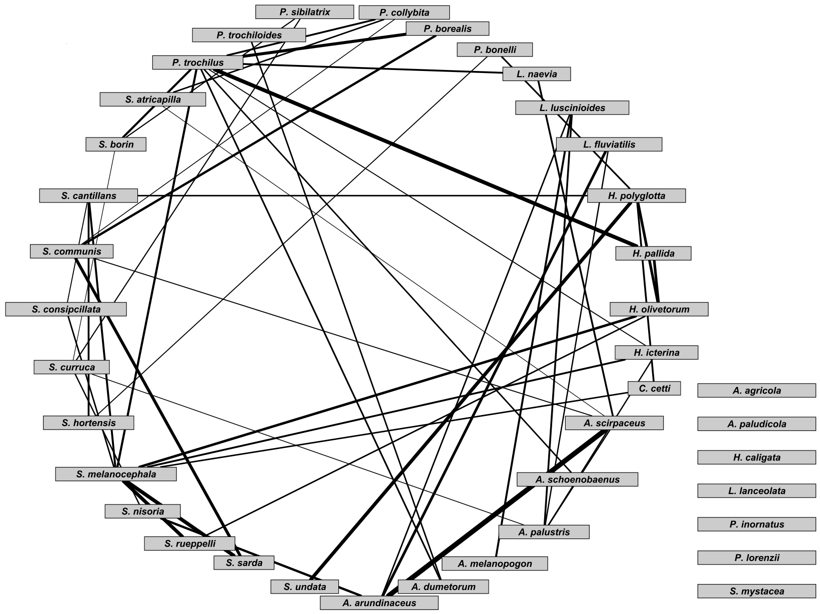


Figure 11: Consensus network recovered from the basic dataset (without spatial autocorrelation or bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

5.6. Recovered Networks

Figures 11-13 shows the consensus networks that were recovered from the warbler data. We get three different networks: one for the basic dataset, one for a dataset where we have modelled spatial autocorrelation as described in Section 3.2.1 of the main paper, and one for a dataset where we have included both spatial autocorrelation and two bio-climate covariates: temperature and availability of water. Details on how the sparsity and the correlation with the ecological measures vary for the different networks can be found in the main paper (Section 4.2).

5.7. Network Characterisation

Studies have shown that molecular regulatory networks have degree distributions that approximately follow a power-law (Wagner, 2001; Guelzim *et al.*,

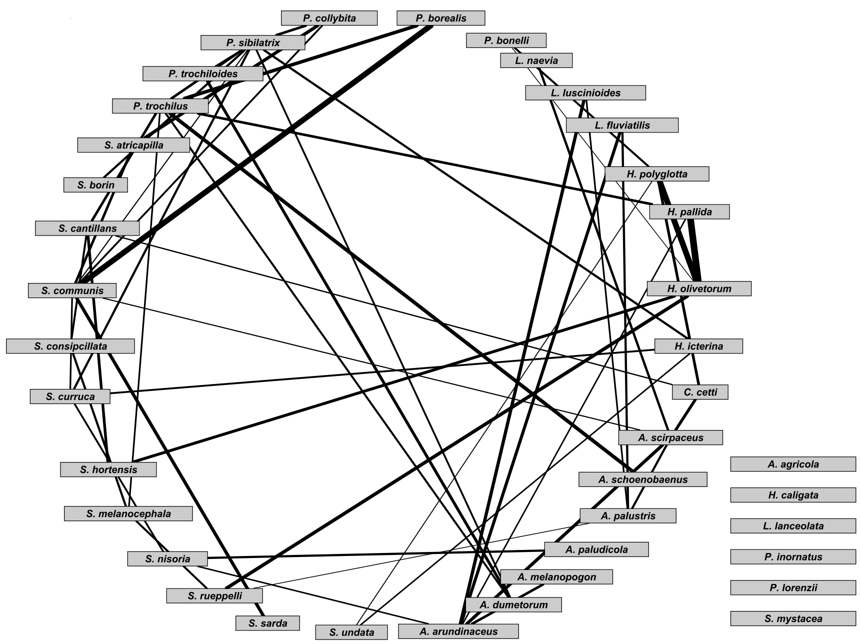


Figure 12: Consensus networks recovered from the dataset with spatial autocorrelation included (but without bio-climate covariates). The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

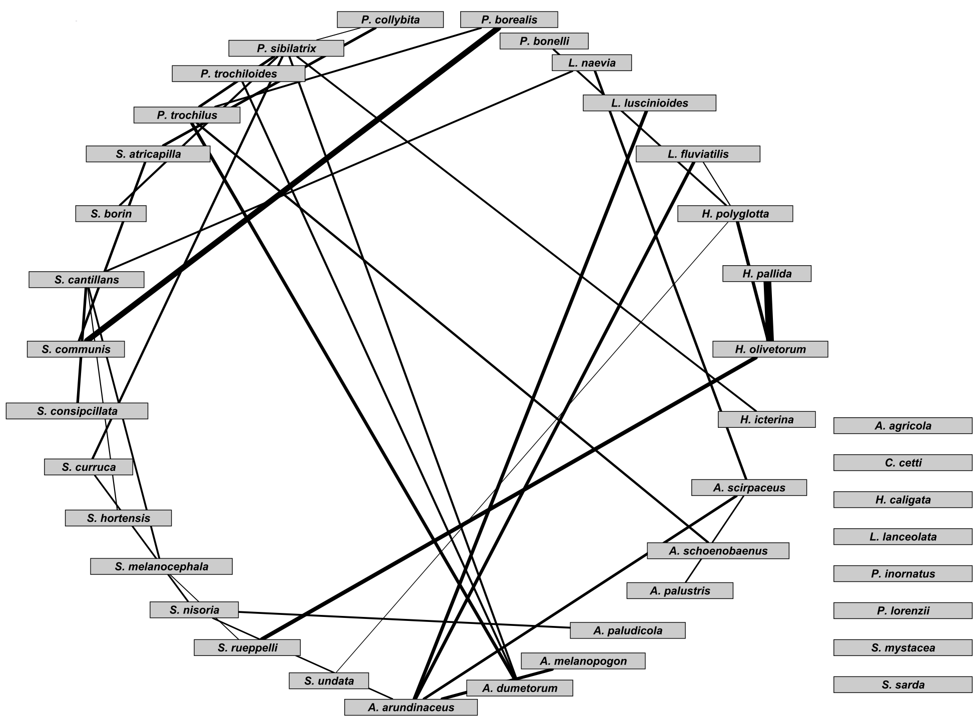


Figure 13: Consensus networks recovered from the dataset with both spatial autocorrelation and bio-climate covariates included. The edges are pruned by placing a threshold value of 0.5 on the original consensus network, which corresponds to a p-value of 0.01. See Section 5.5 for a description of how these p-values were calculated. The boxes on the right show unconnected species.

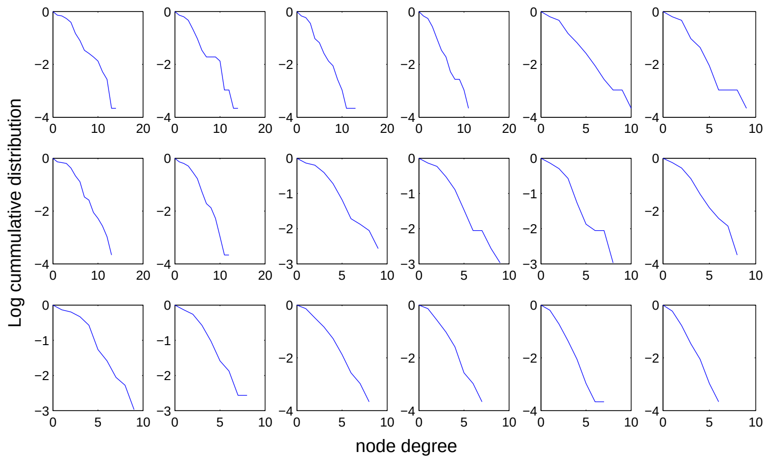


Figure 14: Cumulative degree distribution for the consensus networks on the **log-linear scale** as the threshold varies. (Top) Basic bird data, (Middle) Bird data with spatial autocorrelation model added, (Bottom) Birds with spatial autocorrelation and bio-climate covariates. From left to right the thresholds are set at p-values 0.2, 0.15, 0.1, 0.05, 0.02, and 0.01.

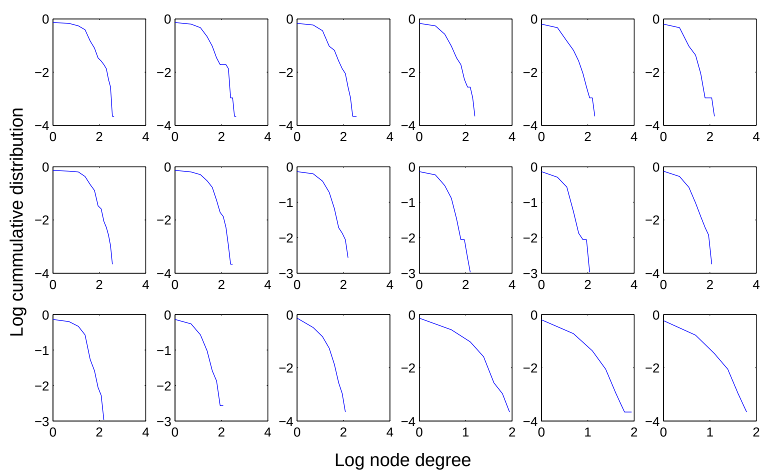
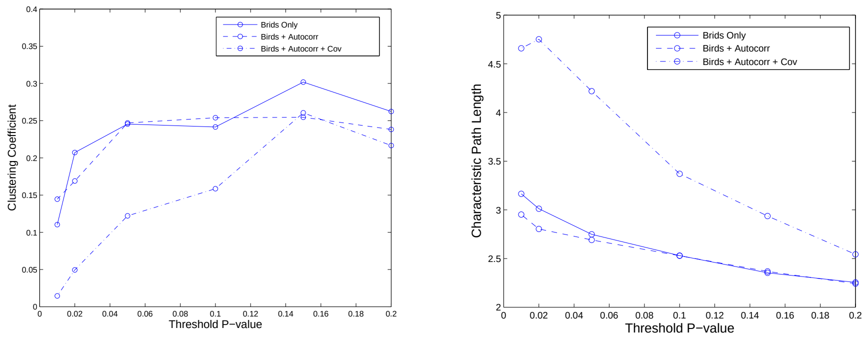


Figure 15: Cumulative degree distribution for the consensus networks on the **log-log scale** as the threshold varies. (Top) Basic bird data, (Middle) Bird data with spatial autocorrelation model added, (Bottom) Birds with spatial autocorrelation and bio-climate covariates. From left to right the thresholds are set at p-values 0.2, 0.15, 0.1, 0.05, 0.02, and 0.01.



(a) Clustering Coefficient

(b) Network Diameter

Figure 16: Variation of the clustering coefficient and network diameter for the consensus networks as the threshold varies.

2002; May, 2006). Loosely speaking, this means that there are many nodes with only one or few connections, but also some nodes with many more connections than the average degree. Studies on food webs generally agree that the degree distribution is not Poisson (Proulx *et al.*, 2005), however they disagree on whether the degree distributions are best fit by a power-law or by some other distribution. The existence of a variety of distributions has been shown, including power-law, truncated power-law and exponential (Dunne *et al.*, 2002; Jordano *et al.*, 2003; Laird and Jensen, 2006). In our study we observe that the distributions are closer to linear on the log-linear plot of the cumulative degree distribution (Fig. 14), than on the log-log plot (Fig. 15). Linearity on the log-log plot would be characteristic of a power-law distribution, but linearity on the log-linear plot shows that the network exhibits a near exponential distribution. The data also displays the insensitivity of this behaviour to varying the threshold.

Figure 16 shows the variation of the clustering coefficient and the network diameter (characteristic path length) as the threshold varies. There is no discernable trend, which may mean that these particular statistics are not useful characterisations of the types of networks that we are considering.

Acknowledgement

This work was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). We thank Jonathan Yearsley for providing the initial community simulation model and Marco Grzegorczyk for providing the Bayesian Network code and being available for questions. We would also like to thank Prof. Samuel Kaski for supporting Ali Faisal on this project.

References

Alstroem, P., Ericson, P., Olsson, U., Sundberg, P., 2006. Phylogeny and classification of the avian superfamily *Sylvioidea*. Mol. Phylogenet. Evol. 38 (2), 381–397.

Dunne, J., Williams, R., Martinez, N., 2002. Food-web structure and network theory: the role of connectance and size. Proc. Natl. Acad. Sci. 99 (20), 12917–12922.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comp. Biol. 7, 601–620.

Grafen, A., 1989. The phylogenetic regression. Phil. Trans. R. Soc. B 326, 119–157.

Grzegorczyk, M., Husmeier, D., Edwards, K., Ghazal, P., Millar, A., 2008. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics 24 (18), 2071–2078.

Guelzim, N., Bottani, S., Bourgine, P., Képès, F., 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat. Genet. 31 (1), 60–63.

Jordano, P., Bascompte, J., Olesen, J., 2003. Invariant properties in coevolutionary networks of plant-animal interactions. Ecol. Lett. 6 (1), 69–81.

Laird, S., Jensen, H., 2006. The Tangled nature model with inheritance and constraint: Evolutionary ecology restricted by a conserved resource. Ecol. Complex. 3 (3), 253–262.

Maddison, D., Schulz, K., Maddison, W., 2007. The tree of life web project. Zootaxa 1668, 19–40.

Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. Int. Stat. Rev. 63, 215–232.

May, R., 2006. Network structure and the biology of populations. Trends Ecol. Evol. 21 (7), 394–399.

Nobile, A., Fearnside, A., 2007. Bayesian finite mixtures with an unknown number of components: the allocation sampler. Stat. Comput. 17 (2), 147–162.

Proulx, S., Promislow, D., Phillips, P., 2005. Network thinking in ecology and evolution. Trends Ecol. Evol. 20 (6), 345–353.

Rogers, S., Girolami, M., 2005. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics 21 (14), 3131–3137.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4 (1), Article 32.

van Someren, E. P., Vaes, B. L. T., Steegenga, W. T., Sijbers, A. M., Dechering, K. J., Reinders, M. J. T., 2006. Least absolute regression network analysis of the murine osterblast differentiation network. Bioinformatics 22 (4), 477–484.

Wagner, A., 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol. 18 (7), 1283–1292.

Wei, Z., Li, H., 2007. A Markov random field model for network-based analysis of genomic data. Bioinformatics 23 (12), 1537.