1 **Title:** Estimating species interactions from observational data with Markov networks

2 **Author:** David J. Harris: Population Biology; 1 Shields Avenue, Davis CA, 95616

3 **Abstract:** Estimating species interactions from observational data is one of the most

4 controversial tasks in community ecology. One difficulty is that a single pairwise interaction

5 can ripple through an ecological network and produce surprising indirect consequences. For

6 example, the negative correlation between two competing species can be reversed in the

7 presence of a third species that is capable of outcompeting both of them. Here, I apply

8 models from statistical physics, called Markov networks or Markov random fields, that can

9 predict the direct and indirect consequences of any possible species interaction matrix.

10 Interactions in these models can be estimated from observational data via maximum

11 likelihood. Using simulated landscapes with known pairwise interaction strengths, I

12 evaluated Markov networks and six existing approaches. The Markov networks consistently

13 outperformed other methods, correctly isolating direct interactions between species pairs

14 even when indirect interactions largely overpowered them. Two computationally efficient

15 approximations, based on linear and generalized linear models, also performed well. Indirect

16 effects reliably caused a common null modeling approach to produce incorrect inferences,

17 however.

18 **Key words:** Ecological interactions; Occurrence data; Species associations; Markov network;

19 Markov random field; Ising model; Biogeography; Presence–absence matrix; Null model

20 **Introduction**

21 To the extent that nontrophic species interactions (such as competition) affect community

22 assembly, ecologists might expect to find signatures of these interactions in species

23 composition data (MacArthur 1958, Diamond 1975). Despite decades of work and several

24 major controversies, however (Lewin 1983, Strong et al. 1984, Gotelli and Entsminger 2003,

25 Connor et al. 2013), existing methods for detecting competition's effects on community

26 structure are unreliable (Gotelli and Ulrich 2009). In particular, species' effects on one

27 another can become lost in the complex web of direct and indirect interactions in real

28 assemblages. For example, the competitive interaction between the two shrub species in

29 Figure 1A can become obscured by their shared tendency to occur in unshaded areas (Figure

30 1B). While ecologists have long known that indirect effects can overwhelm direct ones at the

31 landscape level (Dodson 1970, Levine 1976), the vast majority of our methods for drawing

32 inferenes from observational data do not control for these effects (e.g. Diamond 1975, Strong

33 et al. 1984, Gotelli and Ulrich 2009, Veech 2013, Pollock et al. 2014). To the extent that

34 indirect interactions like those in Figure 1 are generally important (Dodson 1970), existing

35 methods will thus not generally provide much evidence regarding species' direct effects on

36 one another. The goal of this paper is to resolve this long-standing problem.

37   While competition doesn't reliably reduce co-occurrence rates at the whole-landscape

38 level (as most of our methods assume), it nevertheless does leave a signal in the data (Figure

39 1C). Specifically, after partitioning the data set into shaded sites and unshaded sites, there

40 will be co-occurrence deficits in each subset that might not be apparent at the landscape level.

41 More generally, we can obtain much better estimates of the association between two species

42 from their conditional relationships (i.e. by controlling for other species in the network) than

43 we could get from their overall co-occurrence rates. This kind of precision is difficult to

44 obtain from null models, which begin with the assumption that all the pairwise interactions

45 are zero and thus don't need to be controlled for. Nevertheless, null models have dominated

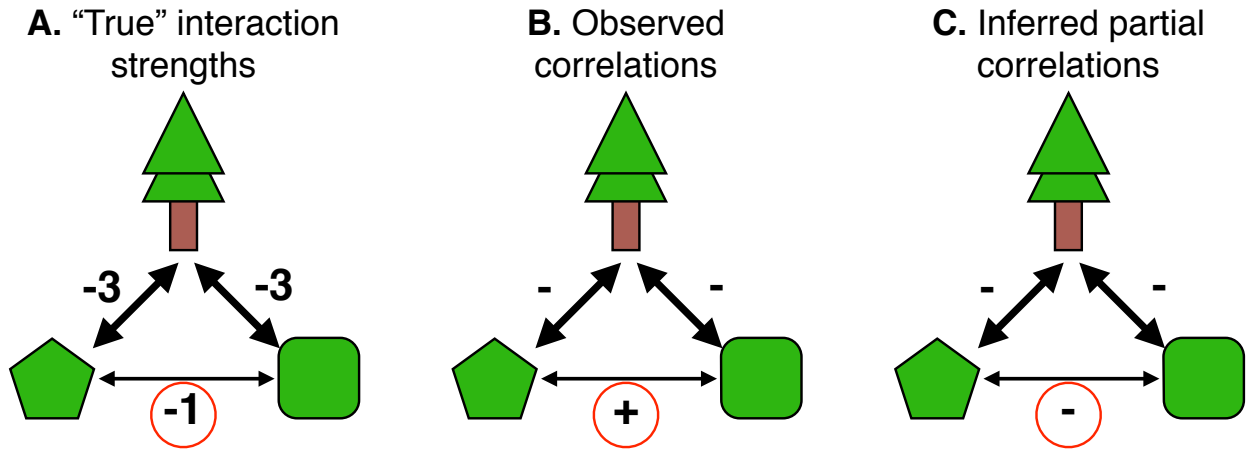46 this field for more than three decades (Strong et al. 1984, Gotelli and Ulrich 2009).

Figure 1: **Figure 1. A.** A small network of three competing species. The tree (top) tends not to co-occur with either of the two shrub species, as indicated by the strongly negative coefficient linking them. The two shrub species also compete with one another, but more weakly (circled coefficient). **B.** In spite of the competitive interactions between the two shrub species, their shared tendency to occur in locations without trees makes their occurrence vectors positively correlated (circled). **C.** Controlling for the tree species' presence with a conditional method such as a partial covariance or a Markov network leads to correct identification of the negative shrub-shrub interaction (circled).

47 Following recent work by Azaele et al. (2010) and Fort (2013), this paper shows that

48 Markov networks (undirected graphical models also known as Markov random fields; Murphy

49 2012) can provide a framework for understanding the landscape-level consequences of

50 pairwise species interactions, and for detecting them from observed presence-absence

51 matrices. Markov networks have been used in many scientific fields in similar contexts for

52 decades, from physics (where nearby particles interact magnetically; Cipra 1987) to spatial

53 statistics (where adjacent grid cells have correlated values; Harris 1974, Gelfand et al. 2005).

54 While community ecologists explored some related approaches in the 1980's (Whittam and

55 Siegel-Causey 1981), they used severe approximations that led to unintelligible results (e.g.

56 "probabilities" greater than one; Gilpin and Diamond 1982).

57 Below, I introduce Markov networks and show how they can be used to simulate

58 landscape-level data or to make exact predictions about the direct and indirect consequences

59 of possible interaction matrices. Then, using simulated data sets where the "true"

60 interactions are known, I compare this approach with several existing methods. Finally, I

61 discuss opportunities for extending the approach presented here to other problems in

62 community ecology, e.g. quantifying the overall effect of species interactions on occurrence

63 rates (Roughgarden 1983) and disentangling the effects of biotic versus abiotic interactions

64 on species composition (Kissling et al. 2012, Pollock et al. 2014).

65 **Methods**

66 **Markov networks.** Markov networks provide a framework for translating back and forth

67 between the conditional relationships among species (Figure 1C) and the kinds of species

68 assemblages that these relationships produce. Here, I show how a set of conditional

69 relationships can be used to determine how groups of species can co-occur. Methods for

4

70 estimating conditional relationships from data are discussed in the next section.

71 A Markov network defines the relative probability of observing a given vector of

72 species-level presences (1s) and absences (0s), $\vec{y}$, as

73 $$p(\vec{y}; \alpha, \beta) \propto exp(\sum_i \alpha_i y_i + \sum_{i \neq j} \beta_{ij} y_i y_j).$$

74 Here, α_i is an intercept term determining the amount that the presence of species i

75 contributes to the log-probability of $\vec{y}$; it directly controls the prevalence of species i.

76 Similarly, β_{ij} is the amount that the co-occurrence of species i and species j contributes to

77 the log-probability; it controls the conditional relationship between two species, i.e. the

78 probability that they will be found together, after controlling for the other species in the

79 network (Figure 2A, Figure 2B). For example, β_{ij} might have a value of $+2$ for two

80 mutualists, indicating that the odds of observing one species are e^2 times higher in sites

81 where its partner is present than in comparable sites where its partner is absent. Because the

82 relative probability of a presence-absence vector increases when positively-associated species

83 co-occur and decreases when negatively-associated species co-occur, the model tends—all

84 else equal—to produce assemblages that have many pairs of positively-associated species and

85 relatively few pairs of negatively-associated species (exactly as an ecologist might expect).

86 Of course, if all else is *not* equal (e.g. Figure 1, where the presence of one competitor is

87 associated with release from another competitor), then species' marginal association rates can

88 differ from this expectation. Determining the marginal relationships between species from

89 their conditional interactions entails summing over the different possible assemblages (Figure

90 2B). This becomes intractable when the number of possible assemblages is large, though

91 several methods beyond the scope of this paper can be employed to keep the calculations

92 feasible (Salakhutdinov 2008, Lee and Hastie 2012). Alternatively, as noted below, some
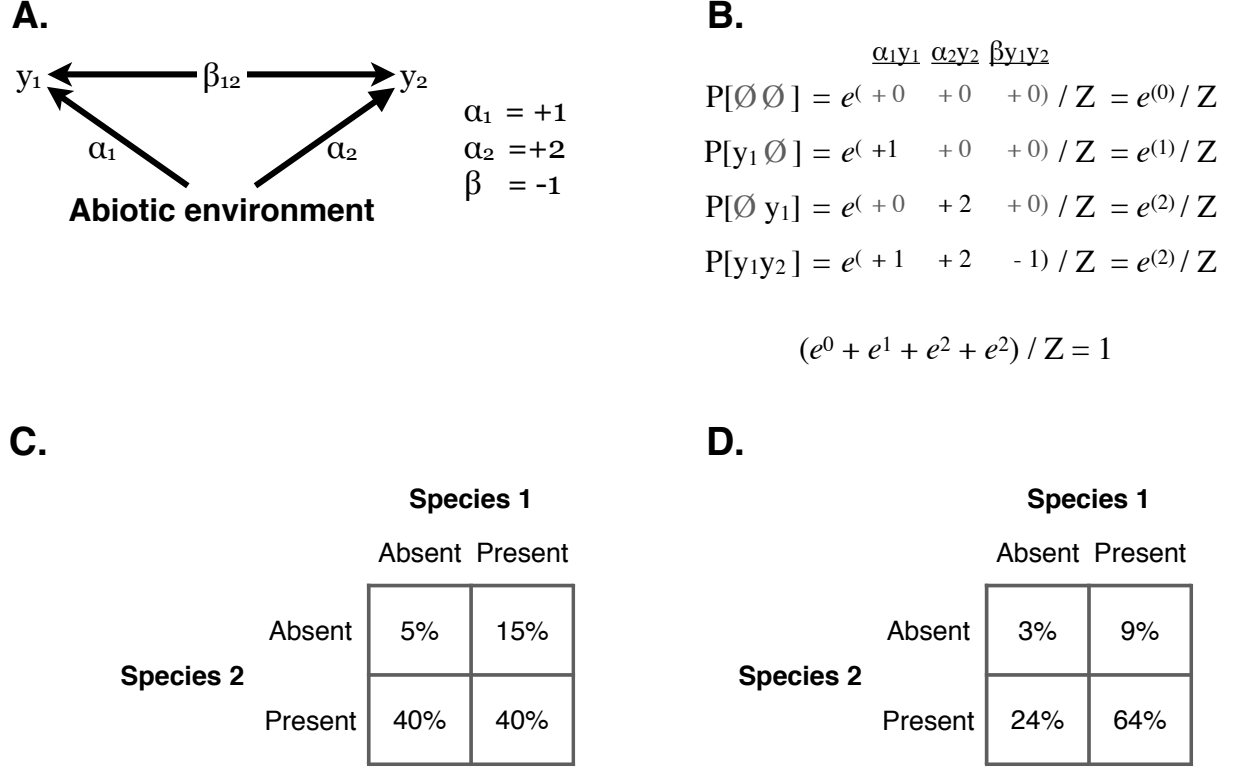
A.

$$y_1 \xleftrightarrow{\quad \beta_{12} \quad} y_2$$

$$\alpha_1 \qquad \alpha_2$$

Abiotic environment

$\alpha_1 = +1$
$\alpha_2 = +2$
$\beta = -1$

B.

$$\underline{\alpha_1 y_1} \ \underline{\alpha_2 y_2} \ \underline{\beta y_1 y_2}$$

$$P[\varnothing \varnothing] = e^{(\ +0 \quad +0 \quad +0)} / Z = e^{(0)} / Z$$

$$P[y_1 \varnothing] = e^{(\ +1 \quad +0 \quad +0)} / Z = e^{(1)} / Z$$

$$P[\varnothing \ y_1] = e^{(\ +0 \quad +2 \quad +0)} / Z = e^{(2)} / Z$$

$$P[y_1 y_2] = e^{(\ +1 \quad +2 \quad -1)} / Z = e^{(2)} / Z$$

$$(e^0 + e^1 + e^2 + e^2) / Z = 1$$

C.

	Species 1	
	Absent	Present
Species 2 Absent	5%	15%
Present	40%	40%

D.

	Species 1	
	Absent	Present
Species 2 Absent	3%	9%
Present	24%	64%

Figure 2: **Figure 2. A.** A small Markov network with two species. The abiotic environment favors the occurrence of both species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$). The negative β coefficient linking these two species implies that they co-occur less than expected under independence. **B.** Relative probabilities of all four possible presence-absence combinations for Species 1 and Species 2. The exponent includes α_1 whenever Species 1 is present ($y_1 = 1$), but not when it is absent ($y_1 = 0$). Similarly, the exponent includes α_2 only when species 2 is present ($y_2 = 1$), and β only when both are present ($y_1 y_2 = 1$). The normalizing constant Z, ensures that the four relative probabilities sum to 1. In this case, Z is about 18.5. **C.** We can find the expected frequencies of all possible co-occurrence patterns between the two species of interest. **D.** If β_{12} equaled zero (e.g. if the species no longer competed for the same resources), then the reduction in competition would allow each species to increase its occurrence rate and the co-occurrence deficit would be eliminated.

93 common linear and generalized linear methods can also be used as computationally efficient

94 approximations to the full network (Lee and Hastie 2012, Loh and Wainwright 2013).

95 **Estimating α and β coefficients from presence-absence data.**   In the previous

96 section, the values of α and β were known and the goal was to make predictions about

97 possible species assemblages. In practice, however, ecologists will often need to estimate the

98 parameters from an observed co-occurrence matrix (i.e. from a matrix of ones and zeros

99 indicating which species are present at which sites). When the number of species is

100 reasonably small, one can compute exact maximum likelihood estimates for all of the α and

101 β coefficients given a presence-absence matrix by optimizing $p(\vec{y}; \alpha, \beta)$. Fully-observed

102 Markov networks like the ones considered here have unimodal likelihood surfaces (Murphy

103 2012), ensuring that this procedure will always converge on the global maximum. This

104 maximum represents the unique combination of α and β coefficients that would be expected

105 to produce exactly the observed co-occurrence frequencies on average (i.e. maximizing the

106 likelihood matches the sufficient statistics of the model distribution to the sufficient statistics

107 of the data; Murphy 2012). I used the rosalia package (Harris 2015a) for the R programming

108 language (R Core Team 2015) to optimize the Markov network parameters. The package was

109 named after Santa Rosalia, the patron saint of biodiversity, whose supposedly miraculous

110 healing powers played an important rhetorical role in the null model debates of the 1970's

111 and 1980's (Lewin 1983).

112 **Simulated landscapes.**   In order to compare different methods, I simulated two sets of

113 landscapes using known parameters. The first set included the three competing species shown

114 in Figure 1. For each of 1000 replicates, I generated a landscape with 100 sites by sampling

115 from a probability distribution defined by the figure's interaction coefficients (Appendix 1).

116 Each of the methods described below was then evaluated on its ability to correctly infer that

117 the two shrub species competed with one another, despite their frequent co-occurrence.

118   I also simulated a second set of landscapes using a stochastic community model based

119 on generalized Lotka-Volterra dynamics, as described in Appendix 2. In these simulations,

120 each species pair was randomly assigned to either compete for a portion of the available

121 carrying capacity (negative interaction) or to act as mutualists (positive interaction). Here,

122 mutualisms operate by mitigating the effects of intraspecific competition on each partner's

123 death rate. For these analyses, I simulated landscapes with up to 20 species and 25, 200, or

124 1600 sites (50 replicates per landscape size; see Appendix 2).

125 **Recovering species interactions from simulated data.**   I compared seven techniques

126 for determining the sign and strength of the associations between pairs of species from

127 simulated data (Appendix 3). First, I used the rosalia package (Harris 2015a) to fit Markov

128 newtork models, as described above. For the analyses with 20 species, I added a very weak

129 logistic prior distribution on the α and β terms with scale 2 to ensure that the model

130 estimates were always finite. The bias introduced by this prior should be small: the 95%

131 credible interval on β only requires that one species' effect on the odds of observing a

132 different species to be less than a factor of 1500 (which is not much of a constraint). The

133 logistic distribution was chosen because it is convex and has a similar shape to the Laplace

134 distribution used in LASSO regularization (especially in the tails), but unlike the Laplace

135 distribution it is differentiable everywhere and does not force any estimates to be exactly

136 zero. To confirm that this procedure produced stable estimates, I compared its estimates on

137 50 bootstrap replicates (Appendix 4).

138   I also evaluated six alternative methods: five from the existing literature, plus a novel

8

139 combination of two of these methods. The first alternative interaction metric was the sample

140 correlation between species' presence-absence vectors, which summarizes their marginal

141 association. Next, I used partial correlations, which summarize species' conditional

142 relationships (Albrecht and Gotelli 2001, Faisal et al. 2010). In the context of non-Gaussian

143 data, the partial correlation can be thought of as a computationally efficient approximation

144 to the full Markov network model (Loh and Wainwright 2013). This sort of model is very

145 common for estimating relationships among genes and gene products (Friedman et al. 2008).

146 Because partial correlations are undefined for landscapes with perfectly-correlated species

147 pairs, I used a regularized estimate based on James-Stein shrinkage, as implemented in the

148 corpcor package's `pcor.shrink` function with the default settings (Schäfer et al. 2014).

149 The third alternative, generalized linear models (GLMs), can also be thought of as a

150 computationally efficient approximation to the Markov network (Lee and Hastie 2012).

151 Following Faisal et al. (2010), I fit regularized logistic regression models (Gelman et al. 2008)

152 for each species, using the other species on the landscape as predictors. To avoid the

153 identifiability problems associated with directed cyclic graphs (Schmidt and Murphy 2012), I

154 then symmetrized the relationships within species pairs via averaging.

155 The next method, described in Gotelli and Ulrich (2009), involved simulating new

156 landscapes from a null model that retains the row and column sums of the original matrix

157 (Strong et al. 1984). I used the *Z*-scores computed by the Pairs software described in Gotelli

158 and Ulrich (2009) as my null model-based estimator of species interactions.

159 The last two estimators used the latent correlation matrix estimated by the

160 BayesComm package (Golding and Harris 2015) in order to evaluate the recent claim that

161 the correlation coefficients estimated by "joint species distribution models" provide an

9

162 accurate assessment of species' pairwise interactions (Pollock et al. 2014, see also Harris

163 2015b). In addition to using the posterior mean correlation (Pollock et al. 2014), I also used

164 the posterior mean *partial* correlation, which might be able to control for indirect effects.

Evaluating model performance. For the simulated landscapes based on Figure 1, I

166 assessed whether each method's test statistic indicated a positive or negative relationship

167 between the two shrubs (Appendix 1). For the null model (Pairs), I calculated statistical

168 significance using its Z-score. For the Markov network, I used the Hessian matrix to

169 generate approximate confidence intervals and noted whether these intervals included zero.

170 I then evaluated the relationship between each method's estimates and the "true"

171 interaction strengths among all of the species pairs from the larger simulated landscapes.

172 This determined which of the methods provide a consistent way to know how strong species

173 interactions are—regardless of which species were present in a particular data set or how

174 many observations were taken. Because the different methods mostly describe species

175 interactions on different scales (e.g. correlations versus Z scores versus regression

176 coefficients), I used linear regression through the origin to rescale the different estimates

177 produced by each method so that they had a consistent interpretation. After rescaling each

178 method's estimates, I calculated squared errors between the scaled interaction estimates and

179 "true" interaction values across all the simulated data sets. These squared errors determined

180 the proportion of variance explained for different combinations of model type and landscape

181 size (compared with a null model that assumed all interaction strengths to be zero).

Results

Three species. As shown in Figure 1, the marginal relationship between the two shrub

184 species was positive—despite their competition for space at a mechanistic level—due to

indirect effects of the dominant tree species. As a result, the correlation between these

species was positive in 94% of replicates, and the randomization-based null model falsely

reported positive associations 100% of the time. Worse, more than 98% of these false

conclusions were statistically significant. The partial correlation and Markov network

estimates, on the other hand, each correctly isolated the direct negative interaction between

the shrubs from their positive indirect interaction 94% of the time (although the confidence

intervals overlapped zero in most replicates).

Twenty species.   Despite some variability across contexts (Figure 3A), the four methods

that controlled for indirect effects clearly performed the best: the Markov network explained

the largest portion of the variance in the "true" interaction coefficients (35% overall),

followed by the generalized linear models (30%), partial correlations from the raw

presence-absence data (28%), and partial correlations from BayesComm, the joint species

distribution model (26%). The benefit of choosing the full Markov network over the other

three methods was largest on the smaller landscapes, which are also the ones that are most

representative of typical analyses in this field (Gotelli and Ulrich 2009).

The three methods that did not attempt to control for indirect interactions all

explained less than 20% of the variance. Of these, the sample correlation matrix based on

the raw data performed the best (19%), followed by the null model (15%) and BayesComm's

correlation matrix (11%). Although these last three methods had different R^2 values, there

was a close mapping among their estimates (especially after controlling for the size of the

simulated landscapes; Figure 3B). This suggests that the effect sizes from the null model

(and, to a lesser extent, the correlation matrices from joint species distribution models) only

contain noisy versions of the same information that could be obtained more easily and
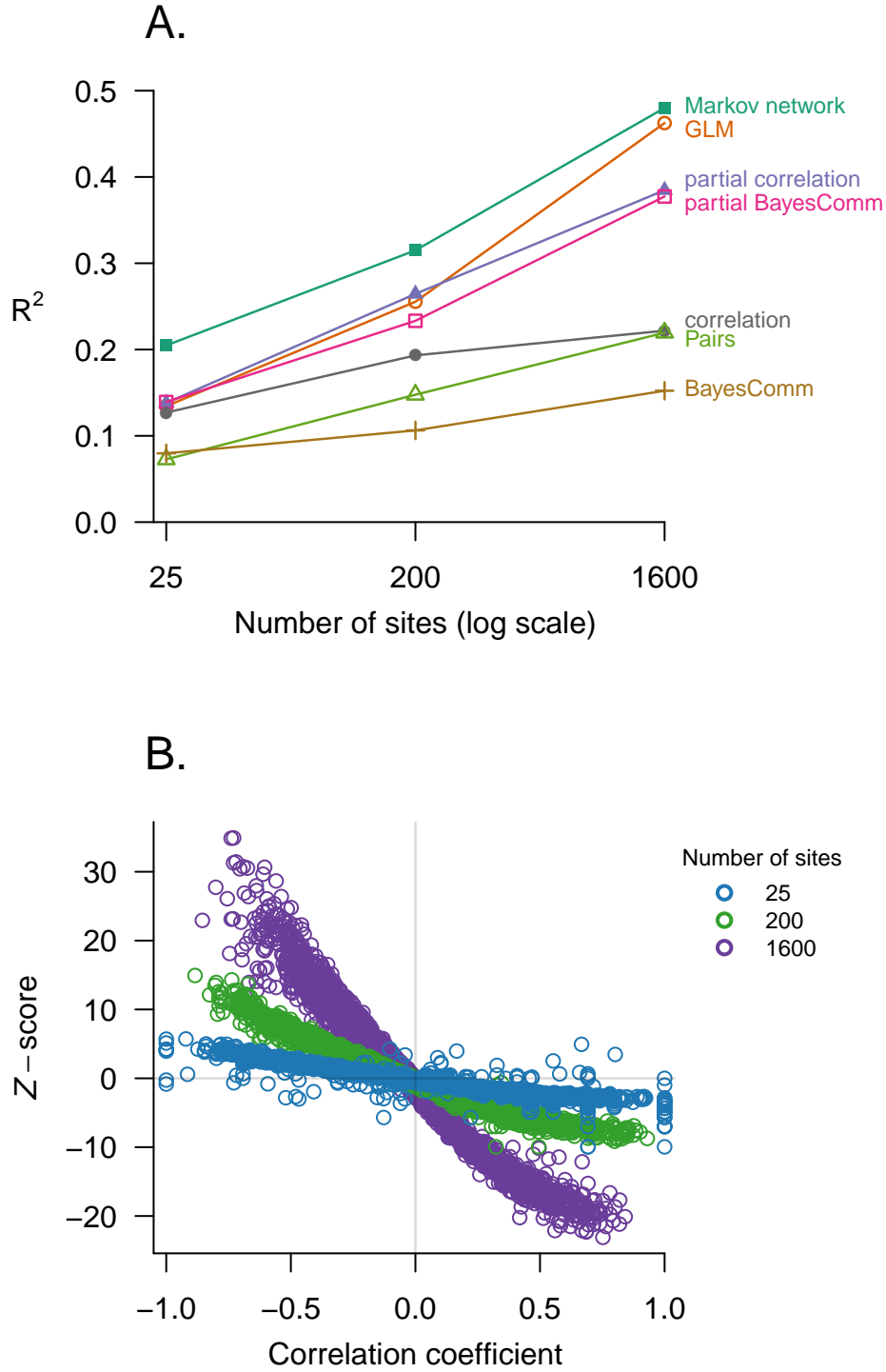
A.

B.

Figure 3: **Figure 3. A.** Proportion of variance in interaction coefficients explained by each method versus number of sampled locations. **B.** The Z-scores produced by the null model ("Pairs") for each pair of species can be predicted using the correlation between the presence-absence vectors of those same species and from the number of sites on the landscape.

208 interpretably by calculating correlation coefficients between species' presence-absence vectors.

209 Bootstrap resampling indicated that the above ranking of the different methods was

210 robust (Appendix 3). In particular, the 95% confidence interval of the bootstrap distribution

211 indicated that the Markov network's overall R^2 value was between 14 and 18 percent higher

212 than the second-most effective method (generalized linear models) and between 2.12 and 2.38

213 times higher than could be achieved by the null model (Pairs). Bootstrap resampling of a

214 200-site landscape also confirmed that the rosalia package's estimates of species' conditional

215 relationships were robust to sampling variation for reasonably-sized landscapes (Appendix 4).

216 **Discussion**

217 The results presented above show that Markov networks can reliably recover species'

218 pairwise interactions from observational data, even for cases where a common null modeling

219 technique reliably fails. Specifically, Markov networks were successful even when direct

220 interactions were largely overwhelmed by indirect effects (Figure 1). For cases where fitting a

221 Markov network is computationally infeasible, these results also indicate that partial

222 covariances and generalized linear models (the two methods that estimated conditional

223 relationships rather than marginal ones) can both provide useful approximations. The

224 partial correlations' success on simulated data may not carry over to real data sets, however;

225 Loh and Wainwright (2013) show that the linear approximations can be less reliable in cases

226 where the true interaction matrix contains more structure (e.g. guilds or trophic levels).

227 Similarly, the approximation involved in using separate generalized linear models for each

228 species can occasionally lead to catastrophic overfitting with small-to-moderate sample sizes

229 (Lee and Hastie 2012). For these reasons, it will usually be best to fit a Markov network

230 rather than one of the alternative methods when one's computational resources allow it.

It's important to note that none of these methods can identify the exact nature of the pairwise interactions (e.g. which species in a positively-associated pair is facilitating the other; Schmidt and Murphy 2012), particularly when real pairs of species can reciprocally influence one another in multiple ways simultaneously (Bruno et al. 2003); with compositional data, there is only enough information to provide a single number describing each species pair. To estimate asymmetric interactions, such as commensalism or predation, ecologists would need other kinds of data, as from time series, behavioral observations, manipulative experiments, or natural history. These other sources of information could also be used to augment the likelihood function with an informative prior distribution, which could lead to better results on some real data sets than was shown in Figure 3A.

Despite their limitations, Markov networks have enormous potential to improve ecological understanding. In particular, they are less vulnerable than some of the most commonly-used methods to mistakenly identifying positive species interactions between competing species, and can make precise statements about the conditions where indirect interactions will overwhelm direct ones. They also provide a simple answer to the question of how competition should affect a species' overall prevalence, which was a major flashpoint for the null model debates in the 1980's (Roughgarden 1983, Strong et al. 1984). Equation 1 can be used to calculate the expected prevalence of a species in the absence of biotic influences ($\frac{e^{\alpha}}{1+e^{\alpha}}$; Lee and Hastie 2012). Competition's effect on prevalence in a Markov network can then be calculated by subtracting this value from the observed prevalence (cf Figure 2D). This kind of insight would have been difficult to obtain without a generative model that makes predictions about the consequences of species interactions; null models (which presume *a priori* that interactions do not exist) have no way to make such predictions.

254   Markov networks—particularly the Ising model for binary networks—have been studied

255 for nearly a century (Cipra 1987), and the models' properties, capabilities, and limits are

256 well-understood in a huge range of applications. Using the same framework for species

257 interactions would thus allow ecologists to tap into an enormous set of existing discoveries

258 and techniques for dealing with indirect effects, stability, and alternative stable states.

259 Numerous other extensions are possible: for example, the states of the interaction network

260 can be modeled as a function of the local abiotic environment (Lee and Hastie 2012), which

261 would provide a rigorous and straightforward approach to the difficult and important task of

262 incorporating whole networks of biotic interactions into species distribution models (Kissling

263 et al. 2012, Pollock et al. 2014), leading to a better understanding of the interplay between

264 biotic and abiotic effects on community structure. There are even methods (Whittam and

265 Siegel-Causey 1981, Tjelmeland and Besag 1998) that would allow one species to affect the

266 sign or strength of the relationship between two other species, tipping the balance between

267 facilitation and exploitation (Bruno et al. 2003).

268   Finally, the results presented here have important implications for ecologists' continued

269 use of null models for studying species interactions. Null and neutral models can be useful

270 for clarifying our thinking about the numerical consequences of species' richness and

271 abundance patterns (Harris et al. 2011, Xiao et al. 2015), but deviations from a particular

272 null model must be interpreted with care (Roughgarden 1983). Even in small networks with

273 three species, it may simply not be possible to implicate individual species pairs or specific

274 ecological processes like competition by rejecting a general-purpose null (Gotelli and Ulrich

275 2009), especially when the test statistic is effectively just a correlation coefficient (Figure 3B).

276 Simultaneous estimation of multiple ecological parameters seems like a much more promising

277 approach: to the extent that the models' relative performance on real data sets is similar to

278 the range of results shown in Figure 3A, scientists in this field could often double their

279 explanatory power by switching from null models to Markov networks (or increase it

280 substantially with linear or generalized linear approximations). Regardless of the methods

281 ecologists ultimately choose, controlling for indirect effects could clearly improve our

282 understanding of species' direct effects on one another and on community structure.

288 **References:**

289 Albrecht, M., and N. J. Gotelli. 2001. Spatial and temporal niche partitioning in grassland

290   ants. Oecologia 126:134–141.

291 Azaele, S., R. Muneepeerakul, A. Rinaldo, and I. Rodriguez-Iturbe. 2010. Inferring plant

292   ecosystem organization from species occurrences. Journal of theoretical biology

293   262:323–329.

294 Bruno, J. F., J. J. Stachowicz, and M. D. Bertness. 2003. Inclusion of facilitation into

295   ecological theory. Trends in Ecology & Evolution 18:119–125.

296 Cipra, B. A. 1987. An introduction to the Ising model. American Mathematical Monthly

297   94:937–959.

298 Connor, E. F., M. D. Collins, and D. Simberloff. 2013. The checkered history of

299   checkerboard distributions. Ecology 94:2403–2414.

300 Diamond, J. M. 1975. The island dilemma: Lessons of modern biogeographic studies for the

301   design of natural reserves. Biological conservation 7:129–146.

302   Dodson, S. I. 1970. COMPLEMENTARY FEEDING NICHES SUSTAINED BY

303   SIZE-SELECTIVE PREDATION. Limnology and Oceanography 15:131–137.

304   Faisal, A., F. Dondelinger, D. Husmeier, and C. M. Beale. 2010. Inferring species interaction

305   networks from species abundance data: A comparative evaluation of various statistical and

306   machine learning methods. Ecological Informatics 5:451–464.

307   Fort, H. 2013. Statistical Mechanics Ideas and Techniques Applied to Selected Problems in

308   Ecology. Entropy 15:5237–5276.

309   Friedman, J., T. Hastie, and R. Tibshirani. 2008. Sparse inverse covariance estimation with

310   the graphical lasso. Biostatistics 9:432–441.

311   Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005.

312   Modelling species diversity through species level hierarchical modelling. Journal of the

313   Royal Statistical Society: Series C (Applied Statistics) 54:1–20.

314   Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su. 2008. A Weakly Informative Default

315   Prior Distribution for Logistic and Other Regression Models. The Annals of Applied

316   Statistics 2:1360–1383.

317   Gilpin, M. E., and J. M. Diamond. 1982. Factors contributing to non-randomness in species

318   Co-occurrences on Islands. Oecologia 52:75–84.

319   Golding, N., and D. J. Harris. 2015. BayesComm: Bayesian Community Ecology Analysis.

320   Gotelli, N. J., and G. L. Entsminger. 2003. Swap algorithms in null model analysis.

321   Ecology:532–535.

322   Gotelli, N. J., and W. Ulrich. 2009. The empirical Bayes approach as a tool to identify

323   non-random species associations. Oecologia 162:463–477.

324   Harris, D. J. 2015a. Rosalia: Exact inference for small binary Markov networks. R package

version 0.1.0. Zenodo. http://dx.doi.org/10.5281/zenodo.17808.

Harris, D. J. 2015b. Generating realistic assemblages with a Joint Species Distribution Model. Methods in Ecology and Evolution.

Harris, D. J., K. G. Smith, and P. J. Hanly. 2011. Occupancy is nine-tenths of the law: Occupancy rates determine the homogenizing and differentiating effects of exotic species. The American naturalist 177:535.

Harris, T. E. 1974. Contact Interactions on a Lattice. The Annals of Probability 2:969–988.

Kissling, W. D., C. F. Dormann, J. Groeneveld, T. Hickler, I. Kühn, G. J. McInerny, J. M. Montoya, C. Römermann, K. Schiffers, F. M. Schurr, A. Singer, J.-C. Svenning, N. E. Zimmermann, and R. B. O'Hara. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography 39:2163–2178.

Lee, J. D., and T. J. Hastie. 2012, May. Learning Mixed Graphical Models.

Levine, S. H. 1976. Competitive Interactions in Ecosystems. The American Naturalist 110:903–910.

Lewin, R. 1983. Santa Rosalia Was a Goat. Science 221:636–639.

Loh, P.-L., and M. J. Wainwright. 2013. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. The Annals of Statistics 41:3022–3049.

MacArthur, R. H. 1958. Population ecology of some warblers of northeastern coniferous forests. Ecology 39:599–619.

Murphy, K. P. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and

349 Evolution:n/a–n/a.

350 R Core Team. 2015. R: A Language and Environment for Statistical Computing. R

351 Foundation for Statistical Computing, Vienna, Austria.

352 Roughgarden, J. 1983. Competition and Theory in Community Ecology. The American

353 Naturalist 122:583–601.

354 Salakhutdinov, R. 2008. Learning and evaluating Boltzmann machines. Technical Report

355 UTML TR 2008-002, Department of Computer Science, University of Toronto, Dept. of

356 Computer Science, University of Toronto.

357 Schäfer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmäki, A. P. D. Silva, and K. Strimmer.

358 2014. Corpcor: Efficient Estimation of Covariance and (Partial) Correlation.

359 Schmidt, M., and K. Murphy. 2012. Modeling Discrete Interventional Data using Directed

360 Cyclic Graphical Models. arXiv preprint arXiv:1205.2617.

361 Strong, D. R., D. Simberloff, L. G. Abele, and A. B. Thistle. 1984. Ecological communities:

362 Conceptual issues and the evidence. Princeton University Press.

363 Tjelmeland, H., and J. Besag. 1998. Markov Random Fields with Higher-order Interactions.

364 Scandinavian Journal of Statistics 25:415–433.

365 Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence. Global Ecology

366 and Biogeography 22:252–260.

367 Whittam, T. S., and D. Siegel-Causey. 1981. Species Interactions and Community Structure

368 in Alaskan Seabird Colonies. Ecology 62:1515–1524.

369 Xiao, X., D. J. McGlinn, and E. P. White. 2015. A strong test of the Maximum Entropy

370 Theory of Ecology. The American Naturalist 185:E70–E80.