1 **Title:** Inferring species interactions from co-occurrence data with Markov networks

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

3 **Abstract:** Inferring species interactions from co-occurrence 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 outcompetes both of them. Here, I apply models from
8 statistical physics, called Markov networks or Markov random fields, that can predict the
9 direct and indirect consequences of any possible species interaction matrix. Interactions in
10 these models can be estimated from observed co-occurrence rates via maximum likelihood,
11 controlling for indirect effects. Using simulated landscapes with known interactions, 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 or abiotic factors largely overpowered them. Two
15 computationally efficient approximations, which controlled for indirect effects with partial
16 correlations or generalized linear models, also performed well. Null models showed no
17 evidence of being able to control for indirect effects, and reliably yielded incorrect inferences
18 when such effects were present.

21 **Introduction**

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

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

₂₄ composition data (MacArthur 1958, Diamond 1975). Despite decades of work and several

₂₅ major controversies, however (Lewin 1983, Strong et al. 1984, Connor et al. 2013), existing

₂₆ methods for detecting competition's effects on community structure are unreliable (Gotelli

₂₇ and Ulrich 2009). In particular, species' effects on one another can become lost in a web of

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

₂₉ Figure 1A is obscured by their shared tendency to occur in unshaded areas (Figure 1B).

₃₀ While ecologists have long known that indirect effects can overwhelm direct ones (Levine

₃₁ 1976), most methods for drawing inferences from co-occurrence data do not control for these

₃₂ effects (e.g. Diamond 1975, Strong et al. 1984, Gotelli and Ulrich 2009, Veech 2013, Pollock

₃₃ et al. 2014). As a result, ecologists do not have tools that allow them to isolate direct

₃₄ interactions from indirect effects.

₃₅ While competition doesn't reliably reduce co-occurrence rates at the whole-landscape

₃₆ level (as most methods assume), it does still leave a signal in the data (Figure 1C). After

₃₇ controlling for the presence of the tree species (e.g. by splitting the data set into shaded and

₃₈ unshaded sites or by using a model that estimates conditional relationships amongs species),

₃₉ the two shrubs do have a negative association, with fewer co-occurrences in unshaded areas

₄₀ than would be expected if they occurred independently of one another.

₄₁ Following Azaele et al. (2010), this paper shows that Markov networks (undirected

₄₂ graphical models also known as Markov random fields; Murphy 2012) can provide a

₄₃ framework for understanding the landscape-level consequences of pairwise species

₄₄ interactions, and for estimating them from observed presence-absence matrices. Markov

₄₅ networks have been used in diverse scientific fields for decades, from physics (where nearby

₄₆ particles interact magnetically; Cipra 1987) to spatial statistics (where adjacent grid cells

₄₇ have correlated values; Harris 1974, Gelfand et al. 2005). While community ecologists

2

48 explored some related approaches in the 1980's (Whittam and Siegel-Causey 1981), they

49 used severe approximations that led to unintelligible results (e.g. "probabilities" greater than

50 one; Gilpin and Diamond 1982).

51   Below, I demonstrate Markov networks' ability to produce exact predictions about the

52 direct and indirect consequences of an interaction matrix, and also to make inferences about

53 the interaction matrix based on co-occurrence rates. Using simulated data sets where the

54 "true" interactions are known, I compare this approach with several existing methods.

55 Finally, I discuss opportunities for extending the approach presented here to other problems

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

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

58 on species composition (Pollock et al. 2014).

59 **Methods**

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

61 between the conditional (all-else-equal) relationships among species (Figure 1C) and the

62 kinds of species assemblages that these relationships produce. Here, I show how a set of

63 conditional relationships can determine species composition. Methods for estimating

64 conditional relationships from data are discussed in the next section.

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

66 species-level presences (1s) and absences (0s), $\vec{y}$ at a site, as

67
$$p(\vec{y}; \alpha, \beta) \propto exp(\sum_i \alpha_i y_i + \sum_{<ij>} \beta_{ij} y_i y_j),$$

68 where the second sum is over all $\frac{1}{2}n(n-1)$ pairs of n species. In this model, α_i is an intercept

69 term determining the amount that the presence of species i contributes to the log-probability

70 of $\vec{y}$; it directly controls the prevalence of species i. Similarly, β_{ij} is the amount that the

3

71 co-occurrence of species i and species j contributes to the log-probability; it determines the

72 conditional relationship between two species, i.e. the probability that they will be found

73 together, after controlling for the other species in the network (Figure 2A, Figure 2B). For

74 example, if $\beta_{ij} = +2$, then each species' odds of occurrence would be e^2 times higher when

75 the other one is present (as compared with otherwise equivalent sites). The relative

76 probability of a presence-absence vector increases when positively-associated species co-occur

77 and decreases when negatively-associated species co-occur. As a result, the model tends—all

78 else equal—to produce assemblages where many positively-associated species pairs co-occur

79 and few negatively-associated pairs do (just as an ecologist might expect).

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

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

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

83 parameters from an observed co-occurrence matrix (i.e. from a set of independent $\vec{y}$ vectors

84 indicating which species are present at each site on the landscape). When Equation 1 can be

85 normalized (Figure 2B), one can find exact maximum likelihood estimates for α and β by

86 numerically optimizing $p(\vec{y}|\alpha, \beta)$. Fully-observed networks like the ones considered here have

87 unimodal likelihood surfaces (Murphy 2012), so these estimates will represent the global

88 optimum.

89   When the number of species is larger than about 30, noralizing Equation 1 can become

90 intractable, and researchers will either need to approximate it (Lee and Hastie 2012) or

91 approximate its gradient (Harris 2016) if they want to fit a Markov network model. For the

92 analyses presented here, where the number of species did not exceed 20, approximations were

93 not necessary. Instead, I used the rosalia package (Harris 2015a) for the R programming

94 language (R Core Team 2015) to calculate $p(\vec{y}; \alpha, \beta)$ and its gradients exactly (Murphy

4

95 2012); the package passes these functions to the "BFGS" method in R's general-purpose

96 optimizer, which finds values for α and β.

97 **Simulated landscapes.** I simulated several sets of landscapes using known parameters so

98 that model estimates could be compared with "true" values. The first set of landscapes

99 included the three competing species shown in Figure 1. For each of 1000 replicates, I

100 generated a landscape with 100 sites by sampling from a probability distribution defined by

101 the figure's interaction coefficients (Appendix 1). Each of the methods described below was

102 then evaluated on its ability to correctly infer that the two shrub species competed with one

103 another, despite their frequent co-occurrence in unshaded areas.

104 I then simulated landscapes with up to 20 interacting species at 25, 200, or 1600 sites

105 using three increasingly complex models (50 replicates for each combination of size and

106 model). The simplest set of simulated landscapes were generated with Gibbs samples from

107 Equation 1. For each replicate, I randomly drew the "true" β coefficient magnitudes from an

108 exponential distribution with rate 1 so that most species pairs interacted negligibly but a few

109 interactions were strong enough to propagate through the newtwork. I randomly assigned

110 25% of the interactions to be positive; the remainder were negative.

111 The next set of landscapes provided a way to assess each method's ability to identify

112 direct interactions in the presence of environmental heterogeneity. Here, the β coefficients

113 were calculated as above, but each species' α value depended linearly on two environmental

114 factors, which were drawn from independent Gaussians for each site. Once the local α values

115 were calculated, independent Gibbs samplers determined the species composition for each

116 site based on Equation 1.

117 In the final set of landscapes, I simulated each species' abundance (instead of just

118 presence/absence); furthermore, interactions between species occurred on a per-capita basis

5

119 in these simulations. To prevent runaway mutualisms leading to infinite abundance, all

120 interaction coefficients were negative in these simulations. Good performance on these

121 landscapes would indicate some robustness to the mechanistic details of species interactions.

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

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

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

125 network models, as described above. For the analyses with 20 species, a weakly-informative

126 regularizer (equivalent to a logistic prior with location 0 and scale 2) ensured that the

127 estimates were always finite (Appendix 2).

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

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

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

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

132 relationships. This approach, which is closely related to linear regression, is common in

133 molecular biology (Friedman et al. 2008), but is rare in ecology (see Albrecht and Gotelli

134 (2001) and Faisal et al. (2010) for two exceptions). In the context of non-Gaussian data, the

135 partial correlation (or partial covariance) can be thought of as a computationally efficient

136 approximation to the full Markov network model (Loh and Wainwright 2013). Partial

137 correlations are undefined for landscapes with perfectly-correlated species pairs, so I used the

138 regularized estimate provided by the corpcor package's `pcor.shrink` function with the

139 default settings (Schäfer et al. 2014).

140   The third alternative, generalized linear models (GLMs), also provide a computationally

141 efficient approximation to the Markov network (Lee and Hastie 2012). Following Faisal et al.

142 (2010), I fit regularized logistic regression models (Gelman et al. 2008) for each species, using

143 the other species on the landscape as predictors. This produced two interaction estimates for

144 each species pair; one for the effect of species i on species j and one for the reverse. These

145 two coefficients are not identifiable from the data, however (Schmidt and Murphy 2012), so I

146 used their average as an overall measure of the pairwise relationship.

147 The next method used the Pairs software described in Gotelli and Ulrich (2009). This

148 program simulates new landscapes from a null model that retains the row and column sums

149 of the original matrix (Strong et al. 1984) and calculates Z-scores to summarize a species

150 pair's deviation from this null.

151 The last two estimators used the latent correlation matrix estimated by the BayesComm

152 package (Golding and Harris 2015) in order to evaluate the recent claim that the correlation

153 coefficients estimated by "joint species distribution models" provide an accurate assessment

154 of species' pairwise interactions (Pollock et al. 2014, see also Harris 2015b). In addition to

155 using the posterior mean correlation (Pollock et al. 2014), I also used the posterior mean

156 *partial* correlation, which should control better for indirect effects.

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

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

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

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

161 generate approximate confidence intervals.

162 For the larger landscapes, I evaluated the relationship between each method's estimates

163 and the "true" interaction strengths. To ensure that the different test statistics

164 (e.g. correlations versus Z scores) were on a common scale, I rescaled them using linear

165 regression through the origin. I then calculated the proportion of variance explained for

166 different combinations of model type and landscape size (compared with a baseline model

7

167 that assumed all interaction strengths to be zero).

168   For the null model and the Markov network, the probability of rejecting the null

169 hypothesis of zero interaction was estimated across a range of "true" interaction strengths

170 using a generalized additive model with binomial responses (reject versus not; Appendix 4).

171 The probability of rejection when the "true" value of β was zero was defined as the Type I

172 error rate.

Results

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

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

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

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

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

179 conclusions were statistically significant. The partial correlation and Markov network

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

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

182 intervals overlapped zero in most replicates).

Twenty species. In general, each model's performance was highest for large landscapes

184 with simple assembly rules and no environmental heterogeneity (Figure 3). Despite some

185 variability across contexts, the rank ordering across methods was very consistent. In

186 particular, the four methods that controlled for indirect effects (the Markov network, the

187 generalized linear models, and the two partial correlation-based methods) always matched or

188 outperformed those that did not. The Markov network consistently performed best of all. As

189 anticipated by Lee and Hastie (2012), generalized linear models closely approximated the

190 Markov network estimates (Figure 4A), especially when the data sets were very large (Figure

191 3). As reviewed in Gotelli and Ulrich (2009), however, most analyses in this field of ecology

192 involve fewer than 50 sites; in this context, the gap between the methods was larger. As

193 shown in Appendix 4, the standard errors associated with the estimates in Figure 3 are small

194 (less than 0.01), so the differences among methods should not be attributed to sampling error.

195   Of the methods that did not control for indirect effects, Figure 3 shows that simple

196 correlation coefficients provided a more reliable indicator of species' true interaction strengths

197 than either the joint species distribution model (BayesComm) or the null model (Pairs). 95%

198 of the variance the Pairs test statistic was explained by correlation coefficients (controlling

199 for landscape size; Figure 4B); much of the remaining variance is due to sampling error.

200   Finally, we can evaluate the models' statistical inferences (focusing on the first two

201 simulation types, where the interaction coefficients are easiest to interpret). The Markov

202 network's Type I error rate depended on the simulation type: 0.02 for simulations that

203 matched the model's assumptions, versus 0.14 for simulations that included environmental

204 heterogeneity (see Appendix 4 for confidence interval coverage across a range of β_{ij} values).

205 The null model's Type I error rates were 0.31 and 0.53 for the constant and heterogeneous

206 landscapes, respectively—far higher than the nominal 0.05 rate. Figure 4C shows, across a

207 range of true interaction strengths, the probability that the null model or the Markov

208 network will predict the wrong sign of the interaction with 95% confidence. The null model

209 makes such errors more than 5 times as often as the Markov network, even though it only

210 rejects the null hypothesis twice as often overall (Appendix 4). The Markov network's errors

211 were also more concentrated around 0, and it never misclassified strong interactions like the

212 null model did (Figure 4C).

213 **Discussion**

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

9

215 pairwise interactions from species composition data, even for cases where environmental

216 heterogeneity and indirect interactions cause ecologists' typical null modeling approaches to

217 reliably fail. Partial correlations and generalized linear models can both provide

218 computationally efficient approximations, but with somewhat lower accuracy (especially for

219 typically-sized data sets with small numbers of sites; Gotelli and Ulrich 2009). The difference

220 in accuracy may be even larger for real data sets than for the simulated landscapes in Figure

221 3; linear approximations to the Markov network make larger errors when the interaction

222 matrix is structured (e.g. due to guilds or trophic levels; Loh and Wainwright 2013).

223 Similarly, the separate generalized linear models for each species can severely overfit in some

224 cases (Lee and Hastie 2012). The full Markov network should thus be preferred to the

225 approximations when it is computationally tractable.

226   Compositional data only contains enough degrees of freedom to estimate one interaction

227 per species pair (Schmidt and Murphy 2012), so none of these methods can identify the

228 exact nature of the pairwise interactions (e.g. which species in a positively-associated pair is

229 facilitating the other). To estimate asymmetric interactions, such as commensalism or

230 predation, ecologists could use time series, behavioral observations, manipulative

231 experiments, or natural history. These other sources of information could also be used to

232 augment the likelihood function with a more informative prior distribution, reducing

233 ecologists' error and uncertainty relative to Figure 3's results.

234   Despite their limitations, Markov networks have enormous potential to improve

235 ecological understanding. In particular, they make many fewer errors than existing

236 approaches, and can make precise statements about the conditions where indirect

237 interactions will overwhelm direct ones. They also provide a simple answer to the question of

238 how competition should affect a species' overall prevalence, which has important implications

for community-level modeling (Strong et al. 1984). Specifically, Equation 1 can be used to calculate the expected prevalence of a species in the absence of biotic influences as $e^{\alpha_i}/(e^0 + e^{\alpha_i})$. Competition's effect on prevalence can then be estimated by comparing this value with the observed prevalence (e.g. comparing Figure 2D with Figure 2C). This novel quantitative result undermines most of our null models, which unreasonably assume that prevalence would be the exactly same in the absence of competition as it is in the observed data (Roughgarden 1983).

Markov networks—particularly the Ising model for binary networks—are very well understood, having been studied for nearly a century (Cipra 1987). Tapping into this framework would thus allow ecologists to take advantage of into a vast set of existing discoveries and techniques for dealing with indirect effects, stability, and alternative stable states. Numerous extensions to the basic network are possible as well. For example, the states of the interaction network can be modeled as a function of the local abiotic environment (Lee and Hastie 2012, Harris 2016), which would lead to a better understanding of the interplay between biotic and abiotic effects on community structure. Alternatively, models could allow one species to alter the relationship between two other species (Whittam and Siegel-Causey 1981, Tjelmeland and Besag 1998, cf Bruno et al. 2003).

Finally, the results presented here have important implications for ecologists' continued use of null models for studying species interactions. Null and neutral models can be useful for clarifying our thinking (Harris et al. 2011, Xiao et al. 2015), but deviations from a particular null model must be interpreted with care (Roughgarden 1983). Even in small networks with three species, it may simply not be possible to implicate specific ecological processes like competition by rejecting a general-purpose null (Gotelli and Ulrich 2009), especially when the test statistic is effectively just a correlation coefficient (Figure 4B).

When the non-null backdrop is not controlled for, Type I error rates can skyrocket, the apparent sign of the interaction can change, and null models can routinely produce misleading inferences (Figure 1, Figure 4C, Gotelli and Ulrich (2009)).

Controlling for indirect effects via simultaneous estimation of multiple ecological parameters seems like a much more promising approach: to the extent that the models' relative performance on real data sets is similar to the range of results shown in Figure 3, scientists in this field could often triple their explanatory power by switching from null models to Markov networks (or increase it nearly as much with linear or generalized linear approximations). Regardless of the methods ecologists ultimately choose, controlling for indirect effects could clearly improve our understanding of species' direct effects on one another and on community structure.

References:

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

287   occurrence vectors positively correlated (circled). **C.** Controlling for trees with a

288   conditional (all-else-equal) approach such as a partial correlation or a Markov network

289   leads to correct identification of the negative shrub-shrub interaction (circled). See

290   Appendix 1 and the results for "three species" for more details.

291 **Figure 2: A.** A small Markov network, defined by its α and β values. The abiotic

292   environment favors the occurrence of each species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$).

293   The negative β_{12} coefficient is consistent with competition between the two species. **B.**

294   The coefficients determine the probabilities of all four possible presence-absence

295   combinations for Species 1 and Species 2. α_1 is added to the exponent whenever Species 1

296   is present ($y_1 = 1$), but not when it is absent ($y_1 = 0$). Similarly, the exponent includes α_2

297   only when species 2 is present ($y_2 = 1$), and includes β_{12} only when both are present

298   ($y_1 y_2 = 1$). The normalizing constant Z, ensures that the four probabilities sum to 1. In

299   this case, Z is about 18.5. **C.** The expected frequencies of all possible co-occurrence

300   patterns between the two species of interest, as calculated in the previous panel. **D.**

301   Without competition (i.e. with $\beta_{12} = 0$, each species would occur more often.

302 **Figure 3:** Proportion of variance in interaction coefficients explained by each method versus

303   number of sampled locations across the three simulation types. For the null model (Pairs),

304   two outliers with $|Z| > 1000$ were manually adjusted to $|Z| = 50$ to mitigate their

305   detrimental influence on R^2 (Appendix 5).

306 **Figure 4: A.** The Markov network's estimated interaction coefficients were generally very

307   similar to the GLM estimates. **B.** The null model's estimates typically matched the

308   (negative) correlation coefficient, after controlling for landscape size. **C.** For any given

309   interaction strength, the null model was much more likely to misclassify its sign with 95%

310   confidence than the Markov network was.

Albrecht, M., and N. J. Gotelli. 2001. Spatial and temporal niche partitioning in grassland ants. Oecologia 126:134–141.

Azaele, S., R. Muneepeerakul, A. Rinaldo, and I. Rodriguez-Iturbe. 2010. Inferring plant ecosystem organization from species occurrences. Journal of theoretical biology 262:323–329.

Bruno, J. F., J. J. Stachowicz, and M. D. Bertness. 2003. Inclusion of facilitation into ecological theory. Trends in Ecology & Evolution 18:119–125.

Cipra, B. A. 1987. An introduction to the ising model. American Mathematical Monthly 94:937–959.

Connor, E. F., M. D. Collins, and D. Simberloff. 2013. The checkered history of checkerboard distributions. Ecology 94:2403–2414.

Diamond, J. M. 1975. The island dilemma: Lessons of modern biogeographic studies for the design of natural reserves. Biological conservation 7:129–146.

Faisal, A., F. Dondelinger, D. Husmeier, and C. M. Beale. 2010. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. Ecological Informatics 5:451–464.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54:1–20.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su. 2008. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2:1360–1383.

Gilpin, M. E., and J. M. Diamond. 1982. Factors contributing to non-randomness in species co-occurrences on islands. Oecologia 52:75–84.

Golding, N., and D. J. Harris. 2015. BayesComm: Bayesian community ecology analysis.

Gotelli, N. J., and W. Ulrich. 2009. The empirical bayes approach as a tool to identify non-random species associations. Oecologia 162:463–477.

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. 2016. Estimating species interactions in large, abiotically structured communities using markov networks and stochastic approximation. bioRxiv.

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.

Lee, J. D., and T. J. Hastie. 2012. 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 Evolution:n/a–n/a.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roughgarden, J. 1983. Competition and theory in community ecology. The American Naturalist 122:583–601.

Schäfer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmäki, A. P. D. Silva, and K. Strimmer. 2014. Corpcor: Efficient estimation of covariance and (partial) correlation.

Schmidt, M., and K. Murphy. 2012. Modeling discrete interventional data using directed cyclic graphical models. arXiv preprint arXiv:1205.2617.

Strong, D. R., D. Simberloff, L. G. Abele, and A. B. Thistle. 1984. Ecological communities: Conceptual issues and the evidence. Princeton University Press.

Tjelmeland, H., and J. Besag. 1998. Markov random fields with higher-order interactions. Scandinavian Journal of Statistics 25:415–433.

Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence. Global Ecology and Biogeography 22:252–260.

Whittam, T. S., and D. Siegel-Causey. 1981. Species interactions and community structure in alaskan seabird colonies. Ecology 62:1515–1524.

Xiao, X., D. J. McGlinn, and E. P. White. 2015. A strong test of the maximum entropy theory of ecology. The American Naturalist 185:E70–E80.