Title: A network approach for inferring species associations from co-occurrence data

Authors:

Naia Morueta-Holme[1,2,*] (morueta-holme@berkeley.edu)

Benjamin Blonder[1,3] (bblonder@gmail.com)

Brody Sandel[1] (brody.sandel@bios.au.dk)

Brian J. McGill[4] (mail@brianmcgill.org)

Robert K. Peet[5] (peet@unc.edu)

Jeffrey E. Ott[5] (jeott@live.unc.edu)

Cyrille Violle[6] (cyrille.violle@cefe.cnrs.fr)

Brian J. Enquist[7] (benquist@email.arizona.edu)

Peter M. Jørgensen[8] (peter.jorgensen@mobot.org)

Jens-Christian Svenning[1] (svenning@bios.au.dk)

*Corresponding author.

NMH and BB contributed equally to this project.

Affiliations:

[1]Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark.

[2]Integrative Biology, University of California – Berkeley, CA, USA.

[3]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK.

[4]School of Biology and Ecology/Sustainability Solutions Initiative, University of Main, Orono, ME, USA.

[5]Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA.

[6]CEFE UMR 5175, CNRS - Université de Montpellier - Université Paul-Valéry Montpellier – EPHE -1919 route de Mende, F-34293 Montpellier, CEDEX 5, France.

[7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA.

[8]Missouri Botanical Garden, P.O. Box 299, St. Louis, MO 63166-0299, USA.

Abstract

Positive and negative associations between species are a key outcome of community assembly from regional species pools. These associations are difficult to detect and can be caused by a range of processes such as species interactions, local environmental constraints and dispersal. We integrate new ideas around species distribution modeling, covariance matrix estimation, and network analysis to provide an approach to inferring non-random species associations from local- and regional-scale occurrence data. Specifically, we provide a novel framework for identifying species associations that overcomes three challenges: 1) finding associations not driven by regional-scale distributions, 2) avoiding spurious associations caused by spatial autocorrelation in regional null models, and 3) correcting for indirect effects from other associated species. We highlight a range of research questions and analyses that this framework is able to address. The approach is statistically robust using simulated data. In addition, we present an empirical analysis of >1,000 North American tree communities that gives evidence for weak positive associations among small groups of species. Finally, we discuss several possible extensions for identifying drivers of associations, predicting community assembly, and better linking biogeography and community ecology.

Introduction

An unresolved question in ecology is how species assemblages come together at different spatial scales from regional species pools to form local communities (Weiher et al. 2011, HilleRisLambers et al. 2012). Some subsets of species may be consistently associated or dissociated in assemblages due to multiple processes including chance, species interactions, and the indirect effect of showing the same (or opposite) response to environmental conditions. While the effect of broad-scale environmental filtering and dispersal limitation can be assessed using niche modeling techniques (De Marco et al. 2008, Guisan and Rahbek 2011, Normand et al. 2011), the influence of species interactions and local environmental conditions on community assembly is more challenging to measure and incorporate into predictions (Kissling et al. 2012, Pottier et al. 2013, Wisz et al. 2013, Thuiller et al. 2013, Araújo and Rozenfeld 2014). If species interact with each other or share resource or scenopoetic requirements (Soberón 2007) not adequately described by broad-scale models, then stacking independent species distribution models to predict species assemblages (sensu Guisan and Rahbek 2011, Calabrese et al. 2014) will provide misleading predictions of fine-scale community assembly. Thus, a better understanding of species associations across scales could improve predictions of the dynamics of local community composition in changing environments.

The goal of this paper is to improve the tools needed to detect interspecific associations from co-occurrence data. We first briefly describe the development of co-occurrence methods and then draw from different lines of research to present a more complete and flexible general framework for inferring species associations that overcomes multiple challenges faced by previous approaches.

From experiments to co-occurrence methods

Efforts to infer species associations and their role in structuring communities have a long history. Traditionally, associations have been derived from small-scale field observations (e.g. MacArthur 1958, Bullock et al. 2000) or manipulative experiments (e.g. Stewart and Aldrich 1951, Connell 1961). Such methods may be suitable for studying associations among a few species at local scales. However, addressing whether small-scale associations occur consistently across large regions raises major practical issues.

Experimental data is time-consuming to obtain even at small scales (e.g. Bullock et al. 2000, Callaway et al. 2002b), so these approaches are unfeasible for large numbers of species because the number of possible interactions grows rapidly with assemblage size. For example, n=100 species have 4,950 possible pairwise interactions and 161,700 possible three-way interactions.

An alternative approach for detecting associations is to gather occurrence data from field observations and analyze species co-occurrence matrices by comparing the observed patterns to null models. Such approaches were widely used during discussions of the "checkerboard" paradigm for forbidden combinations of species on islands, with extensive debate over the need for null models to compare observed association patterns to random expectation (Diamond 1975, Connor and Simberloff 1979). The recent increase in availability of occurrence data has brought a renewed interest in using correlations and algorithms to infer species associations and the mechanisms behind them (e.g. Bruelheide 2000, Blick and Burns 2009, Blois et al. 2014). Since the 1970s, statistical approaches have been refined to simultaneously analyze co-occurrence patterns of multiple species pairs (e.g. Gotelli and Ulrich 2010). The main idea is to create a community matrix where rows are species, columns are sites and elements represent the observed presence/absence or abundance of each species at each site. The matrix is then compared to a set of randomized matrices in order to detect non-random co-occurrence patterns (Connor and Simberloff 1979, Gotelli and Ulrich 2010). Modern updates to null-model-based co-occurrence approaches can test the effects of environmental drivers, species interactions, or both in structuring communities; for approaches based on randomized null models see e.g. Gotelli et al. 2010, Ulrich et al. 2012; for approaches based on analytical null models see Araújo et al. 2011, Veech 2013.

Three missing ideas

One challenge that most previous co-occurrence approaches ignore is the potential effect of other species on a particular pairwise association, i.e. indirect effects (Brown et al. 2004, Harris 2015). If two competing species share a positive (or negative) relationship to a third species, their occurrences could be positively correlated, and thus an indirect effect (correlation) is inferred when the true effect (partial correlation) is actually negative (Brown et al. 2004, Schäfer and Strimmer 2005). A similar idea is implemented in ecology

for joint species distribution models (JSDMs; Pollock et al. 2014), where species associations are inferred after accounting for the environment. However, JSDMs usually do not resolve indirect effects from other species (but see the inversion approach of Harris 2015). The indirect association problem is well known in other fields, such as association detection in large genomic and cell-signaling datasets. One solution is to use Gaussian graphical models (Schäfer and Strimmer 2005) or modifications of them, which estimate the partial correlations between e.g. each gene pair after taking into account the remaining genes (e.g. Dobra et al. 2004, Friedman et al. 2008). The approach ensures estimation of conditional instead of joint associations and has recently been extended to ecological association networks (Harris 2015).

A second challenge is the need for more robust null models. Species are not distributed randomly across sites, but rather have regional geographic distributions that are constrained by climate and dispersal limitations. In the past, most null models have been defined by simply resampling the observed species × site community matrix **O** (e.g. Gotelli and Ulrich 2010, Borthagaray et al. 2014). However, this approach fails to account for the additional broad-scale constraints, introducing unrealistic null expectations of spatial independence across randomized sites that disregard spatial autocorrelation in species' distributions (Legendre 1993, Lennon 2000). In the framework we present here, we suggest that alternative null models can be defined to simulate a specific process. By incorporating regional structure that is contingent on e.g. climate-based species distribution models, we can thus gain more confidence that associations are the outcome of species interactions rather than shared environmental requirements. The integration of regional and local community occurrence data with modern network methods has the potential for improving predictive power in community ecology (see Baselga 2010, 2012 for a related approach).

A third missing idea is the incorporation of network theory. Network approaches have become common in the study of protein interactions, social structure, etc. (Newman 2010), but have not been applied widely to non-bipartite species association networks. The central idea is that any individual association between species can be better understood in the context of the network of associations between all other species. For example, species with many associations may be preferentially linked to those with few associations, suggesting a scenario where some species act as hubs or keystones. Thus, the network of

119 species associations is a useful way to visualize the community as well as quantify changes in species and
120 multi-species patterns via node- and network-level statistics to answer more sophisticated questions about
121 associations.

122 Node-level statistics provide insight into species' roles. For example, the unweighted degree of each
123 species characterizes the number of its association partners. The ratio between the number of positive and
124 negative associations per species can provide a measure of a species' role in the network (e.g. as an attractive
125 "aggregator", if a species has more positive than negative associations, or a repulsive "segregator" of other
126 species, if otherwise). Network-level statistics also provide insight into the overall structure of the species
127 assemblage. For instance, modularity gives a measure of the overall structure of the network, indicating the
128 amount of division of the network into clusters of nodes that are densely connected to each other, but
129 sparsely connected to nodes in other clusters (Newman 2010, Borthagaray et al. 2014). Higher modularity
130 indicates that groups of species are more likely to be mutually associated. Finally, overall non-random
131 numbers of links assigned to each node in the network can be tested by comparing the degree distribution to
132 a binomial distribution, which in the limit of many species is equivalent to a chi-square test for deviation
133 from a Poisson distribution.

134 The approach we suggest thus builds on recent attempts to infer species associations from
135 occurrence data and network theory (e.g. Borthagaray et al. 2014). Here we provide a more flexible
136 framework that includes both positive and negative associations, and has the ability to test for deviations
137 from a range of regional-scale null models. The framework is useful for systems in which associations are
138 possible across any pair of the study species (i.e. not bipartite networks such as pollinator-plant interactions,
139 e.g. Bascompte 2010). We currently implement the framework only for symmetric associations (i.e. (+/+),
140 (0,0) or (-,-)). For asymmetric associations like predation (+/-) or commensalism (+/0), see Discussion.

141 ***Interpreting associations***

142 The potential drivers of species associations can be better assessed using a network framework combined
143 with null modeling. This enables us to identify multi-species modules (i.e. non-random groups) of positively

144 or negatively associated species and assess the importance of particular species in shaping the modules,
145 ultimately improving predictions of community assembly. If we focus on plants, positive associations, or
146 aggregations of species, may be caused by biotic interactions such as facilitation between nurse trees and
147 seedlings, nitrogen fixation by certain species improving soil fertility for other species, or through shared
148 pollinators, seed dispersers, or facultative mutualisms with endophytic fungi (e.g., Afkhami et al. 2014).
149 Alternatively, positive associations may be indicative of shared local environmental requirements or effects
150 of stabilizing niche differences (Chesson 2000, Lasky et al. 2014), or reflect historical dispersal dynamics
151 such as the expansion from glacial refugia (Svenning and Skov 2007). Negative associations, or segregation
152 of species, may also be driven by biotic interactions through competition, or alternatively reflect different
153 local requirements (e.g. variation in microclimate or edaphic conditions).

154 Once we have inferred the patterns of species associations, we can proceed to determine which
155 hypothesized underlying drivers are most important (Box 2), with the aim of understanding the principles
156 that cause associations to occur. Functional trait predictors may help identify mechanisms. For instance, if
157 positive associations reflect complementary niches where each species occupies a different niche in relation
158 to resource use (Bolnick et al. 2011), then phylogenetic (Webb et al. 2002) or functional (Weiher et al. 2011)
159 distances should be largest between positively associated species (Violle et al. 2011). Likewise, by inferring
160 spatial patterns of community-weighted mean associations (Box 2) we can test whether associations are
161 influenced by climate gradients, and are e.g. more common in warmer areas (Brown et al. 1996, Schleuning
162 et al. 2012), or whether positive associations become more prevalent at higher elevations (Callaway et al.
163 2002a).

164 A framework for detecting associations

165 Drawing on the multiple lines of research described above, we here propose a general framework to uncover
166 species associations from co-occurrence patterns. The associations we identify are those not explained by
167 broad-scale climate gradients. The main idea of our approach is to pair broad- and local-scale co-occurrence
168 information with the Gaussian graphical model approach, spatially explicit regional null models (Guisan and

Zimmermann 2000, Gotelli and Ulrich 2010, Borthagaray et al. 2014) and network theory (Newman 2010). Our framework is implemented as the 'netassoc' R package.

We first show how the co-occurrence framework can infer a network of species associations by pairing regional occurrence data with local-scale assemblage data. Second, we perform a simulation analysis demonstrating that the framework has acceptable error rates in realistic use cases. Third, we demonstrate a range of network-based analyses that can describe associations and test hypothesized mechanisms. Lastly, we illustrate the framework with an application to the trees of eastern United States using a large dataset of local co-occurrences and regional occurrences.

In Box 1 we describe the main data inputs and methodological choices for the approach. Briefly, the first step is to identify non-random species associations for an observed and a null dataset. To do this, we compare the partial correlation coefficients inferred for observed local-scale data to those inferred for regional-scale null expectations calculated from independent data. We then reinterpret these co-occurrence effect sizes within a network framework.

Network construction

To compute the species association network we need (i) the observed co-occurrences, i.e. a set of observed presence-absence or abundance data for *n* species found in a random sample of *m* sites, and (ii) the expected co-occurrences for the same species and sites based on a null model. The null expectation can be for presence or abundance of the species. From (i) we generate **O**, the observed species × site community matrix. From (ii) we generate **N**, the expected abundance (or presence) patterns at each local site as predicted by a chosen regional species distribution model (iii). Both the **O** and **N** matrix will have *n* rows (species) and *m* columns (sites).

We first infer the association strength between species *i* and *j* as entries A_{ij} in the *n* x *n* matrix **A**. We calculate an *n* x *n* covariance matrix **Σ** for each of **O** and **N**. From the inverse of this covariance matrix we obtain standard partial correlations between species, i.e. as

193 (1) $$C_{ij}(\mathrm{M}) = \frac{-\Sigma^{-1}_{ij}(\mathrm{M})}{\sqrt{\Sigma^{-1}_{ii}(\mathrm{M}) \cdot \Sigma^{-1}_{jj}(\mathrm{M})}}$$

194 where **M** is the *n* x *m* input community matrix. C_{ij} represents the effect of species *i* on species *j* after
195 correcting for the effects of all other species; it is zero if species *i* is conditionally independent of species *j*.
196 Equation 1 represents the fundamental mathematical approach taken when constructing Gaussian graphical
197 models (Schäfer and Strimmer 2005) for inferring linear associations between random variables. We
198 calculate C(**O**) as well as for *K>>1* resamples of **N**, C(**Ñ**). This distribution simply simulates a weighted
199 lottery model of community assembly where species enter a community based only on their overall
200 abundance in the regional pool (i.e. the probabilities in **N**). To do so, the resamples preserve the total number
201 of individuals within each site, weighting the sampling by the expected abundance of each species based on
202 the original **N** matrix.

203 We then determine if the observed association between each species pair is positive or negative by
204 comparing the observed co-occurrence score to the distribution of expected co-occurrences across **Ñ**. We
205 calculate a standard effect size $SES_{ij}(\mathbf{O},\mathbf{N})$, i.e., by rescaling by the mean (μ) and standard deviation (σ) of the
206 null distributions:

207 (2) $$SES_{ij}(\boldsymbol{O}, \boldsymbol{N}) = \frac{C_{ij}(\boldsymbol{O}) - \mu[C_{ij}(\tilde{\boldsymbol{N}})]}{\sigma[C_{ij}(\tilde{\boldsymbol{N}})]}$$

208 Finally, we distinguish between significant and non-significant associations. We first calculate a two-tailed
209 *p*-value for species *i* and *j* as the number of replicates in which the absolute observed association strength is
210 smaller than the absolute null association strength divided by the total number of replicates. The next step is
211 to then correct for multiple comparisons by specifying a false discovery rate, α, and performing a Benjamini-
212 Hochberg correction (Benjamini and Hochberg 1995) on each p-value, producing a new set of p-values p_{ij}*.
213 The false discovery rate conceptualizes the type I error rate by controlling the expected proportion of false
214 positives, i.e. the proportion of incorrect rejections of the null hypotheses across multiple comparisons.

Together, this process yields a species-by-species adjacency matrix **A** with *n* rows and *n* columns (number of species):

$$A_{ij} = If \begin{cases} p_{ij}^{*} < a & SES_{ij}(\mathrm{O},\mathrm{N}) \\ p_{ij}^{*} \geq a & 0 \end{cases} \quad (3)$$

This matrix *A* is treated as the adjacency matrix (showing which species are connected to each other) used to define the species association network, such that a significantly positive or negative association between species *i* and *j* is established with strength $A_{i,j}$ if A_{ij} is nonzero. The network of associations is used for all subsequent analyses.

A few important decision points in the framework

While our statistical framework for describing species associations is general, the user must choose between multiple definitions and parameters specific to the system and taxa being studied. For example, one important choice in the analysis is the type of null model. The **N** matrix shows the predicted abundance or presence/absence at each local site for each species. Multiple methods can be used to define **N**. For example, a leave-one-out LOESS model on occurrence data can be used to calculate the expected abundance at one site from a distance interpolation of the observed abundances at all other sites. Such a model indicates the expected community produced by a dispersal-environment model. Alternatively, MaxEnt or other species distribution models can be used to calculate the expected abundance based on only broad-scale climate. Other approaches are possible, like calculating the expected occurrence from stacking species' regional geographic ranges sourced from expert-drawn range maps or from mechanistic regional models that predict abundance patterns across space (e.g. demographic or trait-based dispersal models; Jongejans et al. 2008).

A second set of choices that the user must define in the analysis is how to estimate the inverse covariance matrix $\mathbf{\Sigma}^{-1}$, which can be difficult in practice. Two situations can arise: first, the number of species can be much larger than the number of sites; second, most sites can contain very few species. Both cases can lead to $\mathbf{\Sigma}$ becoming singular (i.e. non-invertible) because of very large covariances between some

species pairs. A range of shrinkage estimators for $\mathbf{\Sigma}^{-1}$ have been developed that provide a robust approach to resolving this general problem (e.g., Hoerl and Kennard 1970, Schäfer and Strimmer 2005, Friedman et al. 2008). All the shrinkage estimators increase estimator bias in exchange for reduced mean squared error by introducing additional offset parameters that 'shrink' coefficient estimates and so force the existence of a matrix inverse. These offset parameters can be estimated by cross-validation approaches. A full survey of these methods is beyond the scope of this article, but some popular options include the James-Stein type shrinkage estimator (Schäfer and Strimmer 2005), the graphical lasso (L_1-regularization; Friedman et al. 2008), or ridge regression (L_2-regularization; Hoerl and Kennar 1970). We caution against using the graphical lasso because it produces sparse inverse covariance matrices that can produce singular null partial correlation distributions $C_{ij}(\tilde{\mathbf{N}})$, and instead recommend using the James-Stein estimator because of its good performance and low computational cost (Schäfer and Strimmer 2005).

Finally, we also recommend log-transforming abundance data as $f(x,a): x \mapsto \log\left(x+a\right) - \log\left(a\right)$ for some small number a, e.g. 10^{-6}. This transformation can improve normality of the distribution of abundance data, which can otherwise take either zero or very large values.

Testing the framework with a simulation analysis

To measure the expected performance of the network framework, we simulated co-occurrence matrices with known associations and determined how well network-detected associations matched these.

Consider a scenario involving n species distributed across m sites, of which a fraction h are unsuitable. We first generated an $m \times n$ expected species-by-site matrix $\mathbf{N}$, all of whose abundance entries were independently and identically distributed according to a hurdle model, such that N_{ij} was zero with probability h and Poisson-distributed (mean λ) with probability 1-h. If any of the marginal sums of $\mathbf{N}$ were zero (i.e. a site with no species or a species with no sites; problematic only for small n and m), we re-generated $\mathbf{N}$ until all marginal sums were non-zero. We then independently generated an $m \times n$ observed matrix $\mathbf{O}$ via the same hurdle process.

As a next step, we assumed that there were *Z* associations in the 'true' association network. We chose the *Z* associations by generating a random graph with *n* vertices and *Z* edges (Erdős and Rényi 1959), with weight w_z ($z \in \{1,2,...,Z\}$) set to either 1 or -1 with equal probability. To model this, we iterated over all associations *z*; for each pair of species i_z and j_z for which a true association exists, we chose a random fraction *f* of sites $\{m_z\}$; at each of these sites we either increased the abundance for both species (when $w_z>0$) or increased for one species and decreased for the other (when $w_z<0$) by a factor *s* proportional to the mean abundance of both species at these sites:

$$\begin{aligned} \boldsymbol{O}_{i_z,m_{\{z\}}} &= \boldsymbol{O}_{i_z,\{z\}} + w_z^{\,2} \cdot s \cdot \left(\boldsymbol{O}_{i_z,\{z\}} + \boldsymbol{O}_{j_z,\{z\}}\right)/2 \\ \boldsymbol{O}_{j_z,m_{\{z\}}} &= \boldsymbol{O}_{j_z,\{z\}} + w_z \cdot s \cdot \left(\boldsymbol{O}_{i_z,\{z\}} + \boldsymbol{O}_{j_z,\{z\}}\right)/2 \end{aligned} \quad (5)$$

This process effectively increased the covariance between species when the species were positively associated and decreased it when the species were negatively associated, with the parameters *h* and *s* controlling the strength of the association. We used *h*=0.2 and *s*=0.2 in this analysis.

Next, we applied our network framework using the matrices for all possible parameter combinations of *n*=10, 100; *m*=10, 100, 1,000, *f*=0, 0.5, and *Z*=10, 50, 100 (note that some combinations were not possible, e.g. when n=10, we cannot simulate values of Z>45 because they would exceed the number of interactions in a fully connected network). We used a James-Stein type shrinkage estimator (Schäfer and Strimmer 2005) with significant associations inferred at the α=0.05 significance level. We set the number of null replicates to 1,000 and repeated the entire analysis for each parameter combination 10 times.

In order to calculate error rates for our method, we compared the inferred networks' structure to the true network's structure. We counted a true positive association if it was detected for the correct pair of species and had the correct sign; as a true negative if it was not detected for a pair of species for which an association did not exist. A false positive association was counted if it was detected but was either the incorrect sign or the pair of species did not have a true association. Similarly, a false negative was counted if it was not detected but the pair of species did have a true association. These counts allowed us to calculate

the positive predictive value (PPV; true positives divided by true positives plus false positives) and the negative predictive value (NPV; true negatives divided by true negatives plus false negatives) as summary statistics.

To determine the sensitivity of the method to different parameters, we constructed a random forest regression model for NPV and PPV. The method generates an ensemble of regression trees and therefore allows for multi-way interactions between variables. We calculated the importance of each variable as the residual sum of squares caused by splitting on the variable of interest, averaged over all trees. Random forest models were built using the *randomForest* R package (Liaw and Wiener 2002).

Across all parameter combinations, PPV took a mean value of 12 ± 26 s.d. %, while NPV took a mean value of 87 ± 26 s.d. %. That is, the method was better at detecting the absence of associations than the presence of associations. A random forest model predicting PPV as a function of m, n, f, and Z explained 56% of the variation in the data and showed that m had the largest impact (14.6) on PPV, followed by Z (8.1), with smaller contributions of n (1.7) and f (1.3). Partial dependence plots indicated that larger values of each parameter led to higher values of PPV. A similar random forest model for NPV explained 53% of the variation in the data, and showed that m (11.1) and Z (8.7) had the largest impacts, with smaller contributions of n (1.5) and f (2.6). Partial dependence plots indicated that smaller values of each parameter led to higher values of NPV. Additionally, datasets with large fractions of zero-abundance records did not challenge the model's ability to infer associations.

Overall, this analysis indicates that the method trades off between successfully detecting true associations (high PPV) and successfully detecting the absence of false associations (high NPV). Better PPV occurs for large datasets, while better NPV occurs for small datasets. Increasing the value of the false discovery rate α can further control this tradeoff. We repeated analyses for α=0.5 (results not shown), which approximately doubled PPV and halved NPV in all cases. Thus, the method can yield acceptable performance in a wide range of realistic use cases.

Empirical test of the framework

After establishing the robustness of the framework through the simulations, we can apply it to real datasets and analyze the inferred associations. Here we present an illustration of the approach using communities from temperate forests in North America. In particular, we use the framework to test whether i) there are positive and/or negative associations among tree species that cannot be explained either from a regional dispersal-environmental model, or from broad-scale climate gradients; and whether ii) such associations can be explained by phylogenetic relatedness, functional trait similarity, or environmental gradients. We predict that across null models, association networks will have non-random structure and high modularity, and that positive (negative) associations will be more common among more (less) closely related or more similar species or in warmer (colder) environments.

Species data

We chose local community data to represent eastern North America. We extracted community-level tree species abundance data from the Forest Inventory and Analysis (FIA) database (Gray et al. 2012) (see Appendix S1 for query details). We selected 5,138 0.07 ha plots from the eastern USA (east of the 100° W meridian), and subsampled to a maximum of three plots for each 10,000 km^2 to reduce spatial sampling biases. Each plot consists of four 7.3 m radius subplots located 36.6 m from each other. All plots included were surveyed in the field between 2004 and 2008, used standard sampling protocols, and were marked as natural stands without evidence of artificial regeneration or human disturbance. If a plot was surveyed more than once in the time period chosen, we only included the newest survey. We excluded FIA taxa that were not identified to the species level. A total of m=1,009 plots out of the 5,138 plots and n=137 tree species were included in the analyses.

Point occurrence data for the null models came from BIEN, the Botanical Information and Ecology Network (Enquist et al. 2009, http://bien.nceas.ucsb.edu/bien/) for each of the 137 tree species.

Alternative species distribution models

We computed results based on three definitions of the null model. First, we used the commonly used random-swap algorithm, where the local community matrix **O** is randomized to create the null matrix **N**

(e.g., Connor and Simberloff 1979, Gotelli and Ulrich 2010), keeping row and column sums fixed (i.e. total species and site abundances). Second, we used a leave-one-out LOESS regression of plot abundances with a span parameter of 0.2 (mirroring ter Steege et al. 2013) to compute the expected abundance at each plot from its spatial position and those of the remaining 5,137 plots in the full FIA dataset. Third, we also used species distribution models created with the algorithm MaxEnt (Phillips and Dudík 2008) and based on the BIEN point occurrence data to estimate the climatic potential range of each species. As a test case, we used 19 bioclimatic layers representing "current" climate (average 1950–2000 conditions) as model predictors, extracted from WorldClim 1.4 at 30-arc second resolution (Hijmans et al. 2005). Each model was fit using default parameters to both North and South America to capture the climatic range of the full New World distribution of each species. We converted the model suitability scores to expected abundance values by standardizing them so that the summed suitability scores for each species equaled the total number of individuals across all plots. This procedure assumes a linear relationship between suitability and abundance.

We used 1,000 resamples $\tilde{\mathbf{N}}$ of the expected species × site matrix **N**. As in the simulation analysis, we used a James-Stein shrinkage estimator for the inverse covariance matrix, and specified an overall false discovery rate of $\alpha=0.05$ to exclude non-significant associations. Modules in the network were inferred using a standard fast-greedy algorithm (Clauset et al. 2004).

Potential predictors of species' associations

We obtained data on phylogenetic relatedness between species and functional trait values to illustrate how the network statistics can be used to test hypotheses of drivers of species associations.

We calculated phylogenetic distances from a phylogeny for all the trees of eastern USA using Phylocom's 'phylomatic' tool (Webb et al. 2008). We used the R20120829 backbone tree, with branch lengths adjusted by fossil constraints (Gastauer and Meira-Neto 2013). We then computed the distance between each pair of species.

We also obtained measurements of four traits thought to underlie major axes of ecological strategy variation (Westoby et al. 2002): maximum height (m), specific leaf area (SLA; cm^2/g), seed mass (g) and

wood density (g/cm^3). Functional trait data were extracted from the BIEN database. There was good coverage for trait data: 99% for height, 72% for SLA, 93% for seed mass, and 81% for wood density. We log_{10}-transformed each trait value to reduce skewness, then rescaled all values by subtracting means and dividing by standard deviations. We then computed trait differences between all pairs of species.

We used linear regression to determine whether the links between each species pair (link strength if adjacent; 0 if not adjacent) was predicted by pairwise phylogenetic or trait distance between the species pair. We did not correct for non-independence of predictor distances (e.g. Mantel-type test). This approach should lead to an increased rate of falsely rejecting the null hypothesis, meaning that failing to reject the null hypothesis is more likely to reflect a true absence of relationship.

Network structure of trees of eastern USA

The species association network based on the random-swap algorithm showed a non-random structure (chi-square test for Poisson degree distribution, $p < 10^{-45}$), identifying many positive and negative associations (Table 1). On the other hand, the overall structure of networks based on the two regional null models – the LOESS dispersal-environment model, and the MaxEnt climate model – was not different from random. This result indicates that the overall distribution of local associations between trees in North America can mainly be explained by broad-scale drivers. Looking at the individual associations, we only found a few positive interactions in both of these networks. Interestingly, all five modules of 2-3 associations identified using the LOESS regional model also appeared when applying the MaxEnt model, although sometimes with an additional species in the module (Fig. S1). We found additional associations deviating from the MaxEnt regional model, and modules were larger in general (Table 1, Fig. 1). In the random-swap network most species were inferred to be 'segregators', while in the MaxEnt and LOESS network most species were inferred to be 'aggregators' (Fig. S2).

What predicts species' associations?

The positive and negative associations we found were not predicted by phylogenetic or trait distances (Fig. S3) regardless of null model. The network structure explained by phylogeny in any network was no more than 0.05%, and the variation explained by all four traits together was no more than 0.12% in any network.

We did find weak spatial gradients in abundance-weighted mean values of degree for the species comprising each community (Fig. S4). For each of the random-swap and MaxEnt networks, mean degree was negatively correlated with mean annual temperature but not with mean annual precipitation (multiple regression; both $p < 0.002$, both $R^2 < 0.06$). The LOESS network was not correlated with either mean annual temperature or precipitation ($p=0.26$).

Can we infer non-random species associations?

Our simulation analysis showed that the co-occurrence framework is indeed able to identify known associations within the parameter regimes we explored. However, there is a trade-off, where large datasets lead to better success in detecting true associations, and smaller datasets are better are identifying the absence of false associations. When including explicit regional null models, the framework is able to pinpoint which species are associated after broad-scale drivers have been accounted for. This not only allows testing for multiple regional models. Another advantage is that spatial autocorrelation in species distributions can be taken into consideration. Indeed, when looking at the results from our case study, it is apparent that the random-swap algorithm identifies spurious positive and negative associations that stem from ignoring the spatial dependence of species occurrences. This important implication shows that the null model from a hypothesized process such as dispersal or broad-scale climate patterns gives better control of the type of associations identified.

Another challenge addressed by the framework is that of indirect effects from multiple species on pairwise associations. The result is that when applying the framework to tree communities of eastern North America, we only found a small number of positive associations deviating from either the LOESS or the MaxEnt based regional models after correcting for indirect effects (Fig. 1). Ecologically, this may reflect that tree species distributions are largely controlled by environment and dispersal, with little importance of

interspecific interactions, even if the latter matter for local abundances. The associations we do identify are those deviating from the broad-scale expectations. The LOESS model can be interpreted as a dispersal and environment model, since it simply interpolates abundance as a function of distance and implicitly includes environmental conditions that are spatially autocorrelated. Alternatively, the MaxEnt model computes the expected co-occurrence as a function of broad-scale climate variables, ignoring dispersal constraints, and instead representing the potential climate range of each species. It is thus not entirely surprising that the associations identified with the dispersal-environment model are all a subset of those identified with the climate-only model. Contrary to our predictions, functional traits, phylogenetic relatedness and environmental gradients did not correlate with the associations found across the networks. However, natural history, successional dynamics and missed environmental drivers could explain at least some of the associations identified. Indeed, local habitat requirements not fully captured by the broad scale environmental gradients tested here seem to explain several associations. For instance, *Carya aquatica* and *Taxodium distichum* are both species of Coastal Plain, strongly associated with large river backswamps that are periodically flooded. A similar habitat is preferred by *Nyssa aquatica* and *Planera aquatica*, although this species pair is found more up-stream than the first ones, in areas where flooding is less prolonged. *Quercus palustris*, *Q. bicolor* and *Carya laciniosa* also prefer swamp habitats, though more from interior flatlands on more calcareous soils. Habitat preference does not seem to explain the association between *Carya pallida* and *Quercus michauxii*, which prefer dry/sandy and swampy soils, respectively. These two species rarely occur together, but both are largely confined to the southeastern Coastal Plain and lower Piedmont regions. This geographic signal could instead be driven by dispersal limitation. A few boreal modules such as the one around *Abies balsamea* are identified with the MaxEnt regional model but disappear in the LOESS model. These modules may represent local mosaics of boreal and temperate stands at the boreal-temperate transition zone, reflecting local environmental conditions and/or priority effects (Pastor and Mladenoff 1992).

Given that the MaxEnt model only covers environmental constraints, species might be simulated to co-occur that have similar climatic requirements but are allopatric. In such a case we would expect more

434 negatively associated species in the network derived from the MaxEnt null model. Instead, the modularity of
435 networks based on the MaxEnt models was much higher than that of LOESS-based networks, and in neither
436 case did we find negative associations. The fact that we only identify positively associated groups of species
437 unexplained by broad-scale climate conditions is consistent with the results of the simulation study of Araújo
438 and Rozenfeld (2014), who found that the effect of positive (but not negative) dependencies between species
439 scaled up to biogeographical scales and should be accounted for in range models under climate change.

440 One implication of our results – in particular the larger amount of positive associations deviating
441 from the climate-only model – is that we cannot rely solely on stacked species distribution models (SDMs)
442 (sensu Guisan and Rahbek 2011, Calabrese et al. 2014) to predict the composition of local communities, but
443 need to take species associations into account (Araújo and Rozenfeld 2014) – whether driven by biotic
444 interactions, dispersal, or local environmental filtering. In cases where e.g. local environmental or dispersal
445 data are available, these can be integrated to filter broad-scale predictions and improve the performance of
446 SDMs for communities (Boulangeat et al. 2012).

447 We note that even low error rates can translate into high absolute numbers of incorrect associations
448 for datasets with large species numbers. Given that 100 species have 4,950 potential pairwise interactions,
449 our simulation error rates mean that, on average, approximately 25 links will be inferred that do not actually
450 exist. Our Type 2 error rates are very high for small datasets, but settle to near zero for datasets with at least
451 100 sites, with performance further increasing when species co-occurrence patterns have high association
452 weights. Thus, the algorithm does not often miss real associations, but even then, our Type 2 error rates
453 translate into approximately 500 real links that are missed. These rates and uncertainties are similar to other
454 association-inference approaches (Morales-Castilla et al. 2015) and suggest that these frameworks are most
455 useful for reducing the space of possible associations to significantly more manageable numbers (cf. Figure 2
456 in Morales-Castilla et al. 2015).

457 ***Inferring drivers of associations — a complex challenge***

458 In our example with North American tree communities, the associations we identified can best be explained
459 by ecological (and possibly geographic) groupings. It is thus not surprising that they could not be predicted
460 by markers like functional trait similarity and phylogenetic distance. Similarly, we only found a weak
461 negative correlation between community mean degree and mean annual temperature for the associations
462 identified with the MaxEnt null model, possibly reflecting the geographic signal of boreal communities. The
463 application still shows how network metrics can be used to test for drivers of associations, although in this
464 particular case, more direct measures such as species' flooding tolerance from Ellenberg values would have
465 been more useful.

466 Still, we can imagine several issues that could pose obstacles to the prediction of associations. One
467 issue is that multiple competing processes can be at play. Competition and facilitation may cancel each other
468 out (Callaway and Pennings 2000). Species associations may also vary across environmental gradients
469 (Callaway et al. 2002a, Pottier et al. 2013) or across temporal scales (Martorell and Freckleton 2014, Blois et
470 al. 2014), such that the scale of input data used would be critical. Even in networks of biotic interactions
471 between plants and pollinators, pairwise associations do not always correlate with hypothesized drivers, even
472 when correlations are found with metrics of overall network structure (Olito and Fox 2015). We therefore
473 expect that disentangling the processes driving patterns of co-occurrence will remain an ongoing challenge.

474 ***Extensions of the network framework***

475 The increased availability of regional species occurrence and local-scale co-occurrence data across
476 large extents has been a strong driver behind recent attempts to disentangle the processes promoting species'
477 associations through biotic interactions, dispersal limitations and environmental filtering (e.g. Blick and
478 Burns 2009, Ulrich et al. 2012, Blois et al. 2014). We have presented a flexible framework to infer species
479 positive and negative associations that deviate from expected broad-scale processes, and illustrated how
480 network statistics can be used to test hypothesized drivers of associations. The approach can be readily
481 applied to other datasets and systems for any other type of potential species associations across trophic
482 levels.

There is an ongoing debate on the directionality of interactions, with some studies finding more positive than negative associations (Blick and Burns 2009 and references therein, Blois et al. 2014), while others (including a meta-analysis) have found the opposite (Azaele et al. 2010, Gotelli and Ulrich 2010). Inconsistent associations appear to be the norm rather than the exception in the literature, changing when using species-based rather than individual-based models (Blick and Burns 2009), or when looking for patterns across time in the paleo-record (Blois et al. 2014). Some of the disparate results seen across studies might be the unintended outcome of methodological differences. We suggest that sensitivity analyses and consensus approaches should be used to achieve robust results. Importantly, partial correlations should be widely implemented, as well as taking into account spatial autocorrelation as part of the null model to avoid identifying non-existent interactions.

The framework as currently implemented only accounts for symmetric associations, e.g. those representing (+/+), (0,0) or (-,-) interactions. This limitation is imposed because the matrix **A** is derived from **C**, which is derived from $\mathbf{\Sigma}^{-1}$, which is symmetric. The coefficients in **C** could instead be calculated using other approaches that allow for non-symmetric outcomes, and so also capture effects like predation (+/-) and commensalism (+/0). The network framework allows for such directed linkages, and the R package does allow arbitrary user-specified functions to be used for calculating **C**. However, we are unaware of viable candidate functions to use. Most association metrics are symmetric (cf. Janson and Vegelius 1981) and the few that are not (e.g. Somers' D (Somers 1962), Goodman and Kruskal's lambda (Goodman and Kruskal 1972)) cannot be used with abundance data, do not take both positive and negative values, and do not account for indirect associations. Aside from the excess co-occurrence approach of Araújo and Rozenfeld (2011), the Markov network approach suggested by Harris (2015), and the directed partial correlation coefficients proposed for microarray studies (Yuan et al. 2011), we are unaware of other approaches. Developing association metrics that capture all possible ecological integrations types should be a priority.

For cases in which many strong associations are identified, the network framework could potentially be extended to provide predictions of local community composition, which in turn can be tested in new communities. For instance, with information on the identity of only one species in a community, a "network-

crawling" approach could be used to predict the identity of the remaining species by allowing species that are close to the one known species in the network are more likely to be predicted in the local community (or less likely for species negatively associated). Such an extension would provide transparent and powerful tests of the predictive ability of association networks derived either solely from occurrence data, or combined with experimental or field-based interaction data.

More sophisticated analyses could be applied to our framework to provide a more thorough test of the drivers of associations observed. Indeed, a recent study found modularity analyses useful to identify biological attributes driving the connection of modules of co-occurring species (Borthagaray et al. 2014). Such additional analyses can be readily applied with our framework, potentially combined with other methods to identify network modules (Leger et al. 2015) to e.g. investigate the relationship of traits and phylogenetic relationships for species within each module, and thus explore the definitions of the scale of study.

Looking forward

Mechanistic understandings of the drivers of species distributions across scales are needed to better predict the consequences of rapid global change. The general framework we propose here provides a novel tool to infer local-scale associations that can affect species' distributions. We have presented a flexible framework linking co-occurrence and null-models to network theory for inferring species associations that is easily applicable to other systems and datasets. This approach can resolve challenges related to assumptions of spatial independence in species' distributions and indirect effects of multiple species on associations. Variations in the implementation of our approach can be used to test for association patterns that do not follow the expectation from broad-scale climate, dispersal or other hypothesized processes. With network metrics and analyses we can test for drivers of the patterns, and the approach can be potentially extended to predictions of local community assembly. Combined with field experiments and other methods, our framework can be a powerful tool to move beyond extant concepts in the analysis of co-occurrence data to improve assembly predictions across scales and help merge community ecology and biogeography.

Acknowledgements – This study was conducted as a part of the BIEN Working Group (Principal Investigators: Brian J. Enquist, Brad Boyle, Richard Condit, Steven Dolins, Robert K. Peet, and Barbara M. Thiers) supported by the National Centre for Ecological Analysis and Synthesis, a center funded by the National Science Foundation (NSF Grant EF-0553768), the University of California, Santa Barbara, and the State of California. The BIEN Working Group was also supported by The iPlant Collaborative (NSF Grant DBI-0735191). We thank all the contributors for the invaluable data provided to the BIEN (http://bien.nceas.ucsb.edu/bien/people/data-contributors/). We are grateful to Irena Šímová for preparing the trait data and to Brad Boyle for guidance on SQL queries. We thank D. Harris and two anonymous reviewers for constructive comments that improved this manuscript, in particular D. Harris for pointing us towards the use of partial correlation methods. We further acknowledge support to NMH by an EliteForsk Award, the Aarhus University Research Foundation, and the Villum Foundation. B. Blonder and JCS acknowledge financial support from Centre for Informatics Research on Complexity in Ecology (CIRCE), funded by the Aarhus University Research Foundation under the AU Ideas, the European Research Council (ERC Starting Grant #310886 'HISTFUNC', and the Danish Council for Independent Research – Natural Sciences (grant #12-125079). B. Blonder was also supported by a United States National Science Foundation graduate research fellowship. CV was supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Program (DiversiTraits project, no. 221060).

552 # References

553 Afkhami, M. E. et al. 2014. Mutualist-mediated effects on species' range limits across large geographic
554 scales. - Ecol. Lett. 17: 1265–1273.

555 Araújo, M. B. and Rozenfeld, A. 2014. The geographic scaling of biotic interactions. - Ecography 37: 1–10.

556 Araújo, M. B. et al. 2011. Using species co-occurrence networks to assess the impacts of climate change. -
557 Ecography 34: 897–908.

558 Azaele, S. et al. 2010. Inferring plant ecosystem organization from species occurrences. - J. Theor. Biol. 262:
559 323–329.

560 Bascompte, J. 2010. Structure and dynamics of ecological networks. - Science 329: 765–766.

561 Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. - Glob. Ecol.
562 Biogeogr. 19: 134–143.

563 Baselga, A. 2012. The relationship between species replacement, dissimilarity derived from nestedness, and
564 nestedness. - Glob. Ecol. Biogeogr. 21: 1223–1232.

565 Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful
566 approach to multiple testing. - J. R. Stat. Soc. Ser. B 57: 289–300.

567 Blick, R. and Burns, K. C. 2009. Network properties of arboreal plants: Are epiphytes, mistletoes and lianas
568 structured similarly? - Perspect. Plant Ecol. Evol. Syst. 11: 41–52.

569 Blois, J. L. et al. 2014. A framework for evaluating the influence of climate, dispersal limitation, and biotic
570 interactions using fossil pollen associations across the late Quaternary. - Ecography 37: 1095–1108.

571 Bolnick, D. I. et al. 2011. Why intraspecific trait variation matters in community ecology. - Trends Ecol.
572 Evol. 26: 183–192.

573 Borthagaray, A. I. et al. 2014. Inferring species roles in metacommunity structure from species co-
574 occurrence networks. - Proc. Biol. Sci. 281: 20141425.

575 Boulangeat, I. et al. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of
576 species distributions and their abundances. - Ecol. Lett. 15: 584–593.

577 Brown, J. H. et al. 1996. The geographic range: size, shape, boundaries, and internal structure. - Annu. Rev.
578 Ecol. Syst. 27: 597–623.

579 Brown, J. H. et al. 2004. Constraints on negative relationships. - In: Taper, M. L. and Lele, S. R. (eds), The
580 nature of scientific evidence: Statistical, philosophical, and empirical considerations. University of
581 Chicago Press, pp. 298–324.

582 Bruelheide, H. 2000. A new measure of fidelity and its application to defining species groups. - J. Veg. Sci.
583 11: 167–178.

584 Bullock, J. M. et al. 2000. Geographical separation of two *Ulex* species at three spatial scales: does
585 competition limit species' ranges? - Ecography 23: 257–271.

586 Calabrese, J. M. et al. 2014. Stacking species distribution models and adjusting bias by linking them to
587 macroecological models. - Glob. Ecol. Biogeogr. 23: 99–112.

588 Callaway, R. M. and Pennings, S. C. 2000. Facilitation may buffer competitive effects: Indirect and diffuse
589 interactions among salt marsh plants. - Am. Nat. 156: 416–424.

590 Callaway, R. M. et al. 2002a. Positive interactions among alpine plants increase with stress. - Nature 417:
591 844–848.

592 Callaway, R. M. et al. 2002b. Epiphyte host preferences and host traits: mechanisms for species-specific
593 interactions. - Oecologia 132: 221–230.

594 Chesson, P. 2000. Mechanisms of maintenance of species diversity. - Annu. Rev. Ecol. Syst. 31: 343–366.

595 Clauset, A. et al. 2004. Finding community structure in very large networks. - Phys. Rev. E 70: 066111.

596 Connor, E. F. and Simberloff, D. 1979. The assembly of species communities: chance or competition? -
597 Ecology 60: 1132–1140.

598 De Marco, P. et al. 2008. Spatial analysis improves species distribution modelling during range expansion. -
599 Biol. Lett. 4: 577–580.

600 Diamond, J. M. 1975. Assembly of species communities. - In: Cody, M. L. and Diamond, J. M. (eds),
601 Harvard University Press, pp. 342–444.

602 Dobra, A. et al. 2004. Sparse graphical models for exploring gene expression data. - J. Multivar. Anal. 90:
603 196–212.

604 Enquist, B. J. et al. 2009. The Botanical Information and Ecology Network (BIEN): Cyberinfrastructure for
605 an integrated botanical information network to investigate the ecological impacts of global climate change
606 on plant biodiversity. - iPlant Collaborative, www.iplantcollaborative.org

607 Erdős, P. and Rényi, A. 1959. On Random Graphs I. - Publ. Math. 6: 290 – 297.

608 Friedman, J. et al. 2008. Sparse inverse covariance estimation with the graphical lasso. - Biostatistics 9: 432–
609 441.

610 Gastauer, M. and Meira-Neto, J. A. A. 2013. Avoiding inaccuracies in tree calibration and phylogenetic
611 community analysis using Phylocom 4.2. - Ecol. Inform. 15: 85–90.

612 Goodman, L. A. and Kruskal, W. H. 1972. Measures of association for cross classifications, IV:
613 simplification of asymptotic variances. - J. Am. Stat. Assoc. 67: 415–421.

614 Gotelli, N. J. and Ulrich, W. 2010. The empirical Bayes approach as a tool to identify non-random species
615 associations. - Oecologia 162: 463–477.

Gotelli, N. J. et al. 2010. Macroecological signals of species interactions in the Danish avifauna. - Proc. Natl. Acad. Sci. 107: 5030–5035.

Gray, A. et al. 2012. Forest Inventory and Analysis Database of the United States of America (FIA) (J Dengler, J Oldeland, F Jansen, C M, E J, M Finckh, G F, G Lopez-Gonzalez, RK Peet, and JHJ Schaminee, Eds.). - Biodivers. Ecol. 4: 225–231.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. - Ecol. Modell. 135: 147–186.

Guisan, A. and Rahbek, C. 2011. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. - J. Biogeogr. 38: 1433–1444.

Gutiérrez, E. E. et al. 2014. Can biotic interactions cause allopatry? Niche models, competition, and distributions of South American mouse opossums. - Ecography 37: 741–753.

Harris, D. J. 2015. Estimating species interactions from observational data with Markov networks. - bioRxiv dx.doi.org/10.1101/018861.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. - Int. J. Climatol. 25: 1965–1978.

HilleRisLambers, J. et al. 2012. Rethinking community assembly through the lens of coexistence theory. - Annu. Rev. Ecol. Evol. Syst. 43: 227–248.

Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: biased estimation for nonorthogonal problems. - Technometrics 12: 55–67.

Janson, S. and Vegelius, J. 1981. Measures of ecological association. - Oecologia 49: 371–376.

Jongejans, E. et al. 2008. Dispersal, demography and spatial population models for conservation and control management. - Perspect. Plant Ecol. Evol. Syst. 9: 153–170.

639 Kissling, W. D. et al. 2012. Towards novel approaches to modelling biotic interactions in multispecies
640 assemblages at large spatial extents. - J. Biogeogr. 39: 2163–2178.

641 Lasky, J. R. et al. 2014. Trait-mediated assembly processes predict successional changes in community
642 diversity of tropical forests. - Proc. Natl. Acad. Sci. 111: 5616–5621.

643 Lennon, J. J. 2000. Red-shifts and red herrings in geographical ecology. - Ecography 23: 101–113.

644 Leger, J.-B. et al. 2015. Clustering methods differ in their ability to detect patterns in ecological networks (T
645 Münkemüller, Ed.). - Methods Ecol. Evol. 6: 474–481.

646 Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? - Ecology 74: 1659–1673.

647 Liaw, A. and Wiener, M. 2002. Classification and regression by randomForest. - R News 2: 18 – 22.

648 MacArthur, R. H. 1958. Population Ecology of Some Warblers of Northeastern Coniferous Forests. -
649 Ecology 39: 599.

650 Martorell, C. and Freckleton, R. P. 2014. Testing the roles of competition, facilitation and stochasticity on
651 community structure in a species-rich assemblage (R Brooker, Ed.). - J. Ecol. 102: 74–85.

652 Morales-Castilla, I. et al. 2015. Inferring biotic interactions from proxies. - Trends Ecol. Evol. 30: 347–356.

653 Newman, M. 2010. Networks: an introduction. - Oxford University Press.

654 Normand, S. et al. 2011. Postglacial migration supplements climate in determining plant species ranges in
655 Europe. - Proc. R. Soc. B Biol. Sci. 278: 3644–3653.

656 Olito, C. and Fox, J. W. 2015. Species traits and abundances predict metrics of plant-pollinator network
657 structure, but not pairwise interactions. - Oikos 124: 428–436.

658 Pastor, J. and Mladenoff, D. J. 1992. The southern boreal-northern hardwood forest border. - In: Shugart, H.
659 H. et al. (eds), A systems analysis of the global boreal forest. Cambridge University Press, pp. 216–240.

660 Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a
661 comprehensive evaluation. - Ecography 31: 161–175.

662 Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint
663 Species Distribution Model (JSDM). - Methods Ecol. Evol. 5: 397–406.

664 Pottier, J. et al. 2013. The accuracy of plant assemblage prediction from species distribution models varies
665 along environmental gradients (R Field, Ed.). - Glob. Ecol. Biogeogr. 22: 52–63.

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

668 Schleuning, M. et al. 2012. Specialization of mutualistic interaction networks decreases toward tropical
669 latitudes. - Curr. Biol. 22: 1925–1931.

670 Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. - Ecol. Lett. 10:
671 1115–1123.

672 Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. - Am. Sociol. Rev. 27:
673 799–811.

674 Stewart, R. E. and Aldrich, J. W. 1951. Removal and Repopulation of Breeding Birds in a Spruce-Fir Forest
675 Community. - Auk 68: 471–482.

676 Svenning, J.-C. and Skov, F. 2007. Could the tree diversity pattern in Europe be generated by postglacial
677 dispersal limitation? - Ecol. Lett. 10: 453–460.

678 ter Steege, H. et al. 2013. Hyperdominance in the Amazonian tree flora. - Science 342: 1243092.

679 Thuiller, W. et al. 2013. A road map for integrating eco-evolutionary processes into biodiversity models. -
680 Ecol. Lett. 16 Suppl 1: 94–105.

681 Ulrich, W. et al. 2012. Null model tests for niche conservatism, phylogenetic assortment and habitat filtering.
682 - Methods Ecol. Evol. 3: 930–939.

683 Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence (P Peres-Neto, Ed.). - Glob.
684 Ecol. Biogeogr. 22: 252–260.

685 Violle, C. et al. 2011. Phylogenetic limiting similarity and competitive exclusion. - Ecol. Lett. 14: 782–787.

686 Webb, C. O. et al. 2002. Phylogenies and community ecology. - Annu. Rev. Ecol. Syst. 33: 475–505.

687 Webb, C. O. et al. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait
688 evolution. - Bioinformatics 24: 2098–2100.

689 Weiher, E. et al. 2011. Advances, challenges and a developing synthesis of ecological community assembly
690 theory. - Philos. Trans. R. Soc. Lond. B. Biol. Sci. 366: 2403–2413.

691 Westoby, M. et al. 2002. Plant ecological strategies: Some leading dimensions of variation between species.
692 - Annu. Rev. Ecol. Syst. 33: 125–159.

693 Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of
694 species: implications for species distribution modelling. - Biol. Rev. Camb. Philos. Soc. 88: 15–30.

695 Yuan, Y. et al. 2011. Directed partial correlation: inferring large-scale gene regulatory network through
696 induced topology disruptions. - PLoS One 6: e16835.

Tables and Figures

Table 1. Descriptive statistics for association networks constructed from the same dataset using different null models. We report an overall p-value for deviations from a Poisson distribution of edges, and for the subsets of the network for positive or negative associations, the mean ± standard deviation of degree, as well as the number of non-singleton modules (i.e. with more than one member) detected in the network.

Regional model	P-value	Degree		Positive modules		Negative modules	
		Positive	Negative	Number	Mean size	Number	Mean size
Swap	2.44E-46	20.31±10.81	56.26±15.42	12	11.4	5	27.4
LOESS	0.230	0.09±0.31	0±0	5	2.2	0	NA
MaxEnt	0.565	0.82±0.89	0±0	24	3.2	0	NA

Figure captions

Box 1. Concepts and data flow underlying the framework. Each step illustrates how the network of non-random species association is derived from the observed co-occurrence matrix and the expectation based on a chosen null model. See detailed description in main text.

Box 2. Examples of possible analyses to explore species associations based on the network. Network-level analyses (left) give information on the overall structure of the network. Examples include 1) testing for overall-nonrandom number of links comparing the degree distribution to e.g. a binomial distribution, and 2) quantifying modularity, i.e. groups of species more likely to be associated. Node-level analyses (right) serve to 1) identify the roles of individual species (e.g. as aggregators or segregators by looking at the proportion of negative and positive links per species), 2) quantify the centrality of each species, or 3) test for correlations of node metrics such as network distance and hypothesized drivers such as phylogenetic or functional trait distance. Finally, a hybrid of network and node-level measures can be used for spatial

analyses (bottom) to test for trends in e.g. mean degree of communities across space and their correlation to climate gradients.

Figure 1. Empirical association network for North American trees. The network was constructed using the MaxEnt regional null model and a shrinkage inverse covariance estimator using 1,000 null replicates and a false discovery rate of α=0.05. Gray envelopes indicate distinct modules. Positive associations are shown as blue lines; negative would be shown as red lines but none were found.

Figure S1. Association networks constructed from the same data, using the random-swap algorithm, the LOESS, and the MaxEnt regional null models, respectively. Left panel, positive associations (blue); right panel, negative associations (red). Gray envelopes indicate distinct modules.

Figure S2. Ratio of positive and negative associations across networks. The role of a species in the network can be indicated by the degree of each node (i.e. number of associations) for the positive subset of the network relative to the degree of each node for the negative subset of the network. Species below the 1:1 line (purple) are 'aggregators' because they have proportionately more positive associations, while those above the line are 'segregators'. This designation shifts between methods, as does the identity of the most highly associated species. The five species with highest degree are labelled in each panel.

Figure S3. Trait and phylogenetic predictors of network structure. Points indicate pairwise distances between all pairs of species. Relationships are shown for the random-swap network (pink), LOESS network (orange), and MaxEnt network (purple).

741

742 **Figure S4.** Distribution of network statistics across space. Each point indicates the abundance-

743 weighted mean network degree for a different community.