The macroevolution of bipartite networks

Timothée Poisot *1,2,3* Daniel B. Stouffer *1*

1: Centre for Integrative Ecology, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand; **2**: Université de Montréal, Département de Sciences Biologiques; **3**: Québec Centre for Biodiversity Sciences

TODO

Keywords: ecological networks - macroecoevolution - approximate Bayesian computation

Correspondence to Timothée Poisot – timothee.poisot@umontreal.ca — Latest update on November 23, 2017

The extant structure and distribution of biodiversity is the outcome of macro-evolutionary processes, and the modelling of these processes has stimulated a large variety of approaches (**???**, **???**). At their core, these approaches are all essentially birth-death processes, in that they model the rate of speciation and extinction to generate a prediction about both the temporal dynamics of species richness and its predicted current state. Surprisingly, these models tend to consider species as isolated entities; even though they share ancestry, they are not explicitly linked via inter-specific interactions. This fact is problematic from both an ecological (**???**) and evolutionary (**???**, **???**, **???**, **???**) standpoint since it is widely accepted that interactions serve as an essential scaffold for biodiversity and its emergent properties such as community persistence or ecosystem function (**???**). After all, predators invariably require prey, hosts require parasites, flowering plants require pollinators, and so on.

Although modern macro-ecological models give an increasingly central role to interactions (**???**), such models are still unable to predict the structure of complex interacting communities (**???**). Nevertheless, there are two key observations upon which solutions to overcome this limitation can be devised. First, extant networks are decidedly non-random with regard to their structure, and their structure is equally non-random with regards to macro-evolutionary processes (**???**). Second, the structure of ecological networks is dynamic over evolutionary timescales (**???**). Both these points are strongly suggestive of perpetual and ongoing action of macro-evolutionary processes. It stands to reason then that models of macro-evolution with explicit consideration of species interactions will therefore provide an appropriate theoretical framework to understand how networks evolve. Notably, such a framework enables the estimation of how much of extant network structure originated through macro-evolution, as opposed to reflecting extant opportunities and constraints (**???**).

If one assumes that the conservatism of interactions across phylogenies can be explained by the fact that an incipient species inherits its ancestor's interactions upon speciation (**???**, **???**), even a simple model with relatively few parameters can describe the possible evolutionary rules that shape a community's interaction network. Ideally, the parameters of any model such as this—no matter how simple or complex—ought to be calibrated against real-world evolutionary dynamics, similar to how the fossil and molecular record has been used to study species diversification (**???**). Unfortunately, the dearth of well-resolved, long-term time series of species interactions rules out such a comparison to temporal network dynamics. Therefore, we instead addressed the question of network macro-evolution here by using extant ecological networks to calibrate the end points of an interaction-centric birth-death simulation model under the assumption that the best-fitting models will provide insight into the network's likely evolutionary history. Among the variety of ecological networks types, bipartite ones are the most appropriate family to test this model: they have well partitioned interactions between guilds with no complex feedback loops, are present in a variety of systems and types of biological interactions, and there is a wealth of well-studied data available (**???**). Moreover, taxa from both guilds of a bipartite ecological network are usually tightly evolutionarily linked and require interactions to persist, making them ideal to elucidate evolutionary rules of community structure.

1 METHODS

1.1 *Model description*

We posit that four simple rules govern the evolution of networks. First, every network originally consists of just two species sharing a single interaction; for example, a plant and its herbivore. Second, a speciation event happens at the top level (*e.g.* the herbivore) with probability p, or at the bottom level with probability $1 - p$. Third, the incipient species starts with all interactions of its ancestor. Fourth, some of these interactions are lost with probability $\varepsilon(\lambda, k, c)$, which allows interactions—that are gained through speciation—to be lost either at a fixed rate λ or as a function of the incipient species' degree k. The c parameter modulates this relationship further by influencing whether high degree of an ancestor increases, or decreases, the probability of the incipient species losing interactions. We have used the following formulation for ε:

$$\varepsilon(\lambda, k, c) = \left(1 + \left(\frac{1}{\lambda} - 1\right) \times c^{k-1}\right)^{-1}. \tag{1}$$

In this formulation, k is the number of interactions of the incipient species, λ is the *basal* rate of interaction loss, and c is a parameter regulating whether species with more interactions tend to gain or lose interactions over time. Negative values of c imply that *rich get richer*, *i.e.* species with more interactions tend to conserve them more over speciation. The special case of $c = 0$ corresponds to no relationship between the degree of a species and its probability of losing or retaining an interaction over speciation. The resulting probability of interaction loss, and its consequences on degree, is shown in fig. 1.1. The values of ε belong to $]0; 1[$. Note that, because species are duplicated upon a speciation event, the network still grows over time. If an incipient species should lose all of its interactions, then it fails to establish.

Probability that each *interaction of the ancestor is lost by the incipient species during speciation.* $\lambda = 0.15$*, and c varied between 1.3 (purple) and 0.7 (green).*

These four rules translate directly into steps for the model: pick a level at random, select a species to duplicate, assess the survival of interactions of the incipient, and add the incipient to the network. These are performed a fixed number of time – we impose an upper limit to the richness at each level, and when this limit is reached, the incipient species replaces one of the resident species at random. An equilibrium for the measures of network structure (see next section) is reached within 1000 timesteps. For all situations, we recorded the network after 5000 iterations.

1.2 *Network measures*

1.2.1 *Connectance*

Connectance, defined as the ratio of realized interactions on the total number of potential interactions, is one of the most common descriptor of network structure. In a bipartite network

with T species at the top, and B at the bottom, having a total of L interactions, it is defined as $Co = L/(T \times B)$. Connectance has a lower bound, as the network cannot have fewer interactions that the number of species in its more speciose level – the minimal connectance is therefore $c_m = \max(T, B)$. This makes the connectance of networks of different sizes difficult to compare, especially since bipartite networks tends to have a low connectance. For this reason, we used a corrected version of connectance, defined as

$$Co^\star = \frac{L - c_m}{T \times B - c_m}\,. \tag{2}$$

This takes values between 0 (the network has the minimal number of interactions) and 1 (all species are connected), but is robust to variations in species richness.

1.2.2 *Nestedness*

We measured nestedness, using the η measure of Bastolla et al. (2009), which returns a global nestedness score based on the fact that interactions of relatively specialized species should be a subset of the interactions of more generalized ones. This measure is robust to changes in species richness, and returns values between 0 (not nested) to 1 (perfectly nested).

1.2.3 *Modularity*

We measured modularity using label propagation coupled with the BRIM measure (LP-BRIM; Liu & Murata 2009) (preliminary analyses revealed no qualitative impact of using other methods to optimize modularity). LP-BRIM returns values close to 1 when there are modules in the network, and values closer to 0 otherwise. The value of modularity for each network is the maximal modularity out of 10 replicates.

1.2.4 *Motifs*

Finally, we enumerated six four-species bipartite motifs (Baker et al. 2014). Bipartite motifs are possible conformations of four species spread across two levels, such as for example three consumers sharing one resource, or two consumers both exploiting resources, *etc.*. The five motifs we used are illustrated in fig. 1.2.4. Because the number of motifs obviously varies with species richness, we corrected it in the following way. For each level, we enumerated the sets of species with a degree allowing them to be part of the motif. Then we multiplied the number of sets for the top (t_x) and the bottom (b_x) level – this gives the number of possible combinations of species that *could* form a given motif. We then divided the count of observed motifs (m_x) by this product, so that

$$m_x^\star = \frac{m_c}{t_c \times b_x} \tag{3}$$

is the *proportion* of species that could form motif x which are actually in the correct conformation. This allows to compare the number of motifs between networks of different sizes.

Illustration of the five motifs used in this study. Motifs 21, 22, and 23 have the same number of species but different numbers of interactions; motifs 31 and 32 are flipped version of one another, and should help discriminate top-rich or bottom-rich communities.

1.3 *Simulations*

To explore the behaviour of the model, we conducted a series of simulations using $p = 0.5$, varying λ from 10^{-4} to 10^{-1} (every order of magnitude), and c from 0.05 to 2.5 (by increments of 0.05). For every combination of parameters, we performed 500 simulations, using 25 species maximum on every level. The network was returned after 8000 timesteps. The network measures were applied on the endpoint of the simulation.

1.4 *Data selection*

We used empirical data of plant-pollinator interactions (59 networks), plant-herbivore interactions (23 networks), phage-bacteria networks (38 interactions), plant-dispersers interactions (30 networks), and host-parasite interactions (121 networks). Pollination and seed-dispersal interactions come from the *WebOfLife* dataset (`http://mangal.io/data/dataset/7/`). Phage-bacteria (which are functionally equivalent to host-parasitoid) data are from (**???**). Host-parasite data (**???**) are from (**???**). Plant-herbivore data are from (**???**). Every network was "cleaned" in the following way. First, species with no interactions (if any) were removed. This yields networks in which all species have at least one interaction. Second, interactions strengths (if present) were removed since our model only requires information about the presence or absence of interactions.

todo herbivory, additional ant-plant data

1.5 *Parameter selection*

We used ABC (Approximate Bayesian Computation) to select the parameter values that yielded realistic networks by assessing how closely each replicate of the second numerical experiment resembles empirical communities (Beaumont 2010). To generate a set of appropriate priors, we randomly generated 10^4 networks with random maximal richness, and random parameters p, λ, and c, then removed the combinations that gave (i) entirely connected networks and (ii) networks with fewer than 5 species on the least species-rich level. We then visually inspected the distribution of parameters to determine their shape, and fitted a distribution using check the method. The distribution of p is uniform between 0 and 1. The distribution of λ is an exponential of parameter $\theta \approx 0.09$ (truncated between 0 and 1). The distribution of c is a normal with parameters $\mu \approx 1.28, \sigma \approx 0.34$ (truncated between 0 and 4). We then used these prior distributions to generate 10^6 random networks, with a number of species at each level drawn uniformly between 10 and 50 (both levels have the same maximum species richness, so as not to interfere with p).

For each empirical network, its observed set of summary statistics (all network measures) was compared to each output of the stochastic model. The Euclidean distance between the two arrays

was recorded as the score ρ of the parameter set. Because each empirical network is in practice a different optimization problem submitted to the ABC routine, and because ABC requires to set the rejection threshold $\rho_{\max}$ on a per-problem basis, setting a global value was not meaningful (Sunnåker et al. 2013). To circumvent this problem, we instead selected the posterior distribution as the 100 parameters sets that gave the lowest ρ. Incidentally, the largest value of ρ for every network ($\rho_{\max}$) can be used to estimate how adequately it is described by the model. Because all the measures are ranged in 0; 1, and because all the posteriors distributions have an equal size, the values of $\rho_{\max}$ can be compared across networks.

For every network, in addition to $\rho_{\max}$, we retain the average parameter values (weighted by ρ^{-1}) $\bar{p}$, $\bar{\lambda}$, and $\bar{c}$, as well as the distance between the empirical value and the simulated value for all network measures.

1.6 *Implementation*

The model (and all the data analysis code) was written in `Julia` (**???**) 0.6.1, using the package `EcologicalNetwork.jl` 1.1.0 – **https://doi.org/10.5281/zenodo.595661**. The code, and copies of the raw data and all intermediate computational artifacts used for this article, is available at todo OSF.IO.

2 RESULTS

2.1 *Model behavior*

By varying the parameters λ (base probability of losing and interactions) and c (effect of the degree on ε, eq. 1), we are able to generate a range of scenarios using the model. In (**???**), we report the values of connectance (as measured by $Co^{\star}$, eq. 2), nestedness, and modularity. As expected, low values of λ and values of c below unity result in the largest connectance. Note that because we removed all combinations of parameters for which fewer than 100 (out of 500) simulations generated networks of acceptable quality (at least five species on each level, connectance below 1.0), the set of feasible parameters depends jointly on the value of λ and c.

2.2 *Predictive ability*

In fig. 2.2, we report the distribution of the values of $\rho_{\max}$ for every network, by type of interaction. Lower values indicate that the network is very well described by the model. All networks are described approximately as well, with the exception of some bacteria-phage networks which are difficult to accurately describe. On the right panel, we have represented the relationship between $\rho_{\max}$ and the average absolute error on every network measure, defined as $\sum |n_0 - x|/|\mathbf{x}|$, where n_0 is the value of every measure on the empirical network, and x is the vector containing the weighted average values in the networks retained as part of the posterior distribution. Note that

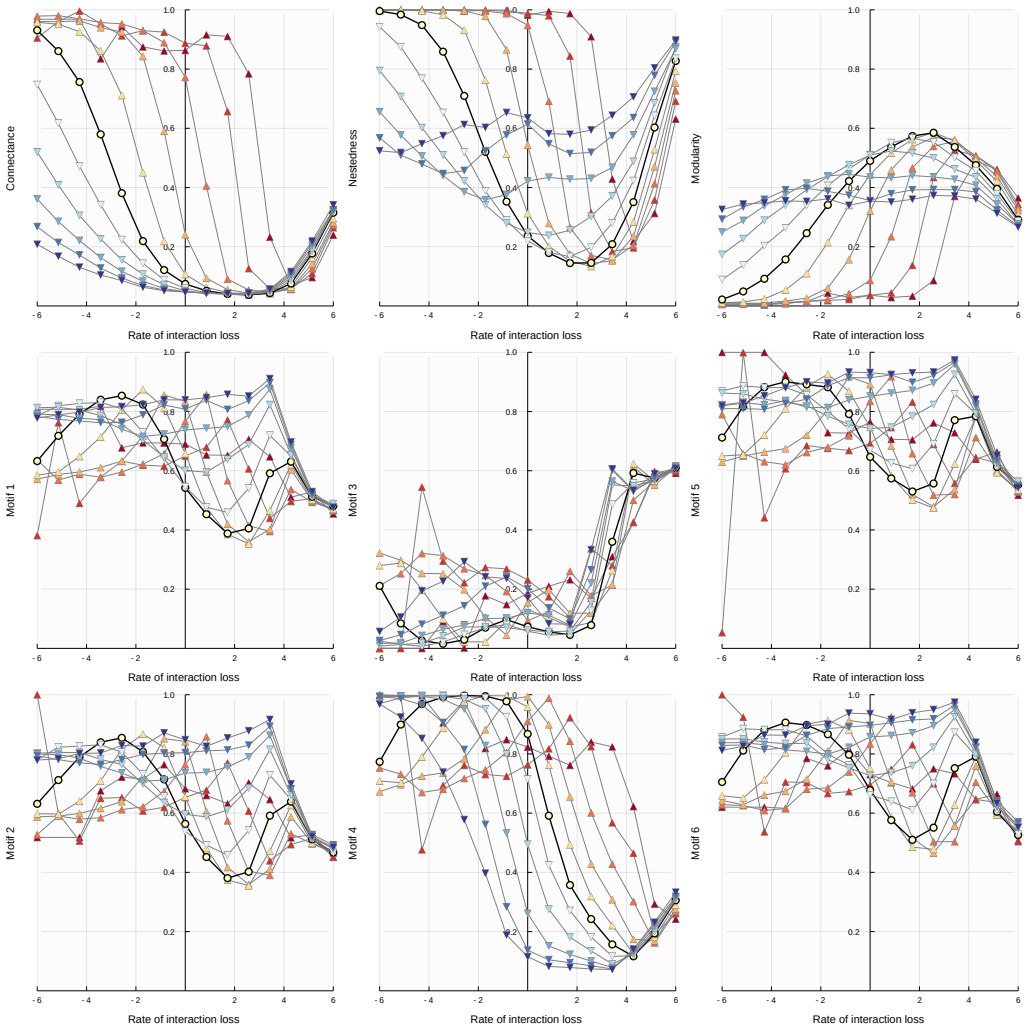

Figure 1 dfff

this quantity is not entirely independent from $\rho_{\max}$, since ρ is defined as the Euclidean distance between the empirical and simulated values. The right panel nevertheless shows that low values of $\rho_{\max}$ accumulate, on average, less error compared with those that are more difficult to fit.

Distribution of the values of ρ (left panel) and average absolute error (right panel) across network types. Note tha the axis for ρ is reversed, as a large value of ρ means that poorly fitting parameter values were accepted when building the posterior distribution. Each point correspond to a network.

2.3 *Evolutionary parameters by network type*

We first observed that the posterior distribution of the parameters differs across interaction types (fig. 2.3). There is no obvious distribution of p by network type, which is expected since the value of p primarily ties into the ratio of top-level to total species, and this not affected by the type of interaction (but see fig. 2.4; the ratio is correctly estimated for most networks). We will focus on the two parameters governing the rate of interaction loss, λ and c. Regardless of interaction types, the values of c were larger than unity, suggesting that on average, ancestors with a high degree tend to have descendants with a lower degree. With the exception of herbivory networks, the basal rate of interaction loss is in $[0, 0.3]$, suggesting that interactions tend to be well conserved over evolutionary timescales. Herbivory networks had best fitting values of λ that were as high as 0.5 – these networks also have a lower connectance, and this is reflected in the high rate of interaction loss.

Scatterplots of the weighted averages for λ and c by network type. Each point represents a network. The ellipses for the 95% confidence interval are represented on each plot.

2.4 *Predictive ability by network measure*

Distribution of the average error on each of the network measures used to estimate fit. Each point represents a network. The corrected connectance tends to be slightly over-estimated by the model, but all other properties are well described.

3 DISCUSSION

Finally, all systems show a strong bias towards moderately high values of c; this indicates that the effective probability of a species retaining its ancestor's interactions decreases with its ancestor's degree. That is, the generalism of species over time has an emergent upper bound, a fact that results in the very spectrum of high-degree and low-degree species that is ubiquitous empirically (**???**).

Emergence of generalists / specialists /// parasite data

Our results demonstrate that the structure of extant bipartite networks can be adequately reproduced by a speciation/extinction model that accounts for biotic interactions. The selection

on parameters related to interaction diversification and persistence was stronger than on the parameter related to the rate of speciation, suggesting that the importance of biotic interactions in macro-evolution may have been understated. Our results also highlight that, while the evolutionary persistence of interactions is undeniably important in the macro-evolution of community structure, different type of ecological interactions respond in largely different ways. This offers a very stimulating possibility – namely, that because the mode of coevolution *within* the interaction between two species differ as a function of their ecological interactions (**???**), this can cascade up to the macro-evolutionary scale in the form of a signal of long-term interaction persistence.

REFERENCES

Baker et al. (2014). Species' roles in food webs show fidelity across a highly variable oak forest. *Ecography.* 38:130–9.

Bastolla et al. (2009). The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature.* 458:1018–20.

Beaumont. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics.* 41:379–406.

Liu & Murata. (2009). Community Detection in Large-Scale Bipartite Networks. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.* Institute of Electrical & Electronics Engineers (IEEE);

Sunnåker et al. (2013). Approximate Bayesian Computation. *PLoS Comput Biol.* 9.