THE UNIVERSITY OF CHICAGO

ASSEMBLING LARGE ECOLOGICAL COMMUNITIES: A THEORETICAL EXPLORATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
CARLOS A. MARCELO SERVÁN

CHICAGO, ILLINOIS
AUGUST 2020

Para mis abuelos, MOSS y Tukis

...Caminante, no hay camino, se hace camino al andar...—Antonio Machado

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

My PhD would not have been possible without the support of many people. First and foremost, I want to thank my mentor Walter Cabrera-Fébola. In a country where we are mostly unaware of graduate studies, I had the luck of meeting Walter during my undergraduate years. He taught me what science is, and still serves as a role model for me. Over our many conversations throughout the years, I have learned so much about life, art and science from him. Thanks for always believing in me and pushing me to try my best.

I would like to thank my PhD advisor Stefano Allesina for his immense support during all the years of my PhD. For always be willing to listen and discuss my ideas and make the Allesina Lab such a welcoming and wonderful work environment.

I would like to thank the members of my committee, Mercedes Pascual, Sarah Cobey and José Capitán for their valuable input into my research projects and for forcing me to think more deeply about my work and put it into the bigger context of the field.

To my collaborators, Stefano Allesina, José Capitán, Dan Maynard, Zach Miller, Jacopo Grilli, Matteo Sireci and Kent Morrisson for making our projects so fun to work on. It has been a pleasure and an incredible learning experience to work together.

To the members of the Allesina Lab, past and present, for being okay with being interrupted at random times by me spouting nonsense and be willing to discuss my ideas, Zach Miller in particular has been the subject of many of these events. To the current members of the lab, Paula Lemos, Abby Skwara and Zach Miller for comments on this dissertation.

As for any immigrant, adjusting to this country was rough and I thank Jacopo Grilli, Matt Michalska-Smith, Gyuri Barabás, Liz Sanders, Angelo Monteiro and José Capitán for making the lab such a welcoming environment that helped me to cope with it. From Jacopo I learned that we can be invested in our science and still have a life outside the lab. I thank him for his friendship, for making me feel more at home by talking to me in Spanish and for the many things I learned from him. To Angelo Monteiro, for his friendship and the many

conversations we had during my first summer here. To José Capitán for making the summers in which he visited Chicago some of the most memorable times during my four years here. For his incredible support, for believing in me, his friendship and teaching me the "joy" of calculations (which to his best efforts I still find very painful to do).

To my friends, Pedro Espinoza, Andre Mostacero, miembros de la Tía Jockey, Yeonwoo Park, Stephanie Sang, Soo Ji Kim, Gayani Senevirathne, Wei Liang, Zachary Miller, Amanda Crego-Emly, Vicente Jiménez, Matteo Sireci, Emilio Berti, Shai Pilosof, Victoria Romero, Sergio Alcalá-Corona and Micol Tresoldi for all the fun moments we have had during the last four years which helped me cope with stressful times.

To my partner, Maryn Carlson, for her unwavering support, her love and kindness which have pulled me up so many times when I was about to give up. This dissertation would not have been finished without all her emotional and editorial support.

To my family, starting with my grandparents Manuel Octavio Serván Soplín (a.k.a. MOSS or *yeyo*) and Tula Esperanza Trigoso Ampuero (Tukis or *yeya*) to which this dissertation is dedicated, for helping my mother raise me and for all their unconditional love and support for me. To my other relatives, in particular my mother Zully, aunt Orffa and uncle papá Marino. For their understanding, love and support. It is very difficult to tell your Peruvian family that first you want to be a scientist and second that you want to go live thousands of kilometers away from them. I had the luck to have a supporting and loving family that besides the distance I can always feel close to me.

ABSTRACT

Ever since the work of May, almost 50 years ago, ecologists have sought to find the mechanisms underlying the stability of species-rich communities. In order to reconcile empirical observations of highly diverse systems with theoretical models, many such mechanisms were proposed. Yet, recent refinements of May's work suggest that stability in large systems requires a large degree of fine-tuning. As such, rich ecosystems are, in theory, very susceptible to perturbations in the model parameters, hindering their ability to persist. What has been largely overlooked, however, is that the set of species we observe is actually the end product of a process. This process, called *community assembly*, is a complex interplay between immigration from a regional pool of species to a local habitat and local extinctions fueled by species interactions. Explicitly considering the whole process allows us to gain a deeper understanding of how highly diverse communities exist. In this work, we study community assembly models under Lotka-Volterra dynamics and a constant regional pool of invaders. Under different parameterizations and modes of invasion, we show that the observation of highly speciose communities is possible without any need of fine-tuning—simply by letting the system assemble itself. Thus, while an arbitrary set of species cannot coexist, the process of assembly allows an arbitrary regional pool of species to give rise to species-rich local communities. In other words, a diverse regional pool gives rise to diverse local communities.

CHAPTER 1

INTRODUCTION

Ecological communities can be astoundingly diverse: a forest fragment in the Amazon or a tract of the human gut can harbor thousands of species. In the early days of ecology, researchers thought that the amazing diversity displayed by ecological communities would translate into an increased resistance to external perturbations [42, 77]. The intuition behind this hypothesis was as follows: if the different species in the community respond to an external perturbation in uncorrelated ways, then the more the species the more a perturbation would "even out", reducing its impact.

In his seminal paper [81], Robert May challenged this view. Contrary to common belief, he showed that the more diverse a system is, the more unlikely it is to be stable. As such, diversity *per se* does not enhance the stability of the system.

To arrive at this conclusion, May considered the population dynamics of an ecological community resting at an equilibrium point, and modeled the interactions between species as random perturbations. More precisely, he modeled the "community matrix" (i.e., the Jacobian matrix evaluated at equilibrium) of the N-species community as $A = B - I$, where I is the identity matrix and B is a *random matrix* with coefficients assumed to be sampled independently from identical distributions (*i.i.d.*). The distribution from which the coefficients are sampled has mean zero and variance σ^2—and as such belongs to a well-studied class of random matrices [111]. Using fundamental results from random matrix theory, one can compute the probability that a community is locally stable—stable systems recover from (arbitrarily small) perturbations. In particular, the stability bound is given by $\sigma\sqrt{N} = 1$, and a sharp transition in the probability of stability (from almost sure stability to almost sure instability) occurs at this point as $N \gg 1$.

May's analysis presupposes the existence of an equilibrium point where all the species are present at positive densities. The existence of such a point, i.e., the *feasibility* of the

system, plays as crucial a role, together with its stability, in determining the coexistence of species [52]. For example, in the textbook case of competitive dominance, coexistence is not precluded because of instability, but rather because no feasible equilibrium point exists [127]. In general, for a set of species to be observed, both feasibility and stability conditions must be satisfied.

One can argue that ecologists responded to May's work by developing a new research program—which continues to this day—the primary goal of which is to uncover the *features* of ecological communities that make coexistence of highly diverse communities possible. In the ensuing decades, many such features have been proposed [31, 53, 59, 60, 86, 113]. For example, contrary to May's assumption, the non-random arrangement of interactions (i.e., the community's network structure [95]) can influence dynamics. Several studies have shown that the trophic hierarchy typical of food webs, modularity, or nestedness, greatly affect both the stability and the feasibility of the system [5, 51, 52]. Yet, it is not clear if the observation of non-random network structure in natural communities is the product of feasibility or stability constraints, or rather the consequence of how ecological systems are built [83, 116].

Another widely popular explanation for the origination and maintenance of diversity is provided by "Modern Coexistence Theory" (MCT) [9, 29]. This framework highlights the importance between niche differences and fitness differences in promoting species coexistence—as in the well-studied case of the two-species competitive Lotka-Volterra model. Recently, the central tenet of this theory—that fitness and niche differences can be considered independent—was questioned [105]; similarly, the theory does not easily generalize to highly diverse communities [7].

In general, even when models consider features promoting coexistence, they show that when the number of species grows, the region of parameter space allowing for full coexistence shrinks. Thus, full coexistence requires *fine-tuning* of the parameters [52]—an assumption that becomes difficult to reconcile with the biological reality of fluctuating environments,

intra-specific and temporal variation in physiological parameters, etc. In this respect, it is important to consider a different direction with which the field responded to the difficulty of observing the coexistence of speciose communities: *community assembly* [38, 39, 69]. Assembly theory, which builds upon the ideas of succession pioneered by Cowles [35], rests on the observation that the systems we observe in nature are not put together in one go without any extinctions. Rather, they are the outcomes of the complex balance between local extinction and colonization. In the view of community assembly, to explain coexistence one has to consider these processes. While community assembly was a fertile area of research during the late 80's and 90's, progress has slowed in recent years, due to a variety of challenges in both empirical and theoretical domains (see Chapter 3).

By acknowledging the process of community assembly, we can move past the dichotomies of stable/not stable or feasible/non feasible. As May's work and many recent refinements showed, throwing an arbitrary set of species together is unlikely to yield full coexistence [4, 8, 52], but what happens when we perform this experiment? Does the system collapse to a very small community, or does a considerable fraction of species survive? By studying these questions, we move further toward understanding how the coexistence of many species is possible. For example, we can ask to what extent are the numbers of species we observe actually "high" compared to expectations. Consider the following example: suppose that we associate each species in a community with a coin, where heads stand for species that are present, and tails for extinct species. If we flipped 50 unbiased coins, the chance of getting 50 heads (full species coexistence) is almost 0. On the other hand, if we had tossed 100 coins, and subsequently removed all the tails (i.e., removed the extinct species), then the probability of observing 50 heads is quite high! This example highlights how theoretical expectations on full coexistence could not be informative of the processes that led to the build-up of biodiversity in natural systems. Worse, as shown by Maynard et al. [83], we could observe highly non-random structures in the community of extant species—and conclude that

these structures are responsible for coexistence, rather than simply be the by-product of the ecological assembly process.

In this thesis, we explore these questions with models that resemble May's work in their simplicity. We show that coexistence of many species is possible—without any fine-tuning—by simply letting the ecological system assemble itself. The results are of direct relevance to current approaches to build and design large ecological communities [15, 50]. In these laboratory experiments, natural microbial communities are sampled, and challenged with a new, synthetic environment; dynamics lead the system to a final community of coexisting species, often showing that the more species are in the initial sample, the richer the final community [15].

In the next chapter, we focus on communities governed by strongly-stable Lotka-Volterra (LV) systems with random interactions and growth rates. Exactly as in the experiments described above [15, 50], all the species from a regional are included initially at positive densities—what we term *top-down assembly*. While, consistently with May's result, the full system is usually unlikely to coexist, the community collapses to subsystems of moderate size (see also Bunin [21]). In particular, when interactions and growth rates are equally likely to be positive or negative, we find that on average half of the species survive without any saturation with respect to the number of species in the pool (reflecting the fact that a LV model with random coefficients implicitly assumes a potentially infinite set of niches). Then, highly diverse systems can be built without any additional coexistence mechanism—we simply need to start with a large enough species pool, and let the dynamics trim it down to about half of the original size. Further, by deriving the full distribution of the number of coexisting species, we provide expectations for the size of the coexisting subsets in the local communities, conditioned on the size of the regional pool. As mentioned above, the effect of the pool's size is seen very clearly in the work of Bittleston and colleagues [15]. In fact, their work supports a linear relationship between the size of the initial pool and the size of

the final community—exactly what is expected under LV with random interactions.

In the third chapter, we explore a more traditional way of assembling the system, which can be seen at play in the many simulations published in the 90's. In contrast to chapter two, species enter the system by *sequential invasion*. That is, species enter the system one at a time, with enough time between invasions for the community to reach its asymptotic configuration. We term this invasion scheme *bottom-up* assembly. We state sufficient conditions under which the *a priori* different processes of bottom-up and top-down assembly are equivalent, in the sense that the final configuration of species attained by either process is the same. In doing so, we make use of a mathematical representation of assembly known as the "assembly graph" [25, 54, 102]. The assembly graph G of a pool of species encodes all the possible sequences of invasions and coexisting subcommunities which can appear along the process of assembly. Two nodes (feasible states) are connected by a link in G if an invasion induces the shift between two communities (see also Amor et al. [6]). As such, G completely captures the assembly process.

By demonstrating under which conditions top-down and bottom-up assembly lead to the same outcome, we connect decades of literature on bottom-up assembly with a new generation of experimental ecologists working mostly with top-down assembly [15, 50]. Thus, our results open the door to experimental exploration of bottom-up assembly. Early experiments highlighted the challenges, mainly due to the high number of potential assembly sequences, involved in the experimental design of bottom-up assembly [38, 120]. Our results show that a way forward is to use results from top-down assembly experiments to infer the structure of the assembly graph (in a similar way as what was proposed in a purely coexisting setting in Maynard et al. [85]) and use the predictions of possible assembly sequences to guide the development of the experimental protocol. Besides its importance for basic science, this type of experiment has immediate connections to the engineering of ecological communities [71], and the development of restoration practices [112].

In the fourth chapter, we extend the ideas put forward in the first chapter by considering a case in which the random interactions between species are influenced by phylogenetic relatedness. The link between interactions and phylogenies is developed via a trait-based interaction model. For a regional pool of N species, ℓ traits (with $\ell \geq N$) evolve by Brownian motion on the phylogenetic tree relating the species in the regional pool. The competitive interactions between species are modulated by the sample covariance matrix derived from the species' trait values. In this way, the closer in the tree two species are, the stronger they compete in expectation. The number of traits determines the variance of the interactions.

Assuming top-down assembly, we characterize the properties of the surviving communities as a function of the tree structure, the number of traits (ℓ), and the number of species (N) in the regional pool. In particular, we find that when the number of traits is very high relative to the number of species, then full coexistence is guaranteed and the tree structure is fully reflected in the abundance distribution of the communities. Following from the results of chapter three, assembling the communities by bottom-up assembly will give us the same results. In this way, our results provide new baseline expectations for the ways in which the phylogenetic information at the regional pool level is translated into local community properties.

We conclude by highlighting some caveats of our analysis and possible ways to overcome them. For example, one limitation is that we focused on symmetric or strongly stable Lotka-Volterra systems, and always considered a fixed regional pool. We also propose new ways to extend our results. The ideas of the second chapter can be easily extended to study systems where the interactions are parameterized by a "low-dimensional" trait space. Low-dimensional approximations to interaction matrices are useful when trying to infer interactions from data which are not sufficient to parameterize each and every coefficient (for such conditions see Maynard et al. [85]), as for example data coming from Biodiversity-Ecosystem functioning experiments [114]. Similarly, the results of the fourth chapter could be used to

devise statistical tests assessing whether phylogenies influence community patterns.

CHAPTER 2

COEXISTENCE OF MANY SPECIES IN RANDOM ECOSYSTEMS

Abstract

Rich ecosystems harbor thousands of species interacting in tangled networks encompassing predation, mutualism and competition. Such widespread biodiversity is puzzling because in ecological models it is exceedingly improbable to obtain the stable coexistence of large communities. One aspect rarely considered in these models, however, is that coexisting species in natural communities are a selected portion of a much larger pool, which has been pruned by population dynamics.

Here we compute the distribution of the number of species that can coexist when we start from a pool of species interacting randomly, and show that even in this case we can observe rich, stable communities. Interestingly, our results show that, once stability conditions are met, network structure has very little influence on the level of biodiversity attained.

Our results identify the main drivers responsible for widespread coexistence in natural communities, providing a baseline for determining which structural aspects of empirical communities promote or hinder coexistence.

2.1 Introduction

Lotka [75] and Volterra [118] first attempted to mathematize the population dynamics of interacting species, and their model has been eviscerated and refined by countless studies [65]. Analyzing models that include more than a handful of interacting populations has however

0. This work has been published in Nature Ecology & Evolution volume 2, 1237-1242(2018) [103]. This chapter is reproduced from the manuscript under the license to publish of the Nature Publication Group (NPG)

proven remarkably difficult, despite the fact that ecosystems harbor hundreds of populations, interacting through complex networks encompassing consumption, competition, and mutualism [95].

In Lotka-Volterra and similar models, it is exceedingly improbable to obtain the coexistence of all species in a large community without fine-tuning the parameters [48, 52, 81, 98, 104], and such fine-tuning is questionable at best for biological systems [53]. Consider however that in natural communities the extant species we observe are a selected portion of a much larger pool, which has then been pruned by population dynamics [74, 104]. Therefore, to understand the establishment and maintenance of natural communities we need to change our focus: rather than asking what is the probability that all species in a community coexist, here we attempt to predict the number of extant species we obtain when starting from a species pool of n species, and let the dynamics unfold. As a limiting case, we study the behavior of ecological models in which the parameters are randomly drawn from fixed distributions, meaning that species have not had time to co-adapt or co-evolve. While many studies have investigated, numerically [20, 37, 49, 62, 94, 97, 108, 124] or analytically [10, 21], the effect of particular parameterizations and network structure on the average number of coexisting species, here we derive the full distribution.

We start by studying coexistence in random ecological communities, and, having derived the behavior of random networks of interacting species, we probe the effect of particular network structures on coexistence. We find that network structure, which has been shown to have strong influence on the stability properties of ecological communities [5, 7, 51, 95, 98, 126], has instead very little effect on coexistence, once stability conditions are met.

The idea of studying random ecological communities was pioneered by May [81], who determined the local stability properties of large ecosystems through an application of random matrix theory. His work was generalized and refined [3, 4], so that we can now characterize the stability of ecological networks displaying hierarchical [5] or modular [51] structure.

Similarly, "structural stability" (i.e., the range of conditions leading to positive equilibria in ecological systems) has been investigated by letting the growth rate of the species [52, 98], or the interactions between species [107] vary randomly. Clearly, to have robust coexistence we need a combination of the two: species densities must be positive, while a stable attractor is needed to allow densities to rebound when perturbed.

2.2 Results

Our goal is to compute the probability of observing k species stably coexisting when starting with a pool of n interacting populations and random parameters. For example, take the generalized Lotka-Volterra (GLV) system

$$\frac{dX_i(t)}{dt} = X_i(t)\left(r_i + \sum_j A_{ij}X_j(t)\right), \tag{2.1}$$

and sample parameters at random: how many species coexist once the dynamics have elapsed?

We first analyze the case closest to the spirit of May's contribution, which can be thought of as a caricature of a food web: some species can grow in isolation (e.g., producers, with positive intrinsic growth rates), while other species can grow only thanks to their interactions (e.g., consumers, with negative growth rates); all species establish random interactions with each other. More specifically, we sample the intrinsic growth (death) rates (r_i) and the inter-specific interactions (A_{ij}, $i \neq j$) from distributions (not necessarily the same) that are symmetric around zero (such that $P(x) = P(-x)$). For example, we could sample all these entries from a Normal distribution with mean zero. We set the intra-specific interactions (A_{ii}) by summing a mean-zero symmetric random variable and a constant d_i (not necessarily the same for all i). Note that in this way, about half of the species would grow in isolation, while the rest rely on "consumption" for their survival.

We start by presenting a result on the feasibility of equilibria. Under the conditions outlined above, the probability that a system composed of n species has a completely positive equilibrium point (i.e., in which all species have positive density) is $1/2^n$, irrespective of the choice of d_i, and the exact shapes of the distributions (section 2.4.1). Our proof extends previously known mathematical results [89], confirming the conjecture put forward by Goh & Jennings forty years ago [49].

Clearly, feasibility is only necessary, but not sufficient for coexistence. To study coexistence, we make the stronger assumption that the matrix $A + A^T$ is negative definite [43, 63]. This property implies Lyapunov diagonal stability, and is a strong form of stability routinely assumed in studies of feasibility [52, 98] that can be always attained by choosing suitable large and negative d_i. Under these conditions, a GLV model has a single, globally attractive equilibrium, called the non-invasible solution (also known as the saturated rest point [57, 58]): k species have positive density at equilibrium, while all the other $n - k$ species cannot invade this community, and will go extinct irrespective of initial conditions. Surprisingly when we sample the parameters at random as specified above, the non-invasibility and feasibility conditions for each subset of species balance out, such that each species has probability $1/2$ of being included in the non-invasible, globally attractive solution. Hence, the probability $P(k|n)$ of finding k species coexisting when we start with n follows the binomial distribution $B(n, 1/2)$ (Figure 4.1 and Figure 2.4). This beautifully simple result means that if we were to start with a strongly stable (i.e., with $A + A^T$ negative definite) random matrix of interactions and random growth rates, about half of the species would coexist, irrespective of the choice of n. Remarkably, this is exactly what we would expect if species were not to interact with each other at all (i.e., $A_{ii} = d_i < 0$ for all i and $A_{ij} = 0$ for all $i \neq j$).

Extending May's results, Allesina & Tang [3] showed how stability is strongly influenced by the correlation between the inter-specific interactions: if we sample interactions in pairs (A_{ij}, A_{ji}) from a bivariate distribution with mean zero and correlation ρ, then stability is

enhanced by choosing a negative correlation. When analyzing coexistence, breaking the independence among the inter-specific effects by sampling them in pairs from a bivariate distribution has no effect: we recover the same condition for feasibility, and the same distribution for the number of coexisting species (Figure 4.1 and section 2.4.1).

So far, we have assumed that every species interacts with every other. To study the effect of network structure, we set most of the interactions to zero, and choose the position of the nonzero coefficients according to the adjacency matrix of a) an Erdős-Rényi random graph, b) a random graph with power-law degree distribution, c) a graph displaying modular, or d) bipartite structure. Irrespective of the choice of network structure, we always recover the same distribution for the number of coexisting species k (Figure 4.1 and Section 2.4.1). This is interesting, because network structure strongly influences stability [3, 5, 7, 51]. However, because in our analysis stability is assumed, we find that the exact location of the nonzero interactions has no effect on coexistence.

The results above hold when we sample the growth rates and the inter-specific effects from symmetric distributions with mean zero, meaning that positive effects (e.g., contribution of prey to the growth of predators) on average counterbalance negative ones (e.g., effects of predators on prey). Of course this needs not to be the case in natural communities, and therefore we examine the mathematically much more challenging case in which the entries have mean nonzero.

To this end, we consider a simple model of interacting competitors: we set all inter- and intra-specific interactions to be negative, and consider the case of random growth rates. In this case we assume that all species in the pool are sampled from a common habitat, and therefore have growth rates with a well-defined average value. In particular, we sample the intrinsic growth rates from a Normal distribution with mean γ, and, for simplicity, we construct A by setting all inter-specific interaction to be competitive, $A_{ij} = \mu = \hat{\mu}/n < 0$, and all intra-specific effects to $A_{ii} = d_i = \alpha < 0$. Numerical simulations presented below

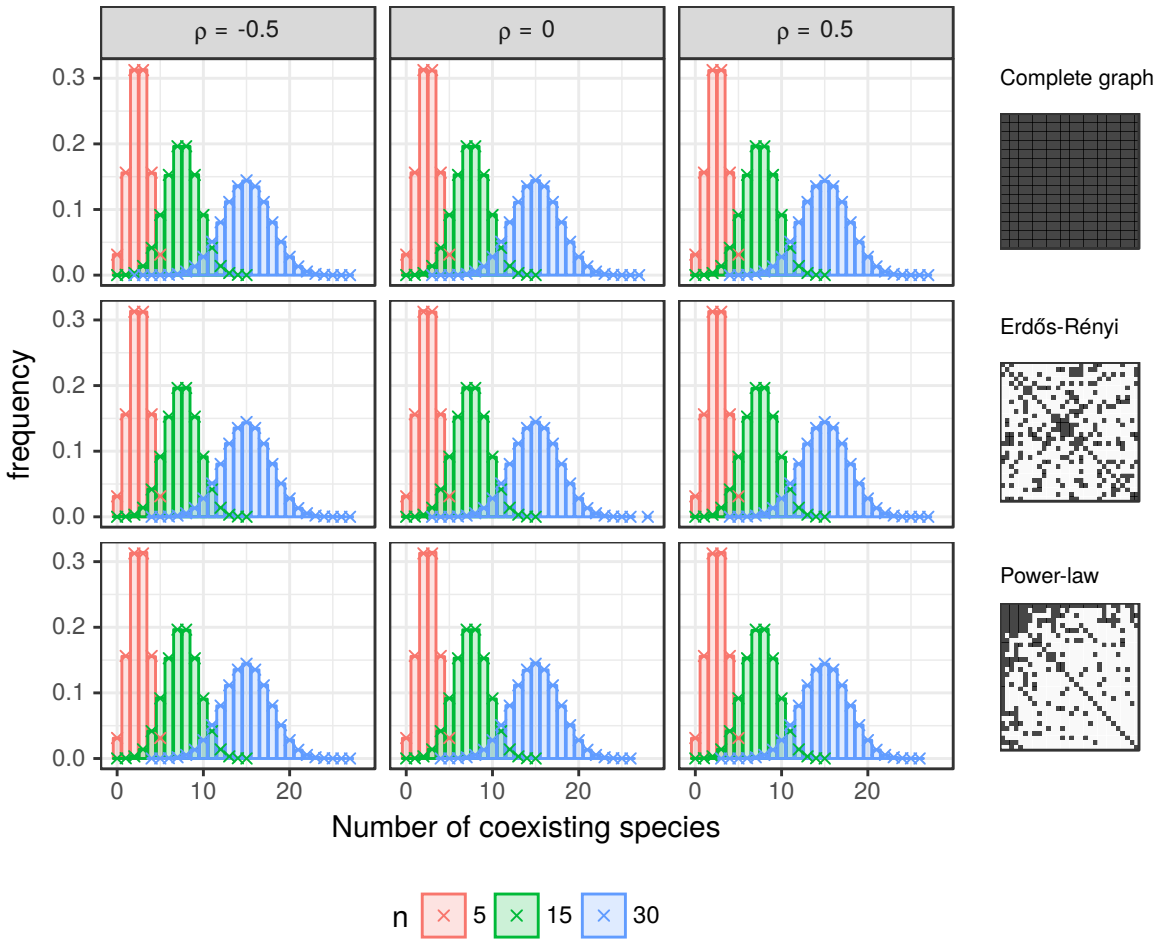

Figure 2.1: **Number of coexisting species when interactions and intrinsic growth rates are randomly sampled from the standard Normal distribution.** For each panel, histograms show the number of coexisting species out of $2 \cdot 10^5$ simulations, when starting from a different number of species n (colors) and interaction matrices A that are strongly stable. Binomial distributions $B(n, 1/2)$ are reported as crosses. In the three rows, different network structures are used to set the positions of the nonzero coefficients (as exemplified by the adjacency matrices on the right); top: complete graphs, middle: Erdős-Rényi graphs; bottom: Power-law graphs; the results for other network structures are presented in Figure 2.4. Sampling the off-diagonal coefficients of matrix A independently ($\rho = 0$, center), or in correlated pairs (A_{ij}, A_{ji}) ($\rho \neq 0$), has no effect on the expected number of coexisting species.

show that our results well-approximate the case in which the elements of A are variable (e.g., when the nonzero elements are arranged in a network).

Again, we consider matrices for which α is sufficiently strong to yield Lyapunov diagonal stability ($\alpha < \mu < 0$). When we sample the growth rates from a Normal distribution, then the equilibrium point $X = -A^{-1}r$ is described by a multivariate Normal distribution. Exploiting this fact, we are able to express the probability that k species form a non-invasible and feasible subset as a double integral that can be used to compute the size of the non-invasible community (Supplementary Information). The double integral can be approximated, for large n, via a saddle-point technique to obtain an accurate analytical approximation for the distribution $P(k|n; \alpha, \hat{\mu}, \gamma)$. Note that in this setting, growth rates need to be positive for species to survive, and therefore we only consider the case of $\gamma \geq 0$. We also show (Section 2.4.6) that the results remain qualitatively unchanged when rates are drawn from a truncated Gaussian distribution, which forces all rates to remain strictly positive.

The results (Figure 4.2) show that a nonzero mean γ in growth rates can yield a larger (red area of parameter space) or smaller (blue) number of coexisting species, compared to the mean-zero case. If

$$\frac{\alpha\gamma}{\hat{\mu}} > \frac{1}{\sqrt{2\pi}}, \tag{2.2}$$

averages are larger than expected in the mean-zero case (and conversely). The distribution $P(k|n; \alpha, \hat{\mu}, \gamma)$ is not binomial anymore, but still retains a strong central tendency. Importantly, the mode of the number of species can be estimated analytically (section 2.4.5).

When we repeat the calculation but position the nonzero elements according to a network structure, we find results that are quite similar to the mean-zero case: though not all network structures yield the same exact distribution, the effect is very modest, such that our analytical approximation well-describes coexistence in all cases (Figure 4.3).

In summary, we have computed the distribution of the number of coexisting species under

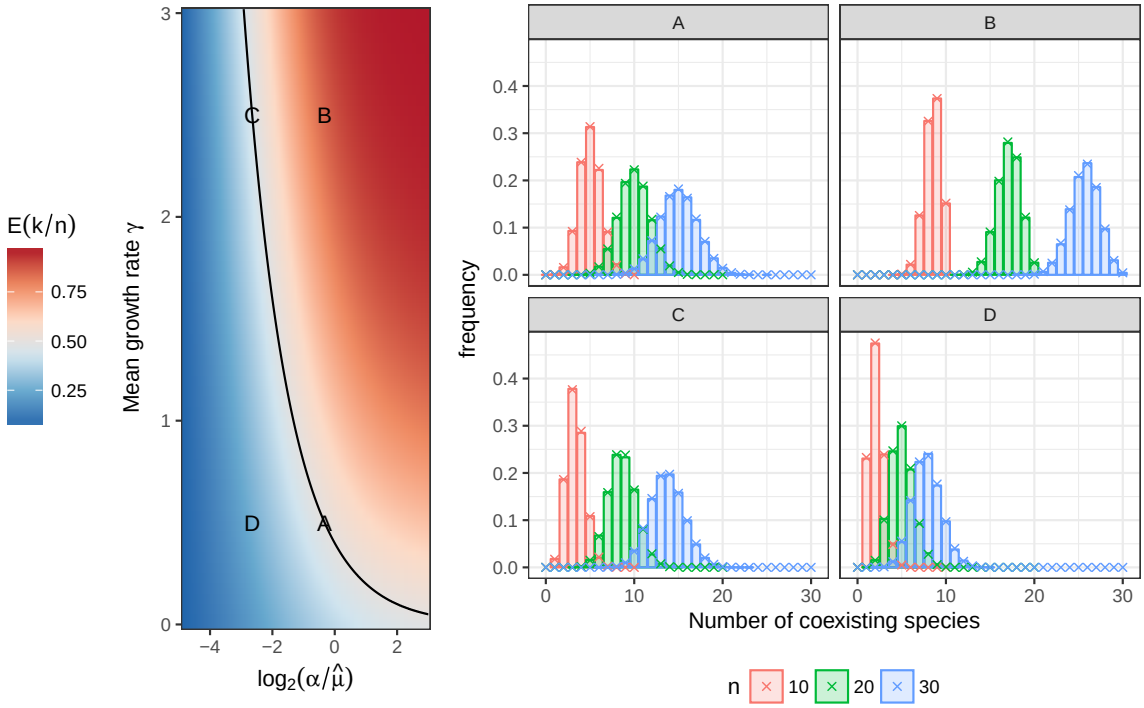

Figure 2.2: **Number of coexisting species for competitive interactions.** When species interact competitively, the histograms deviate from the binomial distribution, but can still be computed using a double integral (crosses, see Eq. S38). Here the interactions are set to $A_{ij} = \hat{\mu}/n < 0$, intra-specific competition to $A_{ii} = \alpha$, and intrinsic growth rates are normally-distributed with mean γ. The expected value of the ratio k/n, $E(k/n)$, is drawn on the left in the relevant parameter space: we chose two points (A, C) for which predictions in the nonzero mean case match closely those for mean zero ($E(k/n) = 1/2$); in case B the number of species coexisting exceeds that for the mean-zero case; for point D the expectation is lower. The analytical prediction in equation (2.2) is also shown (line).

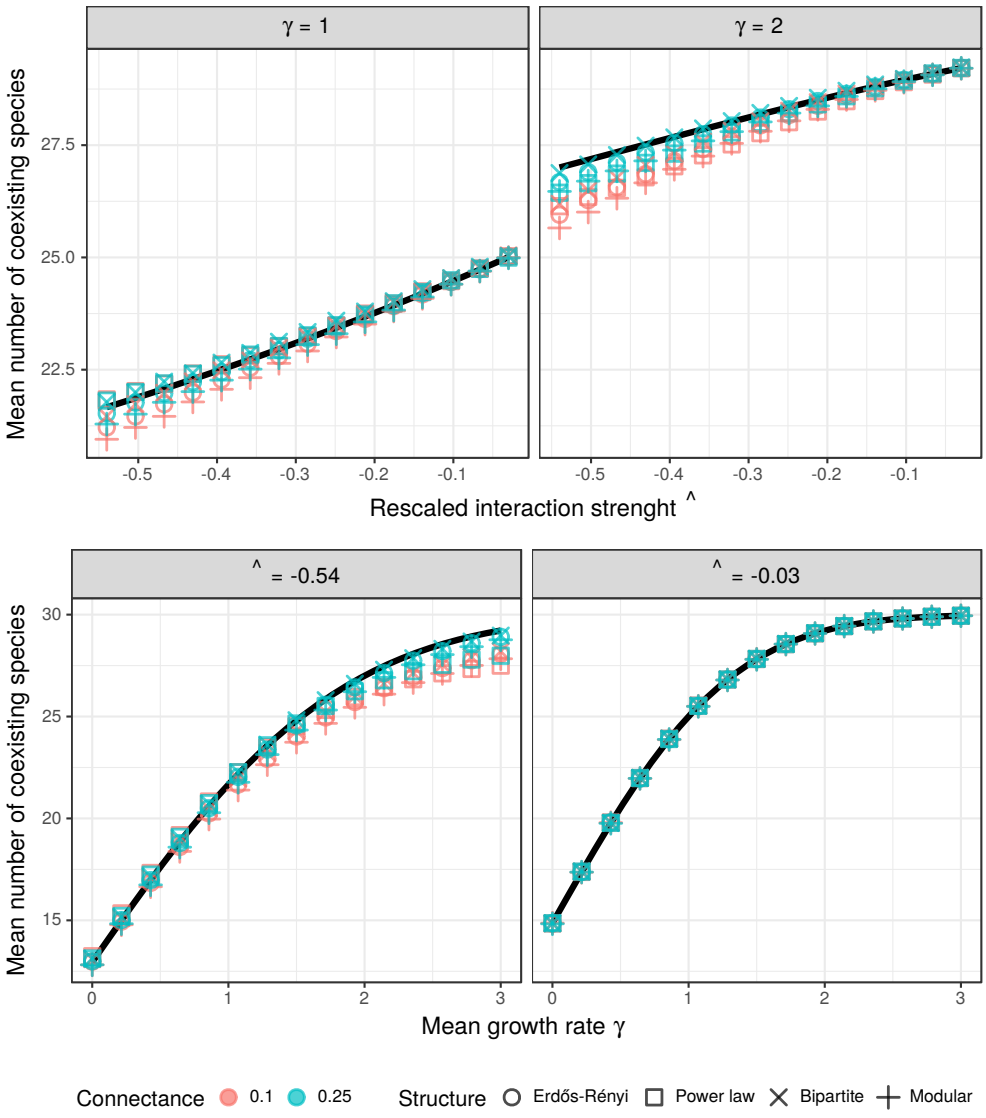

Figure 2.3: **Effect of network structure on coexistence for the case of nonzero means.** The position of the nonzero coefficients is chosen according to one of four structures (shape), and for two levels of connectance (proportion of nonzero coefficients, colors). Because most of the coefficients are zero, one needs to calculate a "rescaled" $\hat{\mu}$ (x-axis in upper panels, see Supplementary Information) in order to contrast the results of the simulations (point) with our analytical approximation for the fully-connected case (line, see Supplementary Information eq. (2.88)). The four panels show that, although both the interaction strength $\hat{\mu}$ and the mean intrinsic growth rate γ interacts with connectance and network structure in nontrivial ways, the overall effect is very modest.

the assumptions of random parameters and strong stability. We have two cases: a) when inter-specific interactions have mean zero, the number of coexisting species follows the binomial distribution with probability 1/2, and network structure has no influence whatsoever—in fact, we would recover the same result if species were not to interact at all; b) when the inter-specific interactions have mean nonzero, the distribution is not binomial anymore, and we can expect either a larger or smaller proportion of populations to survive, depending on the choice of parameters. Also in this case, however, network structure has a very modest effect.

2.3 Discussion

Our results show that large communities can stably coexist thanks to the selection imposed by the dynamical pruning of a large species pool. In practice, we can attain communities of any size (with no saturation) even when setting parameters at random—all we need is to start with a much larger species pool.

When the growth (death) rates and the interactions are sampled from distributions that are symmetric about zero, we find that the number of coexisting species follows the binomial distribution. Given that this result holds for any distribution of interactions (provided it is symmetric about zero), it follows that the same should be found when interactions are all zero (i.e., when species do not interact), which can be thought of as a limiting case.

In our simple model of a food web, we sample growth (death) rates and interactions independently. As such, we could create species that have a negative growth rate (typical of consumers) and yet establish only negative interactions with the rest of the species (typical of producers). Such species will surely go extinct, resulting in a non-trivial correlation between the growth rates of the extant species and the type of interaction they can establish (i.e., for the extant species, negative growth rates are possible only when they consume other species). This explains the shift in the distribution of growth rates found in the extant communities

after pruning (Section 2.4.7).

The picture complicates considerably when interactions and growth rates can have nonzero mean: as found in the case of feasibility [52], nonzero means in growth rates and interactions can have a strong effect on the number of species that can coexist. For the competitive case studied here, this can enhance or depress biodiversity, depending on the parameterization.

Many studies have tracked the average number of species coexisting using simulations [20, 37, 49, 62, 94, 97, 108, 124]. Recently, much progress has been made analytically by borrowing methods from the physics of disordered systems [10, 21]. We add to this growing body of literature by focusing on the full distribution of the number of extant species. Our results can be applied to two cases of interest. First, when we have many local communities composed of subsets of the same pool of species, such as in metacommunities, our methods could model the distribution of the number of species found in local patches. Second, thanks to progresses in molecular biology and imaging it is now possible to assemble increasingly large microbial communities in laboratory setting. Our methods could be used as a baseline distribution for the number of strains coexisting when randomly seeding communities with a subset of a microbial pool.

The study of the stability of large ecological communities started by considering completely random matrices of interactions [81]; further studies included more realistic models in which interactions were paired [3] and organized in patterns [5, 51]. We believe that our results can be similarly extended, and we see three main directions that need to be explored.

First, we have considered here a "weak" form of network structure: the location of the nonzero elements of the matrix is specified, but other than that the coefficient values are randomly determined. A stronger form of network structure would be one in which also the values of the nonzero coefficients are organized in a pattern. For example, a "cascade" structure in which all the positive (negative) elements of the matrix A are confined to the upper (lower) triangular part has been shown to have a strong stabilizing (or destabilizing)

effect [5]. Similarly, arranging the strong/weak competitive interactions in modules or in a nested fashion can greatly influence stability [7]. It would therefore be important to determine whether this "strong" formulation of network structure can indeed influence coexistence as well as stability.

Second, we have determined coexistence under the assumption of strong stability (Lyapunov diagonal stability). Relaxing this constraint will be challenging, but could however shed light on mechanisms of coexistence involving for example limit cycles or chaotic attractors. Recently, Bunin [21] studied coexistence in species pools with random (weak) interactions and identical growth rates, identifying the transitions between systems characterized by a single stable equilibrium, and those displaying multiple attractors. Though this study disregards other types of attractors, it shows that analytical progress in this area is possible.

Third, as pointed out by Sigmund [104], "Mother Nature does not assemble her networks by throwing n species together in one go". Understanding the process of assembly in which communities are built one species at a time is perhaps the greatest challenge ahead for theoretical community ecology [83]. In the Section 2.4.9 we show that, although some of our non-invasible communities cannot be built by a sequential assembly, the probability of finding such cases decreases rapidly with the size of the community. We conjecture that, asymptotically, the probability of finding an assembly sequence for communities built in this way converges to one.

In the last few decades, ecologists have compiled ever more detailed interaction networks [41], documenting the intricate relationships occurring in ecosystems [64, 101]. These networks display interesting patterns, such as broad degree distributions [40], modular organization of interactions [93], hierarchical structure [32], and nestedness [11, 106]. One of the main questions in community ecology is therefore to determine whether these network properties have some bearing for the robust coexistence of ecological communities.

In this context, our results provide a baseline for species coexistence under Lotka-Volterra dynamics—one can use these reference points to prove that certain features of empirical communities promote or hinder coexistence.

Methods

Problem statement We consider n interacting populations, whose dynamics are defined by a system of Generalized Lotka-Volterra (GLV) equations:

$$\frac{dX_i(t)}{dt} = X_i(t)\left(r_i + \sum_j A_{ij}X_j(t)\right), \tag{2.3}$$

where $X_i(t)$ is the abundance of population i at time t, r_i is the intrinsic growth rate of species i, and A_{ij} is the per-capita effect of species j on the growth rate of species i. For notational convenience, we collect the coefficients A_{ij} into the interaction matrix $\boldsymbol{A}$, and X_i and r_i into the (column) vectors $\boldsymbol{X}$ and $\boldsymbol{r}$, respectively.

A vector $\boldsymbol{x}^\star$ is a fixed point (equilibrium) of the system if

$$0 = x_i^\star\left(r_i + \sum_j A_{ij}x_j^\star\right) \quad \text{for } i = 1, 2, ..., n. \tag{2.4}$$

Since $x_i^\star = 0$ is always a possible solution, the system admits up to 2^n fixed points, corresponding to all the combinations of presence and absence of each species.

A fixed point is feasible if $x_i^\star > 0$ for all i. If a feasible fixed point exists, it is the solution of

$$\boldsymbol{r} = -\boldsymbol{A}\boldsymbol{x}^\star. \tag{2.5}$$

If $\boldsymbol{A}$ is invertible, then

$$\boldsymbol{x}^\star = -\boldsymbol{A}^{-1}\boldsymbol{r}. \tag{2.6}$$

Global stability and non-invasible fixed points. In this study, we assume that $\boldsymbol{A}$ is negative definite, and in particular that the matrix $\boldsymbol{A}+\boldsymbol{A}^T$ has only negative eigenvalues [63]. A matrix $\boldsymbol{A}$ is Lyapunov diagonally stable if there exists a positive diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{D}\boldsymbol{A} + \boldsymbol{A}^T\boldsymbol{D}$ is negative definite [43]. Our assumption therefore implies Lyapunov diagonal stability (corresponding to choosing $\boldsymbol{D}$ as the identity matrix).

If $\boldsymbol{A}$ is diagonally stable, then there exists a fixed point of equation (2.3) that is globally attractive: irrespective of the (positive) initial conditions, dynamics always converge to the same fixed point [58]. This globally stable fixed point has k positive entries and $n-k$ entries equal to zero. We define the support $\{S\}_k$ as the set of k persistent species (i.e., those for which at equilibrium $x_i^\star > 0$) and $\{N\}_{n-k} = \{S\}_n \setminus \{S\}_k$ as the set of $n-k$ species with zero abundance. The i^{th} entry of the globally stable fixed point $\boldsymbol{x}^\star$ is equal to zero if $i \in \{N\}_{n-k}$ and equal to $x_i > 0$ if $i \in \{S\}_k$, where $\boldsymbol{x} = (x_i)$ is a k-dimensional (column) vector with positive components. We define the $k \times k$ matrix $\boldsymbol{A}^{(s)}$ as the submatrix of $\boldsymbol{A}$ obtained by considering only rows and columns belonging to $\{S\}_k$. Similarly, we define the $(n-k)\times(n-k)$ matrix $\boldsymbol{A}^{(n)}$ by considering rows and columns in $\{N\}_{n-k}$, the $k \times (n-k)$ matrix $\boldsymbol{A}^{(sn)}$ by considering rows in $\{S\}_k$ and columns in $\{N\}_{n-k}$, and the $(n-k)\times k$ matrix $\boldsymbol{A}^{(ns)}$ by considering rows in $\{N\}_{n-k}$ and columns in $\{S\}_k$. Finally, the entries of the intrinsic growth rate vector can be split into two subvectors $\boldsymbol{r}^{(s)}$, a $k-$dimensional (column) vector with same components of $\boldsymbol{r}$ for the entries in $\{S\}_k$, and $\boldsymbol{r}^{(n)}$, a $(n-k)-$dimensional (column) vector with entries corresponding to $\{N\}_{n-k}$.

If we rearrange the indices of the vectors such that the k persistent species occupy the first k entries, the globally stable fixed point $\boldsymbol{x}^\star$ can be written as the vector $\begin{pmatrix} \boldsymbol{x} \\ \\ \mathbf{0}_{n-k} \end{pmatrix}$, where $\mathbf{0}_{n-k}$ denotes a (column) vector with $n-k$ zero entries, the intrinsic growth rate

vector becomes $\boldsymbol{r} = \begin{pmatrix} \boldsymbol{r}^{(s)} \\ \boldsymbol{r}^{(n)} \end{pmatrix}$, and the interaction matrix reads

$$\boldsymbol{A} = \left(\begin{array}{c|c} \boldsymbol{A}^{(s)} & \boldsymbol{A}^{(sn)} \\ \hline \boldsymbol{A}^{(ns)} & \boldsymbol{A}^{(n)} \end{array} \right) . \tag{2.7}$$

The abundance of the k persistent species is therefore a solution of the equation

$$\boldsymbol{A}^{(s)}\boldsymbol{x} = -\boldsymbol{r}^{(s)} . \tag{2.8}$$

Since we are considering only diagonally stable matrices, this point is also not invasible by any of the remaining $n-k$ species (i.e., none of the species in $\{N\}_{n-k}$ can invade when the system is resting at the equilibrium point) [58]. The condition of non-invasibility can be written by imposing that the growth rate of each of the $n-k$ species is negative for small densities. In the limit of small densities, the per-capita growth rates of the invaders become independent of their densities, and one obtains the following $n-k$ conditions

$$\boldsymbol{r}^{(n)} + \boldsymbol{A}^{(ns)}\boldsymbol{x} < 0 . \tag{2.9}$$

In the case of diagonally stable matrices, the combination of $\{S\}_k$ and $\boldsymbol{x}$ is unique. It is the only one for which the solution $\boldsymbol{x}$ of equation (2.8) has positive components and, simultaneously, equation (2.9) holds.

Distribution of non invasible fixed points Provided that $\boldsymbol{A}$ is diagonally stable, the number of coexisting species k is fully and uniquely determined by the vector of intrinsic growth rates $\boldsymbol{r}$. More precisely, only the direction of the vector $\boldsymbol{r}$, and not its norm, determines coexistence. Our goal is to determine $P(k|n)$, the probability of observing k coexisting species out of n, given a distribution for the entries of the matrix $\boldsymbol{A}$ and a distribution for the intrinsic growth rates $\boldsymbol{r}$. In particular, we parameterize the entries of $\boldsymbol{A}$ as the sum of a

deterministic and a random matrix:

$$A_{ij} = (\alpha - \mu)\delta_{ij} + \mu + B_{ij} \ , \tag{2.10}$$

where $\boldsymbol{B}$ is a random matrix, whose entries are random variables with mean zero, and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. As such, the entry A_{ii} (self-interaction) has mean α, while the off-diagonal entries have mean μ. Similarly, we consider

$$r_i = \gamma + b_i \ , \tag{2.11}$$

where the entries of vector $\boldsymbol{b}$ are random variables with mean zero.

We define $\mathcal{P}(\{S\}_k|\boldsymbol{A})$ as the probability (calculated over the growth rate vectors $\boldsymbol{r}$) that the support of the globally stable fixed points is $\{S\}_k$. By averaging this quantity over the distribution of $\boldsymbol{A}$, we obtain

$$\mathcal{P}(\{S\}_k|n) := \mathbb{E}(\mathcal{P}(\{S\}_k|\boldsymbol{A})) \ . \tag{2.12}$$

The probability that the support has cardinality k is simply

$$P(k|n) := \sum_{\{S\}_k} \mathcal{P}(\{S\}_k|n) \ . \tag{2.13}$$

The derivations are presented in the Supplementary Information. In particular, in section S1 we focus on the case $\mu = 0$ and $\gamma = 0$, showing that if the distributions of the entries $\boldsymbol{B}$ and $\boldsymbol{b}$ are symmetric around zero

$$P(k|n) = \binom{n}{k} \frac{1}{2^n} \ . \tag{2.14}$$

In section 2.4.2 we provide an integral formula for $\mathcal{P}(\{S\}_k|\boldsymbol{A})$ in case of a arbitrary matrix $\boldsymbol{A}$ and in section S3 we exploit this results to compute explicitly $P(k|n)$ in the case

of $\mu \neq 0$, $\gamma \neq 0$, $\boldsymbol{B} = 0$ and normally distributed entries of $\boldsymbol{b}$. Sections S4 and S5 use saddle-point techniques to provide the mode of the distribution of the number of coexisting species. In section 2.4.6, we present a derivation for the case in which growth rates are positive and sampled from a truncated Gaussian distribution. Section 2.4.7 analyzes the differences between the starting and final community, and Section 2.4.8 details the numerical simulations. Finally, in section 2.4.9 we discuss the relationship between our model and the process of ecological assembly.

2.4 Supplementary Information

2.4.1 Mean zero

Toy model: uncoupled logistic equations

Suppose that $\boldsymbol{A}$ is a diagonal matrix, and therefore that species do not interact with each other. For stability, we need $A_{ii} < 0$ for all i (self-regulation). Let p_i be the probability of $r_i > 0$. Then, the probability that a solution $\boldsymbol{x}$ with k positive components $\{S\}_k$ is non-invasible is $\prod_{i\in\{S\}_k} p_i \prod_{i\notin\{S\}_k}(1-p_i)$.

When the distribution of r_i is symmetric around zero, $p_i = \frac{1}{2}$ irrespective of the distribution of $A_{ii} < 0$, and thus the probability of non-invasibility is $\frac{1}{2^n}$ for any particular subsystem. Therefore, the binomial distribution with parameters n and $\frac{1}{2}$ describes the the number of persistent species.

Feasibility

In this and the following section, we show that when the entries of matrix $\boldsymbol{B}$ and vector $\boldsymbol{r}$ are random variables whose distribution is symmetric around 0, and that any n element subset of the columns of $\boldsymbol{B}$ and $\boldsymbol{r}$ are linearly independent (which holds almost surely if the entries of $\boldsymbol{B}$ and $\boldsymbol{r}$ are sampled from a continuous probability distribution function

and are independent of each other), then the probability $P(k|n)$ is still described by the binomial distributions with parameters n and $\frac{1}{2}$—exactly what we found for non-interacting species. Note that this holds true both for the case in which the coefficients B_{ij} are sampled independently, and for the case in which these coefficients are sampled in pairs (B_{ij}, B_{ji}), and the pairs are sampled independently from a bivariate distribution symmetric around $(0, 0)$.

First we show that $P(n|n) = \frac{1}{2^n}$. The proof amounts to showing that, of all the possible 2^n sign $(+, -)$ patterns for the entries of a solution to equation 6 of the main text, each of them is equally probable.

Let $\boldsymbol{x}^\star$ be an arbitrary solution of equation 6 of the main text, and define the matrix $\boldsymbol{D}_k = ((-1)^{\delta_{ik}}\delta_{ij})$. Then, $\boldsymbol{D}_k\boldsymbol{x}^\star$ satisfies $(\boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k)\boldsymbol{D}_k\boldsymbol{x}^\star = -\boldsymbol{D}_k\boldsymbol{r}$. Because of the symmetry assumption, we have that $\boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k$ has the same distribution[1] as $\boldsymbol{A}$, and similarly for $\boldsymbol{D}_k\boldsymbol{r}$ and $\boldsymbol{r}$. Since $\boldsymbol{D}_k$ just flips the sign of the k^{th} component of $\boldsymbol{x}^\star$, by repeating this operation a sufficient number of times we can connect any two sign patterns of solutions to equation 6, and thus the conclusion follows.

Persistent species

As noted before, in the regime of diagonally stable matrices, the final state of the system is the non-invasible (also called saturated) fixed point of the system[58]. With the same assumptions of the previous section the distribution for the number of persistent species follows naturally: the probability of having a non-invasible solution $\boldsymbol{x}$ with k positive components (with support $\{S\}_k$) is the joint probability of the conditions expressed in equations 8 and 9 of the main text, which can be written as $\mathcal{P}(\{S\}_k|n) = P(k|k)[1 - P_{\text{inv}}(\{S\}_n \setminus \{S\}_k|\{S\}_k)]$, where P_{inv} denotes the probability of being invasible by any of the remaining species given that $\boldsymbol{x} > 0$. Let $\boldsymbol{z} = \boldsymbol{r}^{(n)} + \boldsymbol{A}^{(ns)}\boldsymbol{x}$. By following the same procedure illustrated in the

1. This transformation also has the property of preserving the eigenvalues of the matrix, which allows this argument to hold also if we condition on Lyapunov diagonally stable matrices.

previous section (applying the appropriate change of signs to $\boldsymbol{A}$ and $\boldsymbol{r}$), one can show that any sign pattern for $\boldsymbol{z}$ is equally likely, therefore $1 - P_{\text{inv}}(\{S\}_n \setminus \{S\}_k | \{S\}_k) = \frac{1}{2^{n-k}}$. As a consequence, $\mathcal{P}(\{S\}_k|n) = \frac{1}{2^n}$. Because of the uniqueness of this type of solution for a given interaction matrix $\boldsymbol{A}$ and a vector of rates $\boldsymbol{r}$, the binomial distribution with parameters n and $\frac{1}{2}$ describes the distribution of the number of species having positive density at the globally stable equilibrium.

Adding Structure

Let $\boldsymbol{G}$ be the adjacency matrix of an undirected graph, and consider the matrix $\boldsymbol{M} = \boldsymbol{G} \circ \boldsymbol{A}$, where $\circ$ represents the Hadamard (entry-wise) product between $\boldsymbol{G}$ and $\boldsymbol{A}$. Because this type of product is commutative with respect to the multiplication by a diagonal matrix, i.e., $\boldsymbol{D}(\boldsymbol{G} \circ \boldsymbol{A})\boldsymbol{D} = \boldsymbol{G} \circ (\boldsymbol{DAD})$ for $\boldsymbol{D}$ diagonal, the arguments used in the previous two sections still hold. This means that the distribution of $\boldsymbol{M}$ is invariant to $\boldsymbol{D}_k \boldsymbol{M} \boldsymbol{D}_k$ (even when $\boldsymbol{G}$ is also a random matrix) and by restricting ourselves to diagonally stable matrices the linear independence assumption is assured (the matrix is invertible). Consequently, adding a network structure in this way does not change the probability of feasibility nor the distribution of persistent species.

2.4.2 Calculating the distribution of persistent species

If we integrate the GLV dynamics starting from an interaction matrix $\boldsymbol{A}$, a vector of intrinsic growth rates $\boldsymbol{r}$, and an arbitrary (positive) initial condition with n species, we end up with k species with density different from zero and $n-k$ species with density equal to zero. If the matrix $\boldsymbol{A}$ is diagonally stable, the end point of the dynamics always correspond to a fixed point $\boldsymbol{x}^\star$, irrespective of the initial conditions.

The goal of this section is to provide a formula for the probability $P(k|n)$ of finding k persisting species out of n, for an arbitrary matrix $\boldsymbol{A}$, under the assumption that $\boldsymbol{A}$ is

diagonally stable. We assume that the entries of $\boldsymbol{r}$ are drawn from a Normal distribution with mean γ and unit variance. This choice of a variance does not affect the generality of out results, since the coexistence properties of the Generalized Lotka-Volterra equations are independent of the norm of $\boldsymbol{r}$: rescaling all growth rates by a constant simply rescales all equilibrium abundances by the same constant, with no impact on feasibility or stability.

We define the vector $\boldsymbol{z}$ with $n-k$ components as

$$\boldsymbol{z} := \boldsymbol{r}^{(n)} + \boldsymbol{A}^{(ns)}\boldsymbol{x}. \tag{2.15}$$

On the other hand, we have equation 8 of the main text, defining $\boldsymbol{x}$. By imposing feasibility and non-invasibility —equation 9 of the main text—, it must hold that $\boldsymbol{x} > 0$ and $\boldsymbol{z} < 0$.

Using the probability density of the growth rates,

$$\begin{aligned} P(\boldsymbol{r}) &= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\sum_{i=1}^{n} \frac{(r_i - \gamma)^2}{2} \right) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\|\boldsymbol{r}^{(s)} - \gamma \mathbf{1}_k\|^2 - \frac{1}{2}\|\boldsymbol{r}^{(n)} - \gamma \mathbf{1}_{n-k}\|^2 \right) , \end{aligned} \tag{2.16}$$

where $\mathbf{1}_k$ stands for a $k-$dimensional column vector whose entries are all equal to one. Introducing equation 8 of the main text and equation (2.15), we can write the joint probability density as

$$f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{A}) = \frac{|\det \boldsymbol{\Lambda}|}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|\boldsymbol{A}^{(s)}\boldsymbol{x} + \gamma \mathbf{1}_k\|^2 - \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{A}^{(ns)}\boldsymbol{x} - \gamma \mathbf{1}_{n-k}\|^2 \right) , \tag{2.17}$$

where $\boldsymbol{\Lambda}$ is the Jacobian matrix obtained from the change of variables $\boldsymbol{r} \to (\boldsymbol{x}, \boldsymbol{z})$. According to equations 8 and (2.15), it is simple to observe that $\boldsymbol{\Lambda}$ as the following structure:

$$\boldsymbol{\Lambda} := \left(\begin{array}{c|c} \frac{\partial \boldsymbol{r}^{(s)}}{\partial \boldsymbol{x}} & \frac{\partial \boldsymbol{r}^{(s)}}{\partial \boldsymbol{z}} \\ \hline \frac{\partial \boldsymbol{r}^{(n)}}{\partial \boldsymbol{x}} & \frac{\partial \boldsymbol{r}^{(n)}}{\partial \boldsymbol{z}} \end{array} \right) = \left(\begin{array}{c|c} \boldsymbol{A}^{(s)} & \mathbf{0} \\ \hline \boldsymbol{A}^{(ns)} & \boldsymbol{I}_{n-k} \end{array} \right) , \tag{2.18}$$

$\boldsymbol{I}_{n-k}$ being the $(n-k)-$dimensional identity matrix. Therefore $|\det \boldsymbol{\Lambda}| = |\det \boldsymbol{A}^{(s)}|$.

The first term appearing in the exponential in equation (2.17) can be written as

$$\|\boldsymbol{A}^{(s)}\boldsymbol{x} + \gamma \mathbf{1}_k\|^2 = (\boldsymbol{x} - \boldsymbol{\xi})^T \boldsymbol{G} (\boldsymbol{x} - \boldsymbol{\xi}) \ , \tag{2.19}$$

where

$$\boldsymbol{\xi} = -\gamma (\boldsymbol{A}^{(s)})^{-1} \mathbf{1}_k \ , \tag{2.20}$$

and

$$\boldsymbol{G} = (\boldsymbol{A}^{(s)})^T \boldsymbol{A}^{(s)} \ . \tag{2.21}$$

We obtain therefore

$$f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{A}) = \frac{|\det \boldsymbol{A}^{(s)}|}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\xi})^T \boldsymbol{G} (\boldsymbol{x} - \boldsymbol{\xi}) - \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{A}^{(ns)} \boldsymbol{x} - \gamma \mathbf{1}_{n-k}\|^2 \right) \ . \tag{2.22}$$

The probability $\mathcal{P}(\{S\}_k|\boldsymbol{A})$ of observing the globally stable fixed point with support $\{S\}_k$, can be obtained from the joint probability in equation (2.17) by imposing the feasibility condition for the k species ($\boldsymbol{x} > 0$) and the non-invasibility condition for the other $n-k$ species ($\boldsymbol{z} < 0$). The equation reads

$$\mathcal{P}(\{S\}_k|\boldsymbol{A}) \equiv \int d^k \boldsymbol{x} \left(\prod_{i=1}^{k} \Theta(x_i) \right) \int d^{n-k} \boldsymbol{z} \left(\prod_{j=k+1}^{n} \Theta(-z_j) \right) f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{A}) \ . \tag{2.23}$$

2.4.3 Mean non zero

In this section we consider a simplified interaction matrix $\boldsymbol{A}$ whose diagonal coefficients are all equal to α, and all the off-diagonal elements are set to a fixed value μ:

$$\boldsymbol{A} = (\alpha - \mu) \boldsymbol{I}_n + \mu \mathbf{1}_n \mathbf{1}_n^T \ . \tag{2.24}$$

Since the matrix $\boldsymbol{A}$ is a deterministic matrix, in this case $\mathcal{P}(\{S\}_k|\boldsymbol{A}) = \mathcal{P}(\{S\}_k|n)$. By introducing equation (2.24) in equation (2.22) and using equation (2.23), we obtain

$$\mathcal{P}(\{S\}_k|n) = \frac{|\det \boldsymbol{A}^{(s)}|}{(2\pi)^{n/2}} \int d^k\boldsymbol{x} \prod_{i=1}^{k} \Theta(x_i) \int d^{n-k}\boldsymbol{z} \prod_{j=k+1}^{n} \Theta(-z_j)$$
$$\times \exp\left\{-\frac{1}{2}\Big(\boldsymbol{x} - \xi^{(k)}\mathbf{1}_k\Big)^T \boldsymbol{G}\Big(\boldsymbol{x} - \xi^{(k)}\mathbf{1}_k\Big) - \frac{1}{2}\|\boldsymbol{z} - (\mu(\mathbf{1}_k^T\boldsymbol{x}) + \gamma)\mathbf{1}_{n-k}\|^2\right\}, \quad (2.25)$$

where we used the fact that, with the parameterization of equation (2.24), $\boldsymbol{\xi} = \xi^{(k)}\mathbf{1}_k$, where

$$\xi^{(k)} = -\frac{\gamma}{\alpha + (k-1)\mu} \ . \quad (2.26)$$

Again, using equation (2.24) together with equation (2.21), we have

$$\boldsymbol{G} = (\alpha - \mu)^2\boldsymbol{I}_k + \Big[k\mu^2 + 2\mu(\alpha - \mu)\Big]\mathbf{1}_k\mathbf{1}_k^T \ . \quad (2.27)$$

We change variables to $x_i' = x_i - \xi^{(k)}$ to get

$$\mathcal{P}(\{S\}_k|n) = \frac{|\det \boldsymbol{A}^{(s)}|}{(2\pi)^{n/2}} \int d^k\boldsymbol{x} \prod_{i=1}^{k} \Theta(x_i + \xi^{(k)}) e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x}}$$
$$\times \int d^{n-k}\boldsymbol{z} \prod_{j=k+1}^{n} \Theta(-z_j) e^{-\frac{1}{2}\|\boldsymbol{z} - \left[\gamma + k\mu\xi^{(k)} + \mu(\mathbf{1}_k^T\boldsymbol{x})\right]\mathbf{1}_{n-k}\|^2}. \quad (2.28)$$

We now write $z_j' = z_j - \gamma - k\mu\xi^{(k)}$ and obtain

$$\mathcal{P}(\{S\}_k|n) = \frac{|\alpha - \mu|^{k-1}|\alpha + (k-1)\mu|}{(2\pi)^{n/2}} \int d^k\boldsymbol{x} \prod_{i=1}^{k} \Theta(x_i + \xi^{(k)}) e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x}}$$
$$\times \int d^{n-k}\boldsymbol{z} \prod_{j=k+1}^{n} \Theta(-z_j - \gamma - k\mu\xi^{(k)}) e^{-\frac{1}{2}\|\boldsymbol{z} - \mu(\mathbf{1}_k^T\boldsymbol{x})\mathbf{1}_{n-k}\|^2} \ , \quad (2.29)$$

where we used

$$|\det \boldsymbol{A}^{(s)}| = |\alpha-\mu|^{k-1}|\alpha+(k-1)\mu| \ . \tag{2.30}$$

By introducing the expression for $\boldsymbol{G}$ obtained in equation (2.27), we get

$$\mathcal{P}(\{S\}_k|n) = \frac{|\alpha-\mu|^{k-1}|\alpha+(k-1)\mu|}{(2\pi)^{n/2}} \int d^k\boldsymbol{x} \prod_{i=1}^{k} \Theta(x_i+\xi^{(k)}) \\ \times \int d^{n-k}\boldsymbol{z} \prod_{j=k+1}^{n} \Theta(-z_j-\gamma-k\mu\xi^{(k)}) e^{g(\boldsymbol{x},\boldsymbol{z})} \tag{2.31}$$

where

$$g(\boldsymbol{x},\boldsymbol{z}) = -\frac{1}{2}\Big[(\alpha-\mu)^2\boldsymbol{x}^T\boldsymbol{x} + \Big[n\mu^2+2\mu(\alpha-\mu)\Big](\mathbf{1}_k^T\boldsymbol{x})^2 - 2\mu(\mathbf{1}_k^T\boldsymbol{x})(\mathbf{1}_{n-k}^T\boldsymbol{z}) + \boldsymbol{z}^T\boldsymbol{z}\Big]. \tag{2.32}$$

We can express this probability as a double integral by introducing two new variables thanks to a Hubbard-Stratonovich transformation: if $b>0$ and $c>0$, it holds that

$$e^{-bd^2/c^2-de/c} = \frac{c}{2\pi}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-(by^2+ey+idw-icwy)}. \tag{2.33}$$

for any real d and e numbers. Similarly, for $b>0$ and $c>0$,

$$e^{-bd^2/c^2+de/c} = \frac{c}{2\pi}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-(by^2+ey+idw+icwy)}. \tag{2.34}$$

In our case [cf. equations (2.27) and (2.32)], we choose $d = \mathbf{1}_k^T\boldsymbol{x}$ and $e = \mathbf{1}_{n-k}^T\boldsymbol{z}$ and identify the exponents of the l.h.s. of equations (2.33) or (2.34) with the terms in (2.32). If $\mu>0$, we find $\frac{1}{c}=\mu$ and use equation (2.34). If $\mu<0$, we set $\frac{1}{c}=|\mu|$ and consider equation (2.33). In both cases, we set $\frac{b}{c^2} = \frac{1}{2}\big[n\mu^2+2\mu(\alpha-\mu)\big]$. In general, we can choose $c=\frac{1}{|\mu|}$ and $b=\frac{1}{2}\Big[n+2\Big(\frac{\alpha}{\mu}-1\Big)\Big]$. To ensure diagonal stability, all the eigenvalues of matrix $\boldsymbol{A}$ must be negative. This implies the conditions $\alpha-\mu<0$ and $\alpha-\mu+n\mu<0$. If $\mu>0$, the

second restriction can be violated for n sufficiently large. Therefore we limit the discussion to the $\mu < 0$ case (competitive communities) and use equation (2.33). In this case we have $\alpha < \mu < 0$ (hence $|\alpha| > |\mu|$) and $\frac{\alpha}{\mu} - 1 + n > 0$ (hence $b > 0$ and we can apply the Hubbard-Stratonovich transformation). Therefore

$$\mathcal{P}(\{S\}_k|n) = C_k \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}\left[n+2\left(\frac{\alpha}{\mu}-1\right)\right]y^2 + i\frac{yw}{|\mu|}} \int d^k\boldsymbol{x} \prod_{i=1}^{k} \Theta(x_i + \xi^{(k)})$$
$$\times \int d^{n-k}\boldsymbol{z} \prod_{j=k+1}^{n} \Theta(-z_j - \gamma - k\mu\xi^{(k)}) e^{-\frac{1}{2}(\alpha-\mu)^2\boldsymbol{x}^T\boldsymbol{x} - i(\mathbf{1}_k^T\boldsymbol{x})w} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z} - \left(\mathbf{1}_{n-k}^T\boldsymbol{z}\right)y}. \quad (2.35)$$

Where $C_k(\mu, \alpha, n) := \frac{|\alpha-\mu|^{k-1}|\alpha+(k-1)\mu|}{(2\pi)^{n/2+1}|\mu|}$. We complete squares and obtain

$$\mathcal{P}(\{S\}_k|n) = C_k \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}\left[n+2\left(\frac{\alpha}{\mu}-1\right)\right]y^2 + i\frac{yw}{|\mu|}} e^{-\frac{k}{2(\alpha-\mu)^2}w^2 + \frac{1}{2}(n-k)y^2}$$
$$\times \left[\int dx \Theta(x + \xi^{(k)}) e^{-\frac{1}{2}(\alpha-\mu)^2\left(x + \frac{iw}{(\alpha-\mu)^2}\right)^2}\right]^k \left[\int dz \Theta(-z - \gamma - k\mu\xi^{(k)}) e^{-\frac{1}{2}(z+y)^2}\right]^{n-k}. \quad (2.36)$$

Denoting the cumulative distribution function of the standard Normal distribution $N(0,1)$ as $\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$ we can write

$$\mathcal{P}(\{S\}_k|n) = \frac{|\alpha-\mu|^{k-1}|\alpha+(k-1)\mu|}{2\pi|\mu||\alpha-\mu|^k} \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}\left[k+2\left(\frac{\alpha}{\mu}-1\right)\right]y^2 + i\frac{yw}{|\mu|} - \frac{k}{2(\alpha-\mu)^2}w^2}$$
$$\times \left[1 - \Phi\left(\frac{iw}{|\alpha-\mu|} - |\alpha-\mu|\xi^{(k)}\right)\right]^k \left[\Phi\left(y - \gamma - k\mu\xi^{(k)}\right)\right]^{n-k}, \quad (2.37)$$

and therefore we find

$$\mathcal{P}(\{S\}_k|n) = \frac{1}{2\pi}\left|k+\frac{\alpha}{\mu}-1\right| \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}\left[k+2\left(\frac{\alpha}{\mu}-1\right)\right]y^2+i\left|\frac{\alpha}{\mu}-1\right|yw-\frac{1}{2}kw^2} \times \left[1-\Phi\left(iw-|\alpha-\mu|\xi^{(k)}\right)\right]^k \left[\Phi\left(y-\gamma-k\mu\xi^{(k)}\right)\right]^{n-k}. \quad (2.38)$$

Note that $\gamma+k\mu\xi^{(k)} = \gamma\left(1-\frac{k\mu}{\alpha+(k-1)\mu}\right) = \frac{\gamma(\alpha-\mu)}{\alpha+(k-1)\mu}$. We define $s := \frac{\alpha}{\mu}-1$ (which satisfies $s>0$ to ensure diagonal stability) and

$$v := \frac{\gamma(\alpha-\mu)}{\alpha-\mu+k\mu} = \frac{\gamma s}{k+s}. \quad (2.39)$$

Then, given that $\alpha<\mu$, it holds that $|\alpha-\mu|\xi^{(k)} = -\frac{\gamma|\alpha-\mu|}{\alpha+(k-1)\mu} = v$ and we can express the probability in its final form as

$$\mathcal{P}(\{S\}_k|n) = \frac{k+s}{2\pi}\int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}(k+2s)y^2+isyw-\frac{1}{2}kw^2}[1-\Phi(iw-v)]^k[\Phi(y-v)]^{n-k}. \quad (2.40)$$

In this formula, the integration over w must be performed in the complex plane. An alternative way to express it is to consider a path Γ in the complex plane such that $\Gamma = \{w' \in \mathbb{C}|w' = iw+x_0\}$ and then reducing the result to the limit $x_0 \to 0$, so that the integral over the imaginary axis is well defined. Therefore, an equivalent form of writing this equation is

$$\mathcal{P}(\{S\}_k|n) = \frac{k+s}{2\pi i}\int_{-\infty}^{\infty} dy \int_{\Gamma} dw\, e^{-\frac{1}{2}(k+2s)y^2+syw+\frac{1}{2}kw^2}[1-\Phi(w-v)]^k[\Phi(y-v)]^{n-k}, \quad (2.41)$$

where the integral in w has to be evaluated over the contour Γ and then take the limit $x_0 \to 0$.

Note that for the case $k=0$ the probability density of $\boldsymbol{x}=\mathbf{0}$ being non-invasible is simply

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\boldsymbol{z}-\gamma\mathbf{1}_n)^T(\boldsymbol{x}-\gamma\mathbf{1}_n)} \quad (2.42)$$

and the condition for non-invasibility reduces to

$$\begin{aligned}\mathcal{P}(\emptyset|n) = P[z_1 < 0, \ldots, z_n < 0] &= \frac{1}{(2\pi)^{n/2}} \int d^n \boldsymbol{z} \prod_{i=1}^{n} \Theta(-z_i) e^{-\frac{1}{2}(\boldsymbol{z}-\gamma \mathbf{1}_n)^T (\boldsymbol{x}-\gamma \mathbf{1}_n)} \\ &= [\Phi(-\gamma)]^n.\end{aligned} \tag{2.43}$$

In addition, for $k = 1$ the integral over w can be actually calculated. Using that

$$\int_{-\infty}^{\infty} dw\, e^{-\frac{1}{2}w^2 - iaw}[1 - \Phi(iw)] = \sqrt{2\pi}\Theta(-a)e^{-\frac{1}{2}a^2} \tag{2.44}$$

we get

$$\mathcal{P}(\{S\}_1|n) = \frac{s+1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, \Theta(sy + v) e^{-\frac{1}{2}(s+1)^2 y^2} [\Phi(y - v)]^{n-1}, \tag{2.45}$$

or, alternatively,

$$\mathcal{P}(\{S\}_1|n) = \frac{1}{\sqrt{2\pi}} \int_{-\gamma}^{\infty} dy\, e^{-\frac{1}{2}y^2} \Big[\Phi\Big(\frac{\mu y}{\alpha} - v\Big)\Big]^{n-1}. \tag{2.46}$$

Numerical evaluation of the double integral

Equation (2.40) can be evaluated numerically via a Fast Fourier Transform (FFT). We can express it as

$$\mathcal{P}(\{S\}_k|n) = \frac{k+s}{2\pi} \int_{-\infty}^{\infty} dy\, e^{-\frac{1}{2}(k+2s)(y^2+2vy)+svy} [\Phi(y)]^{n-k} \widehat{F}(-s(y+v) - kv; k) \tag{2.47}$$

where $\widehat{F}(x; k)$ is the Fourier transform over w of the complex function

$$F(w; k) = e^{-\frac{1}{2}kw^2}[1 - \Phi(iw)]^k \tag{2.48}$$

and the Fourier Transform of the function $F(t; k)$ is defined as $\widehat{F}(x; k) := \int_{-\infty}^{\infty} dt\, f(t; k) e^{-itx}$. Then we first calculate $\widehat{F}(x; k)$ via a FFT algorithm. For that purpose, we assume that $f(t)$

is approximately equal to zero outside the interval $(-T/2, T/2)$ and sample t at m equally spaced points separated a distance $\delta = T/m$ (m is even), so that $t_j = (j - m/2)\delta$, $0 \le j < m$. Then

$$\widehat{F}(x_\ell; k) = \int_{-\infty}^{\infty} dt\, F(t;k) e^{-itx_\ell} \approx \int_{-T/2}^{T/2} dt\, F(t;k) e^{-itx_\ell} \approx \delta \sum_{j=0}^{m-1} F(t_j;k) e^{-itx_\ell}. \quad (2.49)$$

If $x_\ell = 2\pi(\ell - m/2)/T = 2\pi(\ell - m/2)/(m\delta)$, the last expression can be written in terms of the Discrete Fourier Transform, $D_\ell\big(\{z_j\}\big) = \sum_{j=0}^{m-1} z_j e^{-2\pi i j\ell/m}$, as

$$\widehat{F}(x_\ell; k) = \delta e^{i\pi(\ell - m/2)} \sum_{j=0}^{m-1} F(t_j;k) e^{i\pi j(1-2\ell/m)} = (-1)^{\ell - m/2} \delta D_\ell \Big[\{(-1)^j F(t_j;k)\}\Big] \quad (2.50)$$

where $0 \le \ell < m$. Once we have calculated $\widehat{F}(x_\ell; k)$ over the set of sampling points, we interpolate to evaluate numerically the transform at an arbitrary point [see equation (2.47)]. For numerical evaluation over a finite interval, equation (2.47) is more conveniently expressed by changing to the variable $z = \Phi(y)$ as

$$\mathcal{P}(\{S\}_k|n) = \frac{k+s}{\sqrt{2\pi}} \int_0^1 dz\, \widehat{F}\Big(-s\Big[\Phi^{-1}(z) + v\Big] - kv; k\Big) \times e^{-\frac{1}{2}[\Phi^{-1}(z)]\{2v(k+s) + [\Phi^{-1}(z)](k-1+2s)\} + (n-k)\log z}. \quad (2.51)$$

For $k = 1$ from (2.46) we derive the expression

$$\mathcal{P}(\{S\}_1|n) = (s+1) \int_{\Phi(-\gamma)}^1 dz\, e^{-\frac{1}{2}(s+1)^2[\Phi^{-1}(z)+v]^2 + \frac{1}{2}[\Phi^{-1}(z)]^2 + (n-1)\log z}. \quad (2.52)$$

Probability of coexistence

Assuming diagonal stability, the probability of observing k species in stable coexistence out of a pool of n species is given by

$$P(k|n) = \binom{n}{k} \mathcal{P}(\{S\}_k|n), \tag{2.53}$$

with $\mathcal{P}(\{S\}_k|n)$ given by equation (2.41). We now approximate $\mathcal{P}(\{S\}_k|n)$ for large n in order to obtain an analytical formula for the distribution, as well as the mode of the distribution $k^\star$.

We use the saddle point technique from statistical mechanics to evaluate integrals of the form $\int d^n\boldsymbol{u} e^{-nh(\boldsymbol{u})} k(\boldsymbol{u})$ for n large. We define q through $k = qn$ and regard q as a continuous, finite variable such that $0 \leq q \leq 1$. Then equation (2.41) can be written as

$$\mathcal{P}(\{S\}_k|n) = \frac{k+s}{2\pi i} \int_{-\infty}^{\infty} dy \int_{\Gamma} dw\, e^{-sy^2+syw} e^{-n\hat{h}(y,w;q,v)}, \tag{2.54}$$

where

$$\hat{h}(y,w;q) = \frac{q}{2}\Big(y^2 - w^2\Big) - q\log[1-\Phi(w-v)] - (1-q)\log\Phi(y-v). \tag{2.55}$$

In the limit $n \to \infty$, we assume q to take a fixed value (which will be associated to any possible value that k can take in the range $0 \leq k \leq n$). To calculate the limit correctly, at this point we assume that interactions scale with n as $\mu = \hat{\mu}/n$. In this way, the total interaction strength for any species is independent of n. Otherwise, since μ only enters in equation (2.54) through the combination $s = \frac{\alpha}{\mu} - 1$, if we do not assume the scaling in the limit for $n \to \infty$ any dependence on interaction strengths will be lost for n large. Therefore

we write $s = nu - 1$, where $u := \alpha/\hat{\mu}$ and equation (2.54) becomes

$$\mathcal{P}(\{S\}_k|n) = \frac{k + nu - 1}{2\pi i} \int_{-\infty}^{\infty} dy \int_{\Gamma} dw\, e^{y^2 - yw} e^{-nh(y,w;\boldsymbol{\sigma})}, \tag{2.56}$$

where we use the shorthand $\boldsymbol{\sigma} := (q, u, v)$ and

$$h(y, w; \boldsymbol{\sigma}) = \frac{q}{2}\Big(y^2 - w^2\Big) - q \log[1 - \Phi(w - v)] - (1 - q) \log \Phi(y - v) + uy^2 - uyw. \tag{2.57}$$

In this limit of large n, the exponential function $e^{-nh(y,w;\boldsymbol{\sigma})}$ is very peaked around the global minimum of the real part of $h(y, w; \boldsymbol{\sigma})$. Then we can evaluate the integral by approximating the exponent up to second order around the minimum. Note also that w is a complex variable and h is an analytic function of w. Then the Cauchy-Riemann condition holds (i.e., the real part of h satisfies the Laplace equation) and the minimum of $\Re(h)$ calculated along the integration path Γ is given by the maximum of $\Re(h)$ when w is regarded as a real variable. Then we expect a saddle point in the real (y, w) plane.

The conditions for the critical point form a coupled system of non-linear equations for y and w as functions of $\boldsymbol{\sigma}$:

$$\begin{aligned}
\frac{\partial h}{\partial y} &= qy - (1 - q)\frac{\Phi'(y - v)}{\Phi(y - v)} + 2uy - uw = qy - (1 - q)\frac{e^{-(y-v)^2/2}}{\sqrt{2\pi}\Phi(y - v)} + 2uy - uw = 0, \\
\frac{\partial h}{\partial w} &= -qw + q\frac{\Phi'(w - v)}{1 - \Phi(w - v)} - uy = -qw + q\frac{e^{-(w-v)^2/2}}{\sqrt{2\pi}[1 - \Phi(w - v)]} - uy = 0.
\end{aligned} \tag{2.58}$$

This system can be solved numerically for each tuple $\boldsymbol{\sigma} = (q, u, v)$, yielding the functions $y^\star(\boldsymbol{\sigma})$ and $w^\star(\boldsymbol{\sigma})$ as the coordinates of the critical point. We now expand $h(y, w; \boldsymbol{\sigma})$ around these coordinates point up to second order. Using that $\Phi''(y - v) = -(y - v)\Phi'(y - v)$ and

the conditions (2.58), we find

$$\begin{aligned}
\left.\frac{\partial^2 h}{\partial y^2}\right|_{\substack{y=y^\star\\w=w^\star}} &= 2u+q+(1-q)\left[y-v+\frac{\Phi'(y-v)}{\Phi(y-v)}\right]\left.\frac{\Phi'(y-v)}{\Phi(y-v)}\right|_{\substack{y=y^\star\\w=w^\star}}\\
&= 2u+q+(2uy^\star+qy^\star-uw^\star)\left(-v+\frac{y^\star-u(w^\star-2y^\star)}{1-q}\right),\\
\left.\frac{\partial^2 h}{\partial w^2}\right|_{\substack{y=y^\star\\w=w^\star}} &= -q+q\left[-w+v+\frac{\Phi'(w-v)}{1-\Phi(w-v)}\right]\left.\frac{\Phi'(w-v)}{1-\Phi(w-v)}\right|_{\substack{y=y^\star\\w=w^\star}}\\
&= -q+(uy^\star+qw^\star)\left(v+\frac{uy^\star}{q}\right),\\
\left.\frac{\partial^2 h}{\partial y\partial w}\right|_{\substack{y=y^\star\\w=w^\star}} &= -u.
\end{aligned} \tag{2.59}$$

In Section 2.4.4 we show that the critical point obtained by solving the coupled system (2.58) is precisely a saddle point, as stated above. Therefore, up to second order around the saddle point,

$$\begin{aligned}
h(y,w;\boldsymbol{\sigma}) \approx h(y^\star,w^\star;\boldsymbol{\sigma}) &+ \frac{1}{2}\left.\frac{\partial^2 h}{\partial y^2}\right|_{\substack{y=y^\star\\w=w^\star}}(y-y^\star)^2+\\
&\frac{1}{2}\left.\frac{\partial^2 h}{\partial w^2}\right|_{\substack{y=y^\star\\w=w^\star}}(w-w^\star)^2+\left.\frac{\partial^2 h}{\partial y\partial w}\right|_{\substack{y=y^\star\\w=w^\star}}(y-y^\star)(w-w^\star).
\end{aligned} \tag{2.60}$$

Substituting the expansion into equation (2.56) and transforming the integral over Γ back into an integral over a real variable yields, up to first order in the asymptotic expansion of the exponent in powers of $1/n$, the following approximation for the probability $\mathcal{P}(\{S\}_k|n)$ that the support of the globally stable fixed point is $\{S\}_k$:

$$\mathcal{P}(\{S\}_k|n) = \frac{n(q+u)-1}{\sqrt{K(\boldsymbol{\sigma},n)}}e^{-nh(y^\star,w^\star;\boldsymbol{\sigma})+y^\star(y^\star-w^\star)}, \tag{2.61}$$

with

$$K(\boldsymbol{\sigma}, n) := (nu-1)^2 + n^2\Big[-q + (uy^\star + qw^\star)\Big(v + \frac{uy^\star}{q}\Big)\Big] \\ \times \Big[\frac{2}{n} - 2u - q - (2uy^\star + qy^\star - uw^\star)\Big(-v + \frac{y^\star - u(w^\star - 2y^\star)}{1-q}\Big)\Big]. \quad (2.62)$$

We can write equation (2.61) as

$$\mathcal{P}(\{S\}_k|n) = \frac{n(q+u)-1}{\sqrt{K(\boldsymbol{\sigma}, n)}} e^{nH(\boldsymbol{\sigma}) + G(\boldsymbol{\sigma})} \quad (2.63)$$

where

$$H(\boldsymbol{\sigma}) := \frac{q}{2}\Big(w^{\star 2} - y^{\star 2}\Big) + (1-q)\log[\Phi(y^\star - v)] + q\log[1 - \Phi(w^\star - v)] - uy^{\star 2} + uy^\star w^\star, \\ G(\boldsymbol{\sigma}) := y^\star(y^\star - w^\star). \quad (2.64)$$

We now use the Stirling's approximation to get

$$\binom{n}{qn} \approx \frac{e^{-n[q\log q + (1-q)\log(1-q)]}}{\sqrt{2\pi n q(1-q)}}. \quad (2.65)$$

According to equation (2.53), our approximation for the probability of coexistence is

$$P(k|n) = \frac{n(q+u)-1}{\sqrt{2\pi n q(1-q)K(\boldsymbol{\sigma}, n)}} e^{nF(\boldsymbol{\sigma}) + G(\boldsymbol{\sigma})}, \quad (2.66)$$

where

$$F(\boldsymbol{\sigma}) := \frac{q}{2}\Big(w^{\star 2} - y^{\star 2}\Big) + (1-q)\log[\Phi(y^\star - v)] + q\log[1 - \Phi(w^\star - v)] \\ - uy^{\star 2} + uy^\star w^\star - q\log q - (1-q)\log(1-q), \quad (2.67)$$

In the discrete distribution given by equation (2.66) we have to set $k = qn$ for $0 \leq q \leq 1$ (i.e.,

$0 \leq k \leq n$). We can reproduce the original parameterization with non-scaled interspecific interactions (μ) by changing $\hat{\mu}$ back to $n\mu$, i.e, replacing the constant u by $\frac{\alpha}{n\mu}$.

2.4.4 Classification of the critical point

In order to prove that the critical point $(y^\star, w^\star)$ obtained as the solution of Eq. (2.58) is a saddle point, we only have to show that the discriminant satisfies

$$D(y^\star, w^\star) = \left(\frac{\partial^2 h}{\partial y^2}\right)\left(\frac{\partial^2 h}{\partial w^2}\right) - \left(\frac{\partial^2 h}{\partial y \partial w}\right)^2 < 0, \tag{2.68}$$

where all the derivatives are evaluated at the critical point. From Eq. (2.59) we observe that

$$D(y^\star, w^\star) = \left(\frac{\partial^2 h}{\partial y^2}\right)\left(\frac{\partial^2 h}{\partial w^2}\right) - u^2. \tag{2.69}$$

We now show that $\frac{\partial^2 h}{\partial y^2} \geq 0$ and $\frac{\partial^2 h}{\partial w^2} \leq 0$ at the critical point for any combination of parameters $\boldsymbol{\sigma} = (q, u, v)$. This will complete the proof.

First, consider the expression in (2.59) for $\frac{\partial^2 h}{\partial y^2}$. Since $u > 0$ (recall that we study the case $\alpha < \mu < 0$ and $u = \alpha/\hat{\mu} = \alpha/(n\mu) > 0$) and $0 \leq q \leq 1$, we can write

$$\left.\frac{\partial^2 h}{\partial y^2}\right|_{\substack{y=y^\star \\ w=w^\star}} \geq (2uy^\star + qy^\star - uw^\star)\left(-v + \frac{y^\star - u(w^\star - 2y^\star)}{1-q}\right). \tag{2.70}$$

This product is positive or zero. On the one hand, according to (2.58),

$$2uy^\star + qy^\star - uw^\star = \frac{1-q}{\sqrt{2\pi}} \frac{e^{-(y^\star - v)^2/2}}{\Phi(y^\star - v)}, \tag{2.71}$$

which is obviously a non-negative quantity. On the other hand, $y^\star - u(w^\star - 2y^\star) = (1 -$

$q)y^\star + 2uy^\star + qy^\star - uw^\star$, hence

$$-v + \frac{y^\star - u(w^\star - 2y^\star)}{1-q} = y^\star - v + \frac{e^{-(y^\star - v)^2/2}}{\sqrt{2\pi}\Phi(y^\star - v)} = f_1(y^\star - v), \tag{2.72}$$

where we have defined the function $f_1(x) = x + \frac{e^{-x^2/2}}{\sqrt{2\pi}\Phi(x)}$. It increases monotonically and, as $x \to -\infty$, $f_1(x) \approx -\frac{1}{x} > 0$. Therefore $f_1(x) > 0$ for all x and we have shown that $\frac{\partial^2 h}{\partial y^2} \geq 0$.

Now, from (2.58) we obtain

$$uy^\star + qw^\star = \frac{q}{\sqrt{2\pi}} \frac{e^{-(w^\star - v)^2/2}}{1 - \Phi(w^\star - v)}. \tag{2.73}$$

Therefore we can express the term $v + uy^\star/q$ that appears in Eq. (2.59) as

$$v + \frac{uy^\star}{q} = -(w^\star - v) + \frac{e^{-(w^\star - v)^2/2}}{\sqrt{2\pi}[1 - \Phi(w^\star - v)]}. \tag{2.74}$$

Let us define the function

$$f_2(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}[1 - \Phi(x)]}. \tag{2.75}$$

Using the three equations above into (2.59) we find

$$\left.\frac{\partial^2 h}{\partial w^2}\right|_{\substack{y=y^\star \\ w=w^\star}} = -q\{1 + f_2(w^\star - v)[w^\star - v - f_2(w^\star - v)]\}. \tag{2.76}$$

Now we observe that the function

$$f_3(x) := 1 + f_2(x)[x - f_2(x)] \tag{2.77}$$

is equal to the derivative of $f_4(x) = x - f_2(x)$ with respect to x, $f_3(x) = f_4'(x)$. Therefore, to show that $\frac{\partial^2 h}{\partial w^2} \leq 0$ it is sufficient to see that $f_4(x)$ is a monotonically increasing function (hence $f_3(x) > 0$ and $\frac{\partial^2 h}{\partial w^2} = -q f_3(w^\star - v) < 0$). A simple graphical analysis for $f_4(x)$ proves

that this is indeed the case. As a consequence,

$$D(y^\star, w^\star) = \left(\frac{\partial^2 h}{\partial y^2}\right)\left(\frac{\partial^2 h}{\partial w^2}\right) - u^2 \leq -u^2 < 0 \tag{2.78}$$

and $(y^\star, w^\star)$ is a saddle pont.

In summary, we have shown that the solution $(y^\star, w^\star)$ of Eq. (2.58) is a saddle point for the function $h(y, w; \boldsymbol{\sigma})$ defined in Eq. (2.57), when w is regarded as a real variable. This implies, by the Cauchy-Riemman condition, that the real part of h has a minimum along the imaginary w axis (i.e, along the integration contour Γ). Since the saddle point is unique, it yields a global minimum for the exponent in the probability (2.56) of finding the globally stable fixed point with support $\{S\}_k$.

2.4.5 Mode of the distribution for large number of species

For large n, the mode of the distribution (2.66) is recovered at a $q^\star$ value such that F takes its maximum value. We now calculate this $q^\star$ in the limits $\alpha/\hat{\mu} \gg 1$ (the mode has to be close to 1/2) and the ecological case $\alpha/\hat{\mu} \ll 1$.

First recall that, by definition [cf. equation (2.39)], $v = \frac{\gamma s}{k+s}$. In the limit of large n, $v = \frac{\gamma u}{q+u}$ is a function of q, so we have to take into account this implicit dependence on q. We take the derivative with respect to q on equation (2.67),

$$\begin{aligned}
\frac{\partial F}{\partial q} = {} & \frac{1}{2}\Big(w^{\star 2} - y^{\star 2}\Big) + q\Big(w^\star w^{\star\prime} - y^\star y^{\star\prime}\Big) - \log\Phi(y^\star - v) \\
& + \log(1 - \Phi(w^\star - v)) + (w^\star - 2y^\star) u y^{\star\prime} \\
& + u y^\star w^{\star\prime} + (1-q)\Big(y^{\star\prime} - v'\Big)\frac{\Phi'(y^\star - v)}{\Phi(y^\star - v)} \\
& - q\Big(w^{\star\prime} - v'\Big)\frac{\Phi'(w^\star - v)}{1 - \Phi(w^\star - v)} + \log\frac{1-q}{q}.
\end{aligned} \tag{2.79}$$

Now, according to equation (2.58),

$$\begin{aligned}\frac{\Phi'(y^\star - v)}{\Phi(y^\star - v)} &= \frac{qy^\star + 2uy^\star - uw^\star}{1-q},\\ \frac{\Phi'(w^\star - v)}{1-\Phi(w^\star - v)} &= \frac{uy^\star + qw^\star}{q},\end{aligned} \tag{2.80}$$

so the derivative with respect to q simplifies to

$$\frac{\partial F}{\partial q} = \frac{1}{2}\Big(w^{\star 2} - y^{\star 2}\Big) - v(w^\star - y^\star) - \log\Phi(y^\star - v) + \log(1-\Phi(w^\star - v)) + \log\frac{1-q}{q}. \tag{2.81}$$

Setting the derivative to zero yields the condition

$$(1-q^\star)e^{w^{\star 2}/2 - vw^\star}[1-\Phi(w^\star - v)] = q^\star e^{y^{\star 2}/2 - vy^\star}\Phi(y^\star - v), \tag{2.82}$$

where the functions $y^\star(\boldsymbol{\sigma})$, $w^\star(\boldsymbol{\sigma})$ and $v(q)$ are evaluated at $q = q^\star$. On the other hand,

$$\begin{aligned}\frac{\Phi'(y^\star - v)}{\Phi(y^\star - v)} &= \frac{qy^\star + 2uy^\star - uw^\star}{1-q} = \frac{e^{-(y^\star - v)^2/2}}{\sqrt{2\pi}\Phi(y^\star - v)},\\ \frac{\Phi'(w^\star - v)}{1-\Phi(w^\star - v)} &= \frac{uy^\star + qw^\star}{q} = \frac{e^{-(w^\star - v)^2/2}}{\sqrt{2\pi}[1-\Phi(w^\star - v)]},\end{aligned} \tag{2.83}$$

hence

$$\begin{aligned}(1-q)e^{-(y^\star - v)^2/2} &= \sqrt{2\pi}\Phi(y^\star - v)(qy^\star + 2uy^\star - uw^\star),\\ qe^{-(w^\star - v)^2/2} &= \sqrt{2\pi}[1-\Phi(w^\star - v)](uy^\star + qw^\star).\end{aligned} \tag{2.84}$$

Substituting these expressions into equation (2.82) yields, after some algebra, this simple condition for the mode of the distribution, $q^\star$:

$$y^\star(q^\star, u, v(q^\star)) = w^\star(q^\star, u, v(q^\star)). \tag{2.85}$$

Then, if this condition is satisfied, equation (2.81) reduces to $\log\frac{1-\Phi(y^\star - v)}{\Phi(y^\star - v)} = \log\frac{q^\star}{1-q^\star}$,

which implies

$$\Phi(y^\star - v) = 1 - q^\star. \tag{2.86}$$

From this we get

$$y^\star(q^\star, u, v(q^\star)) = v(q^\star) + \sqrt{2}\text{erf}^{-1}(1 - 2q^\star). \tag{2.87}$$

Finally we take into account the last expression and use equation (2.85) into equation (2.58) to obtain

$$\sqrt{2}\gamma u + 2(q^\star + u)\text{erf}^{-1}(1 - 2q^\star) = \frac{e^{-\left[\text{erf}^{-1}(1-2q^\star)\right]^2}}{\sqrt{\pi}} \tag{2.88}$$

which is a transcendental equation that determines the mode of the distribution $q^\star = \frac{k^\star}{n}$ as a function of interaction strengths and growth rates. Equivalently, the transcendental condition for the mode can be expressed as

$$\frac{\alpha}{\hat{\mu}} = \frac{e^{-\left[\Phi^{-1}(1-q^\star)\right]^2/2} - \sqrt{2\pi}q^\star\Phi^{-1}(1 - q^\star)}{\sqrt{2\pi}\left[\Phi^{-1}(1 - q^\star) + \gamma\right]}, \tag{2.89}$$

with $\Phi^{-1}(q) = \sqrt{2}\text{erf}^{-1}(2q - 1)$. A simple relation arises for the curve that separates left- and right-skewed distributions by choosing the mode to be $q^\star = \frac{1}{2}$:

$$\frac{\alpha\gamma}{\hat{\mu}} = \frac{1}{\sqrt{2\pi}}. \tag{2.90}$$

In terms of the original (non-scaled) parameterization, this expression becomes

$$\frac{\alpha\gamma}{\mu} = \frac{n}{\sqrt{2\pi}} \tag{2.91}$$

via the substitution $\hat{\mu} \to n\mu$.

In the limit of small interaction strengths ($\hat{\mu} \ll \alpha$) of the mean zero case ($\gamma = 0$), condition (2.89) reduces to

$$\frac{k^\star}{n} \approx \frac{1}{2} - \frac{1}{2\pi}\frac{\hat{\mu}}{\alpha} + \frac{1}{4\pi}\left(\frac{\hat{\mu}}{\alpha}\right)^2, \tag{2.92}$$

which reproduces the expected (binomial) behavior.

2.4.6 Truncated-Gaussian distributed rates

In this section we analyze the case that growth rates are drawn from a from a truncated Gaussian distribution,

$$P(\boldsymbol{r}) = \frac{1}{Z_n} \exp\left(-\sum_{i=1}^{n} \frac{(r_i - \gamma)^2}{2} \right) \prod_{j=1}^{n} \Theta\big(r_j\big) \ , \tag{2.93}$$

so that every rate $r_j > 0$ for $j = 1, \ldots, n$ (Z_n is a suitable normalization constant). Then we can express the probability $\mathcal{P}(\{S\}_k|\boldsymbol{A})$ of observing the globally stable fixed point with support $\{S\}_k$ in a simple form:

$$\mathcal{P}_T(\{S\}_k|\boldsymbol{A}) \equiv \int d^k\boldsymbol{x} \left(\prod_{i=1}^{k} \Theta(x_i) \right) \int d^{n-k}\boldsymbol{z} \left(\prod_{j=k+1}^{n} \Theta(-z_j) \right) f_T(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{A}) \ . \tag{2.94}$$

where

$$f_T(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{A}) = \frac{|\det \boldsymbol{A}^{(s)}|}{Z_n} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\xi})^T \boldsymbol{G}(\boldsymbol{x} - \boldsymbol{\xi}) - \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{A}^{(ns)}\boldsymbol{x} - \gamma \mathbf{1}_{n-k}\|^2 \right) \\ \times \prod_{i=1}^{k} \Theta\Big(-(\boldsymbol{A}^{(s)}\boldsymbol{x})_i \Big) \prod_{j=k+1}^{n} \Theta\Big(z_j - (\boldsymbol{A}^{(ns)}\boldsymbol{x})_j \Big) \ . \tag{2.95}$$

We focus on the rank-one competitive case: $\boldsymbol{A}^{(s)} = (\alpha - \mu)\boldsymbol{I}_k + \mu \mathbf{1}_k \mathbf{1}_k^T$, $\boldsymbol{A}^{(ns)} = \mu \mathbf{1}_{n-k} \mathbf{1}_k^T$ for $\alpha < \mu < 0$. Then

$$(\boldsymbol{A}^{(s)}\boldsymbol{x})_i = (\alpha - \mu)x_i + \mu(\mathbf{1}_k^T \boldsymbol{x}) = \alpha x_i + \mu \sum_{\substack{s=1 \\ s \neq i}}^{k} x_s. \tag{2.96}$$

Since Eq. (2.94) forces that $x_i > 0$, and α and μ are both negative, we find that $-(\boldsymbol{A}^{(s)}\boldsymbol{x})_i$

is always positive, i.e., it holds that

$$\Theta\Big(-(\boldsymbol{A}^{(s)}\boldsymbol{x})_i\Big)\Theta(x_i) = \Theta(x_i). \tag{2.97}$$

On the other hand, $\mu(\mathbf{1}_k^T\boldsymbol{x}) < 0$ and we can express

$$\Theta\Big(z_j - (\boldsymbol{A}^{(ns)}\boldsymbol{x})_j\Big)\Theta(-z_j) = \Theta\Big(z_j - \mu(\mathbf{1}_k^T\boldsymbol{x})\Big) + \Theta(-z_j) - 1. \tag{2.98}$$

Now, we apply the same changes of variable leading to Eq. (2.29). Then we can write

$$\begin{aligned}\mathcal{P}_T(\{S\}_k|n) = \frac{|\alpha-\mu|^{k-1}|\alpha+(k-1)\mu|}{Z_n}\int d^k\boldsymbol{x}\prod_{i=1}^{k}\Theta(x_i+\xi^{(k)})e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x}}\\ \times\int d^{n-k}\boldsymbol{z}\prod_{j=k+1}^{n}\Big[\Theta(-z_j-\gamma-k\mu\xi^{(k)})+\Theta(z_j-\mu(\mathbf{1}_k^T\boldsymbol{x})+\gamma)-1\Big]e^{-\frac{1}{2}\|\boldsymbol{z}-\mu(\mathbf{1}_k^T\boldsymbol{x})\mathbf{1}_{n-k}\|^2}\,.\end{aligned} \tag{2.99}$$

Let $\mathcal{K} = \{1,\ldots,n-k\}$. Expanding the product we get

$$\begin{aligned}&\prod_{j=k+1}^{n}\Big\{\Theta(-z_j-\gamma-k\mu\xi^{(k)})+\Big[\Theta(z_j-\mu(\mathbf{1}_k^T\boldsymbol{x})+\gamma)-1\Big]\Big\}\\ &\quad=\sum_{\ell=0}^{n-k}\sum_{\substack{p\in C_\ell^{n-k}\\ b=\mathcal{K}\setminus p}}\prod_{j=1}^{\ell}\Theta(-z_{p(j)+k}-\gamma-k\mu\xi^{(k)})\prod_{i=1}^{n-k-\ell}\Big[\Theta(z_{b(i)+k}-\mu(\mathbf{1}_k^T\boldsymbol{x})+\gamma)-1\Big],\end{aligned} \tag{2.100}$$

where $p = (p(1),\ldots,p(\ell))$ is a combination of ℓ elements taken from $\mathcal{K}$, $p \in C_\ell^{n-k}$, and b is formed by the remaining elements of the set, $b = \{1,\ldots,n-k\}\setminus p$. Without loss of generality,

since integrals are invariant under changes of indices in variable $\boldsymbol{z}$, we can decompose

$$\begin{aligned}\mathcal{P}_T(\{S\}_k|n) = \frac{|\alpha-\mu|^{k-1}|\alpha+(k-1)\mu|}{Z_n}\int d^k\boldsymbol{x}\prod_{i=1}^{k}\Theta(x_i+\xi^{(k)})e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x}} \\ \times\sum_{\ell=0}^{n-k}\binom{n-k}{\ell}\prod_{j=k+1}^{\ell+k}\int dz_j\Theta(-z_j-\gamma-k\mu\xi^{(k)})e^{-\frac{1}{2}\left[z_j-\mu(\mathbf{1}_k^T\boldsymbol{x})\right]^2} \\ \times\prod_{i=\ell+k+1}^{n}\int dz_i\Big[\Theta(z_i-\mu(\mathbf{1}_k^T\boldsymbol{x})+\gamma)-1\Big]e^{-\frac{1}{2}\left[z_i-\mu(\mathbf{1}_k^T\boldsymbol{x})\right]^2}.\end{aligned} \tag{2.101}$$

Note now that

$$\int_{-\infty}^{\infty}dz\Big[\Theta(z-\mu(\mathbf{1}_k^T\boldsymbol{x})+\gamma)-1\Big]e^{-\frac{1}{2}\left[z-\mu(\mathbf{1}_k^T\boldsymbol{x})\right]^2} = -\sqrt{2\pi}\Phi(-\gamma). \tag{2.102}$$

Therefore we can decompose $\mathcal{P}(\{S\}_k|n)$ as the sum

$$\mathcal{P}_T(\{S\}_k|n) = \frac{(2\pi)^{n/2}}{Z_n}\sum_{\ell=0}^{n-k}\binom{n-k}{\ell}[-\Phi(-\gamma)]^{n-\ell-k}\mathcal{P}(\{S\}_k|\ell+k), \tag{2.103}$$

where $\mathcal{P}(\{S\}_k|n)$ is precisely the expression (2.29) obtained for the non-truncated Gaussian distribution. According to Eq. (2.41),

$$\mathcal{P}(\{S\}_k|\ell+k) = \frac{k+s}{2\pi i}\int_{-\infty}^{\infty}dy\int_{\Gamma}dw\,e^{-\frac{1}{2}(k+2s)y^2+syw+\frac{1}{2}kw^2}[1-\Phi(w-v)]^k[\Phi(y-v)]^\ell. \tag{2.104}$$

We introduce (2.104) into (2.103) and use the binomial expansion

$$\sum_{\ell=0}^{n-k}\binom{n-k}{\ell}[-\Phi(-\gamma)]^{n-\ell-k}[\Phi(y-v)]^\ell = [\Phi(y-v)-\Phi(-\gamma)]^{n-k} \tag{2.105}$$

to get the probability $\mathcal{P}_T(\{S\}_k|n)$ expressed as a double integral,

$$\mathcal{P}_T(\{S\}_k|n) = \frac{(2\pi)^{n/2-1}(k+s)}{i\ Z_n}\int_{-\infty}^{\infty} dy \int_{\Gamma} dw\, e^{-\frac{1}{2}(k+2s)y^2+syw+\frac{1}{2}kw^2} \times [1-\Phi(w-v)]^k[\Phi(y-v)-\Phi(-\gamma)]^{n-k}. \quad (2.106)$$

Note that the only difference with Eq. (2.41) is the term $\Phi(-\gamma)$ that appears in the last factor of the integrand. Hence we can easily extend the saddle-point calculation for the truncated Gaussian case. The probability $P_T(k|n) = \binom{n}{k}\mathcal{P}_T(\{S\}_k|n)$ that the support has cardinality k in this case can be written, up to a normalization factor and sub-leading corrections, as $P_T(k|n) \sim e^{nF_T(\boldsymbol{\sigma})}$, where

$$F_T(\boldsymbol{\sigma}) := \frac{q}{2}\Big(w^{\star 2} - y^{\star 2}\Big) + (1-q)\log[\Phi(y^\star - v) - \Phi(-\gamma)] + q\log[1-\Phi(w^\star - v)] - uy^{\star 2} + uy^\star w^\star - q\log q - (1-q)\log(1-q). \quad (2.107)$$

We can compare the mode of the distribution for the truncated and the purely Gaussian cases. The calculation of the mode follows the same steps of the Gaussian case. The equations for the saddle point $(y^\star, w^\star)$ are now

$$\begin{aligned} qy - (1-q)\frac{\Phi'(y-v)}{\Phi(y-v)-\Phi(-\gamma)} + 2uy - uw = 0, \\ qw - q\frac{\Phi'(w-v)}{1-\Phi(w-v)} + uy = 0. \end{aligned} \quad (2.108)$$

As can be easily checked, the condition $\frac{\partial F_T}{\partial q} = 0$ to be satisfied by the mode $q^\star$ leads to the same constraint as in the Gaussian case, $y^\star(q^\star, u, v(q^\star)) = w^\star(q^\star, u, v(q^\star))$, see Eq. (2.85). This implies that

$$\Phi(y^\star - v) = 1 - q^\star + q^\star\Phi(-\gamma), \quad (2.109)$$

which reduces to the Gaussian-case condition for the mode in the limit of large γ, where

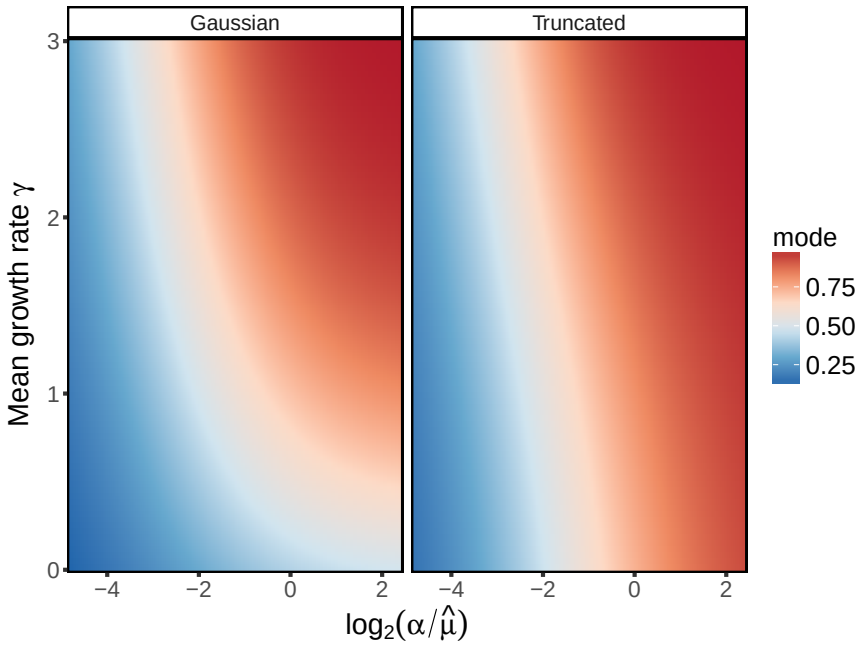

Figure 2.4: Comparison between the modes for purely Gaussian and truncated-Gaussian distributed growth rates.

both the truncated and the Gaussian distributions tend to almost overlap. Finally, after the same algebraic manipulations in the condition above we obtain the following non-linear equation that determines the mode in the truncated-Gaussian case:

$$[1-\Phi(-\gamma)]\Big[\sqrt{2}\gamma u + 2(q^\star+u)\mathrm{erf}^{-1}(1-2q^\star+2q^\star\Phi(-\gamma))\Big] = \frac{1}{\sqrt{\pi}}e^{-\left[\mathrm{erf}^{-1}(1-2q^\star+2q^\star\Phi(-\gamma))\right]^2}. \tag{2.110}$$

Figure 2.4 shows the most probable number of coexisting species obtained for the Gaussian and the truncated Gaussian distributions as function of the parameters γ and $\alpha/\hat{\mu}$. We observe that the expected values for both cases are roughly the same for $\gamma \gtrsim 1$.

2.4.7 Final communities

Figure 2.4.7 shows the properties of the parameters of the communities found after the dynamical pruning, for an starting community of 1000 species, and a final community com-

prising 472 species. As proposed recently [83], the matrix of interactions in the pruned community is a random subset of the original. On the other hand, the distribution of growth rates changes in a nontrivial way, with a larger mean and positive skewness. This change, as explained in the discussion, is related to the interplay between the negative diagonal that we need to add to the matrix in order to ensure stability and the sign and magnitude of the interespecific interactions that each species is assigned, which in the end pushes the growth rates towards the right. As a limiting case to this behavior one can see that if the matrix is purely diagonal, then because we require stability all its entries are negative and the distribution of growth rates will be a truncated version of the original one.

2.4.8 Numerical simulations

In this section, we detail the numerical simulations we used to corroborate our argument, and extend it to cases in which a direct analytic computation is unfeasible. We start by illustrating the Lemke-Howson algorithm that can be used to efficiently search for the non-invasible solution. Applying this algorithm, we were able to determine the non-invasible solution of a system without the need to integrate the dynamics numerically. Then, we detail the parameters for the numerical simulations—how are the matrices constructed, and how the network structure is introduced.

Lemke-Howson algorithm

Because of the equivalence between the Lotka-Volterra and the replicator equation [56], the non-invasible solution in the diagonally stable regime *is* the *unique* symmetric Nash equilibrium for the replicator dynamics in which the last element of the solution is played with non-zero probability [58] (this last element can be interpreted as "the environment" when moving from LV with n equations to a replicator system with $n+1$ equations). We use the Lemke-Howson algorithm [72] to find such a solution. This algorithm is based on

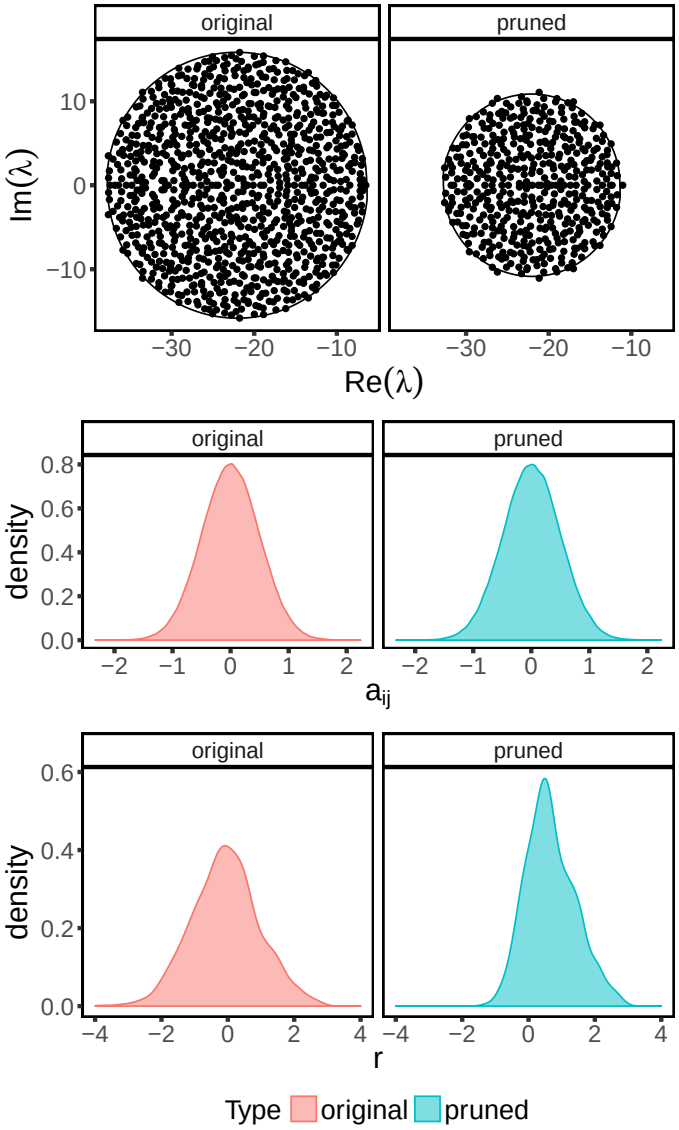

Figure 2.5: Comparison between the properties of the original community with $n = 1000$ and the final community, after dynamical pruning, comprising $n = 472$ species. The first row shows the eigenvalue distribution of the matrix of interaction A; for a matrix in which the entries are i.i.d. samples from a distribution, we expect the eigenvalues to be approximately uniformly distributed in a circle in the complex plane, whose radius depends on the size of the system and the variance of the distribution [4, 81]. In the second panel, we show that indeed the distribution of the off-diagonal elements of A is the same before/after dynamics. Finally, in the third panel we show that instead the distribution of growth rates changes non trivially.

exploring the vertices of the following polytope:

$$P = \{\boldsymbol{z} \in \mathbb{R}^n | \boldsymbol{z} \geq \mathbf{0}, \boldsymbol{C}\boldsymbol{z} \leq \mathbf{1}\}, \tag{2.111}$$

where $\boldsymbol{C}$ is a positive payoff matrix of an $n \times n$ symmetric game—the positivity of the payoffs can be assumed without loss of generality, because adding a suitable constant to all the elements of the payoff matrix does not affect the dynamics.

We say that $\boldsymbol{z} \in P$ has label k if $z_k = 0$ and label $-k$ if $(\boldsymbol{C}\boldsymbol{z})_k = 1$. Let us assume that P is simple (which holds almost surely in the cases we explore), that is, each vertex is adjacent to exactly n facets—a facet is defined by setting to equality one of the inequalities defining the polytope. Say that $\boldsymbol{z}$ represents strategy k if either it has label k or $-k$, then because of the simplicity assumption any $\boldsymbol{z}$ that represents all strategies is either $\mathbf{0}$ or the normalized vector $\widehat{\boldsymbol{z}} = \boldsymbol{z}/\sum_i z_i$ is a *symmetric Nash equilibrium* for the game.

In order to find the solution we move around the vertices of P starting from $\boldsymbol{v_0} = \mathbf{0}$ using a tableaux $T : \boldsymbol{r} = \mathbf{1} - \boldsymbol{C}\boldsymbol{z}$ with a slack variable $\boldsymbol{r}$. Say that r_k is in the basis for a vertex $\boldsymbol{v} \in P$ if and only if $\boldsymbol{v}$ does not have label $-k$, and z_k is in the basis if and only if $\boldsymbol{v}$ does not have label k. Then $\boldsymbol{v_0}$ has basis $\{r_1, \ldots, r_n\}$, bring z_n to the basis and by the min. ratio rule—i.e., by looking at the ratio between the free variable (in this case 1) and the coefficients of z_n in the tableaux—choose r_k to leave the basis and proceed to an adjacent vertex $\boldsymbol{v_1}$. In the next iteration bring z_k to the basis and move to an adjacent vertex $\boldsymbol{v_2}$. We keep repeating this process until we get to a vertex $\boldsymbol{v}$ which represents all strategies, that is, $\boldsymbol{v}$ is a Nash equilibrium which moreover will have z_n in the basis (since by construction the process will stop when the element leaving the basis is r_n). Because of the simplicity assumption the process is going to terminate, having to do in the worst case 2^n iterations. As it often happens, this worst-case scenario is never found in practice, making the algorithm efficient.

Let us illustrate this ideas by a simple example. Take the Lotka-Volterra system with

interactions

$$\boldsymbol{A} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \tag{2.112}$$

and intrinsic growth rates:

$$\boldsymbol{r} = \begin{pmatrix} -1 \\ 3 \end{pmatrix} \tag{2.113}$$

We build the payoff matrix:

$$\boldsymbol{C} = \begin{pmatrix} -2 & 1 & -1 \\ 1 & -2 & 3 \\ 0 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 4 & 2 \\ 4 & 1 & 6 \\ 3 & 3 & 3 \end{pmatrix}, \tag{2.114}$$

where we have added a constant to all entries to make them all positive. At the beginning of the algorithm we have the following tableaux:

$$\begin{aligned} r_1 &= 1 - z_1 - 4z_2 - 2z_3, \\ r_2 &= 1 - 4z_1 - z_2 - 6z_3, \\ r_3 &= 1 - 3z_1 - 3z_2 - 3z_3. \end{aligned} \tag{2.115}$$

We now bring z_3 into the basis, and by the min. ratio rule: the ratio of 1 and the coefficients of z_3, r_2 should leave the basis and the updated tableaux is:

$$\begin{aligned} r_1 &= \frac{2}{3} + \frac{1}{3}z_1 - \frac{11}{3}z_2 + \frac{1}{3}r_2, \\ z_3 &= \frac{1}{6} - \frac{4}{6}z_1 - \frac{1}{6}z_2 - \frac{1}{6}r_2, \\ r_3 &= \frac{1}{2} - z_1 - \frac{5}{2}z_2 + \frac{1}{2}r_2. \end{aligned} \tag{2.116}$$

Now z_2 enters the basis, and in this case r_1 leaves from the basis:

$$\begin{aligned} z_2 &= \frac{2}{11} + \frac{1}{11}z_1 - \frac{3}{11}r_1 + \frac{1}{11}r_2, \\ z_3 &= \frac{3}{22} - \frac{15}{22}z_1 + \frac{1}{22}r_1 - \frac{2}{11}r_2, \\ r_3 &= \frac{1}{22} - \frac{27}{22}z_1 + \frac{15}{22}r_1 + \frac{3}{11}r_2. \end{aligned} \tag{2.117}$$

We bring z_1 into the basis and then we are done because r_3 leaves the basis in this case. So the Nash equilibrium for this game has full support. The final state of the tableaux is :

$$\begin{aligned} z_1 &= \frac{1}{27} + \frac{15}{27}r_1 + \frac{2}{9}r_2 - \frac{22}{27}r_3, \\ z_2 &= \frac{5}{27} - \frac{6}{27}r_1 + \frac{1}{9}r_2 - \frac{2}{27}r_3, \\ z_3 &= \frac{1}{9} - \frac{1}{3}r_1 - \frac{1}{3}r_2 + \frac{15}{27}r_3. \end{aligned} \tag{2.118}$$

By normalizing the free elements in the final tableaux we also get the values at equilibrium, which in this case is $(1/9, 5/9, 3/9)$. Because the last element is positive, then the two species coexist, the second with an equilibrium value that is five times as large as the first.

Sampling the matrices and growth rates

In the following we give the details of the construction of the matrices and growth rates for the cases we explored. For each case we repeat the process 2×10^5 times.

Mean zero

We sample the entries of $\boldsymbol{B}$ in pairs, (B_{ij}, B_{ji}) for $j \neq i$ from a bivariate Normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a covariance matrix with diagonal 1 and off-diagonal ρ. The diagonal elements B_{ii} are chosen from a standard Normal distribution $N(0, 1)$. We then calculate the leading eigenvalue of $\boldsymbol{B} + \boldsymbol{B}^T$: $\lambda_M = \max_\lambda(\Re(\lambda(\boldsymbol{B} + \boldsymbol{B}^T)))$. We define $\boldsymbol{A} = \boldsymbol{B} - d\boldsymbol{I}$, where d is a constant sufficient to make $\boldsymbol{A} + \boldsymbol{A}^T$ negative definite. More precisely, we choose

$d = -\lambda_M - 10^{-6}$ (so that the matrix $\boldsymbol{A} + \boldsymbol{A}^T$ is barely stable). The entries of $\boldsymbol{r}$ are sampled from a standard Normal distribution $N(0, 1)$.

Mean non zero

In this case, the entries of the matrix are fixed and we choose each entry of $\boldsymbol{r}$ from a Normal distribution $N(\gamma, 1)$.

Adding Structure

In order to include a network structure, we generate an adjacency matrix $\boldsymbol{G}$ with a desired connectance level C (we used $C = 0.1$ and $C = 0.25$) and all diagonal elements set to one. In the case of a power-law structure, we use the `sample_fitness_pl` function from the `igraph` package in `R` with an exponent of 2. For the modular and bipartite structures we split the matrix in two blocks, and arrange the connectance levels within and among them such that one is higher than the other—in particular we require two parameters b_r and c_r that determine the ratio of the size among the blocks and the ratio of the connectance within and among blocks (e.g. $c_r > 1$ for a modular structure). The values used were $b_r = 1/3$ for both cases, with $c_r = 3$ for modular, and $c_r = 1/3$ for bipartite. This adjacency matrix is then multiplied element-wise to our original matrix. The results are presented in Figure 4.1 in the main text as well as in Figure 2.6.

In the mean-zero case the matrix is made negative definite by the same process described above.

In the mean non-zero case the fully connected matrix is by construction negative definite ($\alpha < \mu < 0$) but when we add structure we need to restrict the values of μ that keep the negative definiteness.

The prediction shown in Figure 4.3 of the main text is the mode of a fully connected system using the rescaled μ: $\hat{\mu} = n\mu C$.

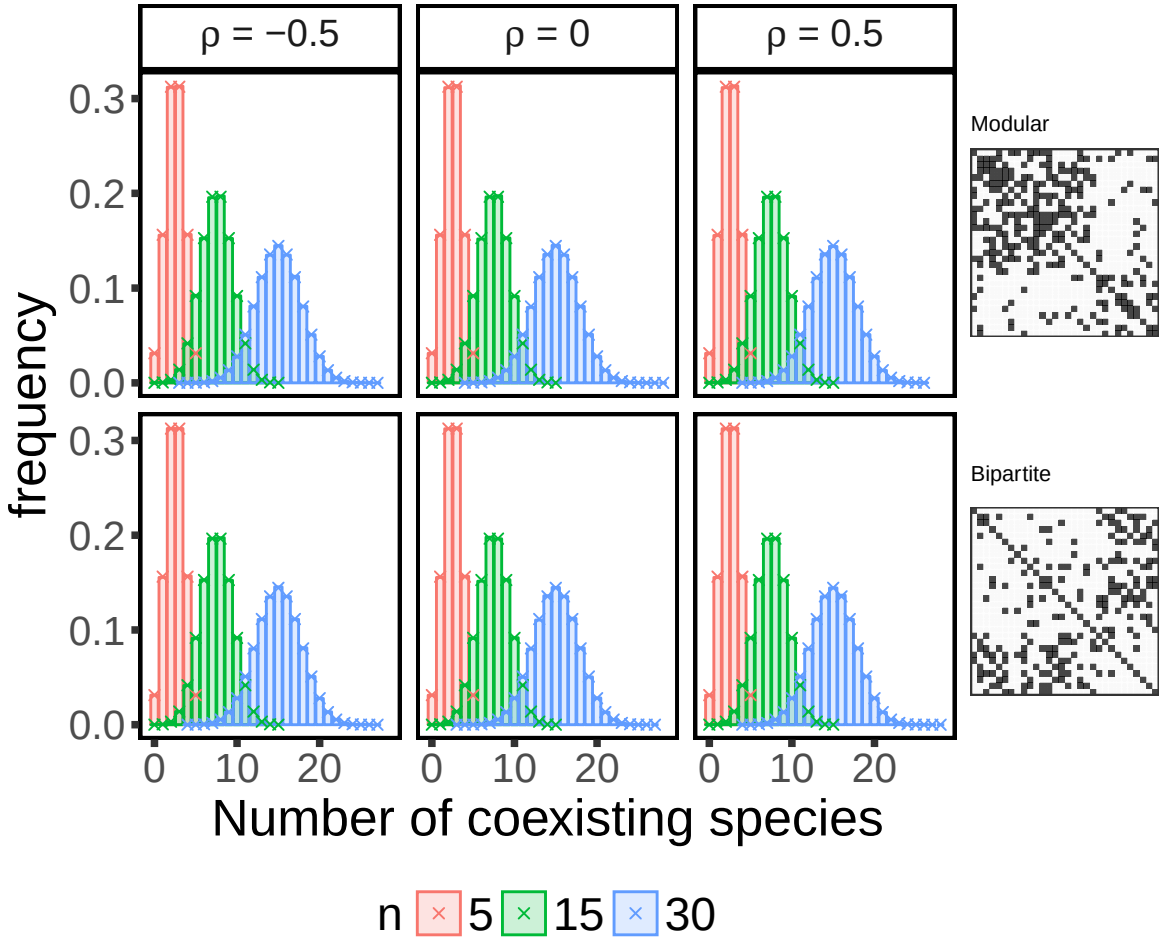

Figure 2.6: As Figure 1 of the main text, but with modular (top) or "anti-modular" (i.e., close to bipartite, bottom) structures.

2.4.9 Assembly

So far, we have described the dynamical process associated with equation 1 of the main text when starting with all n species present. A different view of the problem is to take our original n-dimensional system as a *species pool*, and from that derive the possible states to which one can arrive by adding *one species at a time*. This define a directed graph in which the nodes are the feasible states, and the edges represent invasion events connecting the two states (a subset of which is shown for example in Figure 4.2). In this section we present numerical evidence that suggests that, in the regime of diagonal stability, one can find sets of persistent species satisfying equation 8 of the main text which cannot be assembled (Figure 2.7). In such cases, our end-state with k species cannot be built by adding a species at a time. The probability of finding such a case, however, decreases rapidly with k: when our final community has many species, the probability of finding at least one assembly pathway to build the community approaches one (Figure 2.8).

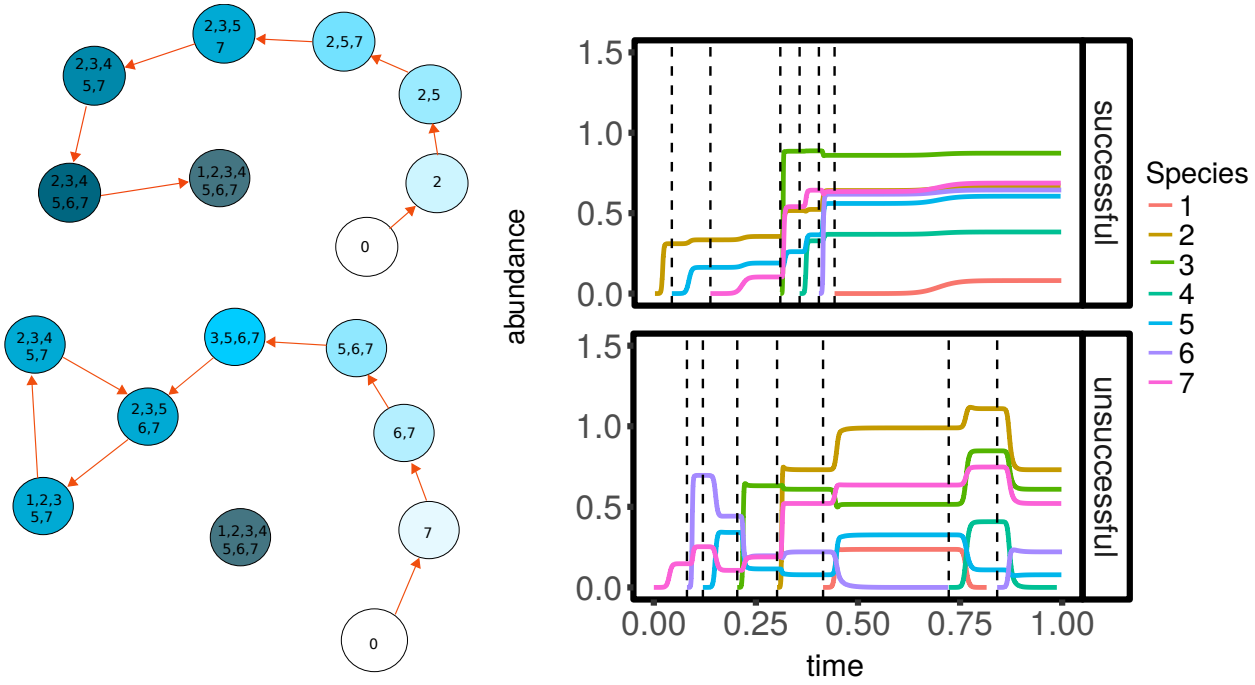

Figure 2.7: **Assembling communities one species at a time**. Top: we want to build the community with species 1, 2, ..., 7 present (darker shades for more speciose communities), by adding a species at a time. Starting from an empty system (state 0), we can try all assembly pathways in which we sequentially add one species at a time, let the dynamics unfold, and reach a new state. In this case, an assembly path exists: by adding species 2, 5, 7, 3, 4, 6 and 1 one at a time, we always recover a feasible and stable community (dynamics are shown on the right). Bottom: again, we would like to build the community with all seven species present. In this case, no assembly path exist. For example, we can add sequentially 7, 6, 5, 3, and 2, reaching a stable community with five species. At this point, however, whenever we add one of the remaining species, we lose another—the state with all species present is unreachable, even when considering all possible assembly paths.

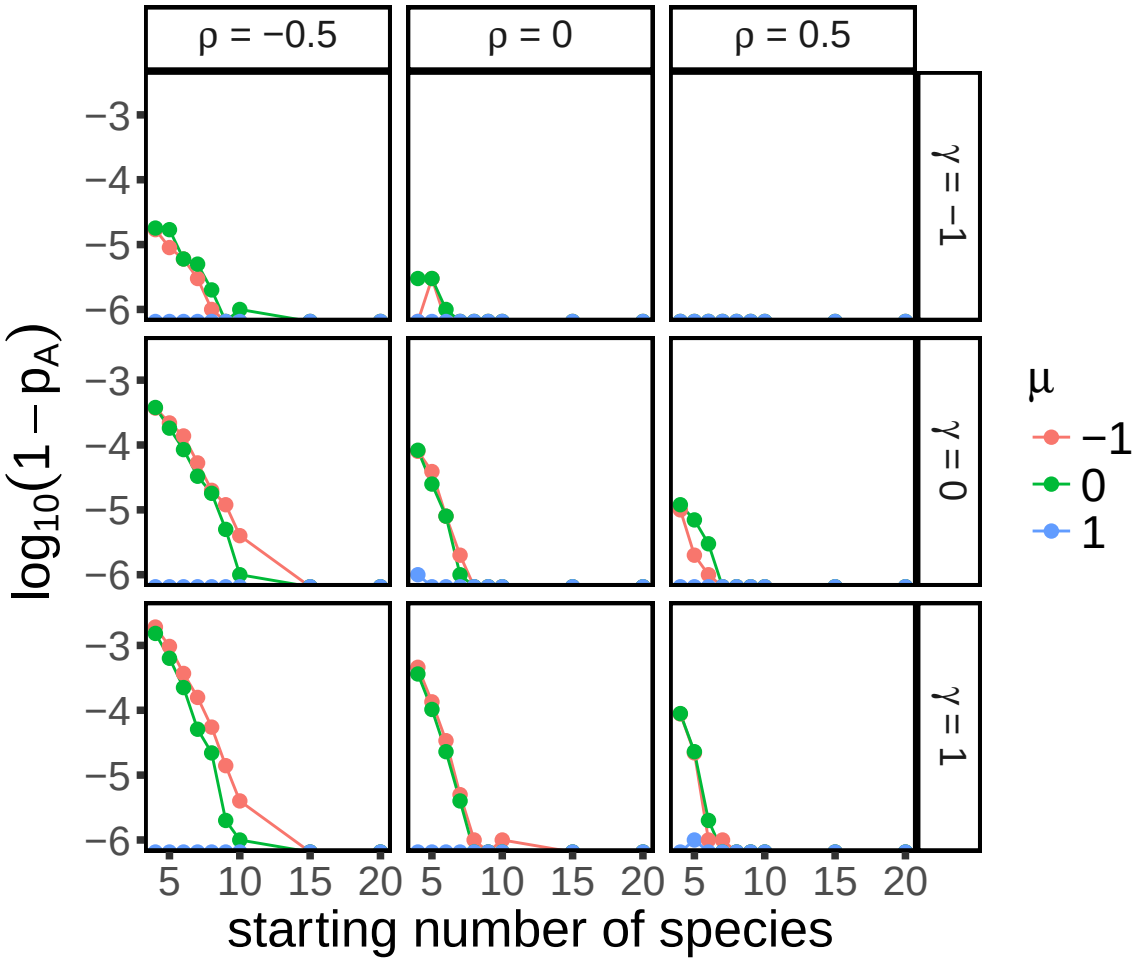

Figure 2.8: Probability p_A of finding an assembly path when starting from n species. For different means and correlations of interactions strengths, sampled from a bivariate Normal with mean μ (colors) and correlation ρ (columns), and for different values of mean intrinsic growth rates (γ, rows), we plot the probability of not finding an assembly path out of 10^6 simulations. While there is a nontrivial effect of all parameters (for example, for $\rho = 0.5$ and $\gamma = 1$ we found an assembly path for all simulations), in all cases we found that for sufficiently large n, all communities could be built by sequential invasions.

CHAPTER 3

TRACTABLE MODELS OF ECOLOGICAL ASSEMBLY

Abstract

Ecological assembly, the process by which natural communities form under ecological time-scales, remains an important and challenging problem. Recent theoretical and empirical approaches to this problem considered a particular scenario in which all species enter the local community at the same time. This "top-down assembly" approach contrasts with earlier models of community assembly, which instead examined a community built from the "bottom-up" by a sequence of invasions. Here, we study a class of assembly models, encompassing the symmetric competitive Lotka-Volterra model, for which both top-down and bottom-up assembly lead the community to the same final state. It follows that analytical predictions derived for the more manageable top-down assembly scenario map to predictions for the bottom-up case. These predictions can thus be used to design novel experiments and to further probe the properties of ecological assembly.

3.1 Introduction

> "Mother Nature, of course, does not assemble her networks by throwing n species in one go. It makes more sense to assume that she adds one species after another through successive invasions."
>
> — K. Sigmund [104]

Understanding how biodiversity emerges and is maintained is a key challenge in community ecology. When considering a local community and ecological time-scales, this process is called "ecological assembly". This process is realized through the interplay between invasions of species from outside the system and the interactions between the resident species,

driving the community towards a state in which a certain subset of species coexist. In the early days of community ecology, assembly—then called succession [35]—was envisioned as an orderly and predictable process by which the community progressed towards a climax, thereby maximizing some ecosystem functions [35, 92].

In the intervening decades, the study of assembly has unveiled a much more complicated picture [38, 69, 102, 120]—in general, assembly is anything but predictable: small changes to the order, size, and timing of invasions can result in completely different local communities [38]. The dependency of the final community on the order of arrival of invaders leads to "historical contingencies" driven by priority effects [45], and much effort went into determining whether and when these alternative histories can emerge [46, 120, 128]. Similarly, the density at which the invader enters the community can result in different outcomes, a complication compounded by the fact that the local community could be coexisting through limit cycles or chaotic attractors—and as such the invader could establish at certain times, but not others.

These complications make the study of assembly in full generality very challenging. From a mathematical standpoint, we would need to consider models accounting for each of the complications outlined above, rendering the problem intractable with current analytical tools [39]. Attacking the problem experimentally is equally complicated, as the experimental design needed to probe the space of possible assembly histories becomes unfeasible when more than a handful of species are considered [120].

To overcome these complexities and improve our understanding of assembly, we need to constrain the problem to make it tractable and yet not trivial. Here we consider two cases that seem promising in terms of mathematical tractability, while still providing a good springboard to study more complex scenarios. We call "top-down assembly" the process in which all the species in a species pool enter a system at the same time. Under these conditions, assembly amounts to the pruning of the community by the dynamics of the

system, resulting in a final set of coexisting species. We call the final set of coexisting species the "endpoint" of the assembly process, and one of our goals is to characterize these endpoints. In recent years, the study of this type of assembly gained traction from both a theoretical [10, 14, 103] and an experimental [15, 50] point of view. As succinctly stated by Karl Sigmund [104], however, top-down assembly could yield results that are quite different from those obtained when the system is built from the ground up [102, 120]. To complement top-down assembly, we therefore study a "bottom-up assembly" process in which species enter the system one at a time, and invasion events are spaced apart far enough so that the local community settles into its asymptotic configuration between any two invasions—a common assumption in early models of community assembly [69, 70]. The main goal of this work is to characterize whether and when bottom-up and top-down assembly will result, asymptotically, in the same set of assembly endpoints.

Here we provide necessary and sufficient conditions for an equivalence between top-down and bottom-up assembly. By equivalence, we mean that the set of assembly endpoints under top-down and bottom-up assembly are the same. A necessary condition for this equivalence is then that each endpoint of top-down assembly is observed during bottom-up assembly. When we constrain the local dynamics so that we only have fixed point attractors, we find that whenever assembly can be modeled as a process maximizing some quantity, bottom-up and top-down assembly are equivalent. Therefore under these conditions, assembly does indeed proceed in an orderly fashion, as envisioned by the pioneers of ecological succession. It is important to note that the equivalence between top-down and bottom-up assembly does not preclude the existence of priority effects; however, equivalence between these two processes does facilitate the exploration of conditions under which priority effects will occur [45].

We formalize the analysis of assembly by studying a mathematical representation of the process, known as the "assembly graph" G associated with a species pool [25, 54, 69]. This object captures both the possible sub-communities of coexisting species that can be formed

under top-down or bottom-up assembly, as well as the transitions between them, triggered by invasions. In this way, each realization of the assembly process corresponds to a walk on G, implying that properties of the assembly process are reflected in properties of the assembly graph. To define the assembly graph, we consider a species pool containing a given set of species, which encompasses all possible invaders that can arrive at the local system [34, 45]. The species pool would therefore correspond to the "mainland community" in an Island Biogeography model, or as the set of species present in a metacommunity in models in which local communities are connected by dispersal [34].

For the sake of concreteness and exposition, the general construction and the proofs of the statements are relegated to the Appendix. Here, we focus on a simple competitive Lotka-Volterra model and show that under certain parameterizations, assembly maximizes the total biomass of the system. This then implies that bottom-up and top-down assembly are equivalent.

3.2 Model

3.2.1 What makes the study of assembly challenging?

We model the assembly process of a local ecological community by considering the sequential arrival of invaders from a fixed regional species pool. While easily stated, the study of this problem in full generality is complicated by three main factors.

Timing of the invasion events The time at which the first species goes extinct after an invasion influences the effect of subsequent invasions. For example, take a "rock-paper-scissor" community with three species $\{x_1, x_2, x_3\}$ [2]. In this case, the three species can coexist, but no pair of species can. We have that if we start with $\{x_1, x_2\}$, the resulting community is $\{x_2\}$; similarly, $\{x_2, x_3\} \to \{x_3\}$, and $\{x_1, x_3\} \to \{x_1\}$. Consider the case in which we start with a bare environment $\{\emptyset\}$, and we introduce species x_1, which can grow

in isolation. As such, we have $\{\emptyset\} \to \{x_1\}$. Now we introduce species x_2, which would eventually send x_1 to extinction $\{x_1, x_2\} \to \{x_2\}$. If we introduce x_3 after x_1 has gone extinct we will have $\{x_2, x_3\} \to \{x_3\}$; however, if x_3 invades before this happens, then we recover the full community $\{x_1, x_2\} \to \{x_1, x_2, x_3\}$. As this simple example highlights, if the speed at which the dynamics of the local community proceed are slow enough compared to the rate of invasion, we have that several species can invade before the community has reached its asymptotic configuration. At the extreme where local dynamics are fast compared to rate of invasion, we have that each invader finds the local community at its asymptotic state (the "bottom-up" process we have outlined above); as the invasion rate increases, the system approaches a point where all the species enter the system before any extinction takes place. If an attractor is reached, then the system conforms to the "top-down" assembly regime. Increasing the invasion rate even further would result in an open system with constant immigration.

Density of the invader at the time of invasion Consider the two-species competitive Lotka-Volterra model with preemptive competition, and suppose that initially we have species x_1 resting at its carrying capacity. If x_2 invades with sufficiently low density, we find $\{x_1, x_2\} \to \{x_1\}$; on the other hand, if x_2 has sufficiently high density, we can cross the separatrix in the phase plane, leading to $\{x_1, x_2\} \to \{x_2\}$.

Type of attractor When the local community coexists at a non-fixed point attractor, the fate of the invader could be very different depending on when it is introduced. For example, a predator requiring its prey to be above a certain level would not be able to invade an oscillating system whenever prey are at low abundance, but would start growing if the invasion happened at a time when prey were abundant.

3.2.2 Ecological assembly without tears

The three features above make the study of assembly in full generality very challenging. To make the problem tractable, here we concentrate on an assembly process that sidesteps these difficulties, and yet is complex enough to generate relevant results.

Invasion events are rare We assume that the invasion rate is low enough such that, after an invasion, the local community has sufficient time to reach its asymptotic configuration before the next invader arrives. In other words, we consider that local dynamics operate at a much faster time-scale than the invasion events. Note that this choice precludes the study of certain models; for example, under these stringent conditions the rock-paper-scissor community described above would never reach the three-species configuration. While this is a strong requirement, it corresponds to assumptions routinely made in the study of population genetics (where often only the wildtype and a single mutant interact [28]), adaptive dynamics [36], and invasion analysis [1, 30].

Invaders arrive at low abundance We assume that the density of the invader is low enough so that any intraspecific competition is negligible at the time of invasion. Under this assumption, the assembly of the Lotka-Volterra preemptive competition model will have two final states, corresponding to each species in isolation. Because the invader can enter the system only at low abundance, the local stability of the points is sufficient to make them "assembly endpoints".

Fixed-point attractors We consider models in which the asymptotic state of the local community is a feasible, stable equilibrium. In the Appendix, we relax this assumption.

Top-down assembly To complement our analysis we also consider top-down assembly, which violates the first two assumptions above. In this scenario, invasion rate is high-enough

that all the species in the pool can attempt to invade before any extinction takes place. As such, the process is equivalent to the scenario in which we initialize the local community with all the species present, at an arbitrary initial condition. The assembly endpoint amounts to the attractor reached by the pruning of the community by means of dynamics.

3.2.3 Local Dynamics

For concreteness, we consider competitive communities of n species and Generalized Lotka-Volterra (GLV) dynamics:

$$\frac{dN_i}{dt} = N_i \left(r_i - \sum_{ij} a_{ij} N_j \right) \tag{3.1}$$

Each species is assigned a positive growth rate $r_i > 0$, so that each can grow in isolation. We take the interaction matrix $A = (a_{ij})$ to be of the form $A = D(v)BD(w)$, where B is a non-singular, symmetric matrix, with nonnegative entries, and $D(v)$ and $D(w)$ are diagonal matrices with diagonal entries equal to v and w, respectively. Further, we assume that $v_i > 0$ and $w_i > 0$ for all i, ensuring that species compete with each other. Notice that $D(v)BD(w)$ can always be expressed such that $b_{ii} = 1$, by reabsorbing the diagonal of B into v and w. As such, we only consider matrices with $b_{ii} = 1$ for all i. The parameters v_i modulate the effect of other species on the growth of species i, and are thus related to resource requirements of i. Conversely, w_i models the resource use of species i. Eq. (3.1) is therefore:

$$\frac{dN_i}{dt} = N_i(r_i - v_i \sum_j b_{ij} w_j N_j) \tag{3.2}$$

By a change of variables $w_j N_j \to x_j$, $\gamma_i = r_i/v_i$, we obtain:

$$\frac{dx_i}{dt} = v_i x_i(\gamma_i - \sum_j b_{ij} x_j) \tag{3.3}$$

This classic system of equations has a global Lyapunov function [78], which is maximized through the dynamics (see also Appendix):

$$V(x) = 2\sum_i \gamma_i x_i - \sum_{ij} x_i x_j b_{ij} \tag{3.4}$$

For any equilibrium x^* we find: $V(x^*) = \sum_i \gamma_i x_i^*$. This quantity has a more intuitive interpretation in the case $r_i = v_i$ (and thus $\gamma_i = 1$), in which case V at equilibrium is simply the total biomass of the system.

Because B is non-singular and symmetric, the model yields only fixed-point attractors. If we take B to be positive definite (stable), we reduce the number of attractors to one [58].

When B is positive definite, the unique attractor is characterized by the following: we can identify a set of species $S \subseteq \{1, \ldots, n\}$ which can coexist at a positive equilibrium (feasibility) and moreover any species not present in S cannot invade when rare (non-invasibility). This type of attractor is called a saturated rest point, or non-invasible solution [58, 103].

Because of the restrictions on Eq. (3.2), any subset of species will follow the same type of dynamics (given that any minor of a positive-definite matrix is also positive definite), and thus any subset of species will have a unique globally attractive fixed point with the same characterization.

This result yields a simple test to assess if the arrival of a new species can push the community to a new state: it is sufficient to determine whether the species can invade when rare. If so, then the resident community, or any of its subcommunities, cannot be the attractor for the augmented community. Thus, the new invader must be part of the new configuration, thereby changing community composition. Consequently, for any invader j, we call the invasion successful if j can invade when rare. While rarely formalized, these conditions are implicitly assumed when performing invasion analysis [1].

When B is not positive definite, the community can have more than one attractor [14]. Nevertheless, as shown in the Appendix, each attractor S is characterized by the same

conditions as for the case of B stable: the equilibrium with the species S is a feasible and non-invasible set of species, whose interaction matrix $B^{(S)}$ (i.e., the matrix B in which we retain only the rows and columns corresponding to the species in S) is stable.

Example: consumer-resource model

Lotka-Volterra systems with an interaction matrix as in Eq. (3.2) arise naturally from consumer-resource models. Let C_i and R_k be a set of consumers and resources whose dynamics are defined by the MacArthur's model [78]:

$$\begin{aligned} \frac{dC_i}{dt} &= C_i\left(-d_i + \sum_k v_i p_{ik} R_k\right) \quad , \; i = 1, \ldots, n. \\ \frac{dR_k}{dt} &= R_k\left(r_k - b_k R_k - \sum_j w_j p_{jk} C_j\right) \quad , \; k = 1, \ldots, m. \end{aligned} \tag{3.5}$$

Assuming R_k is at equilibrium, then:

$$R_k = \frac{1}{b_k}\left(r_k - \sum_j w_j p_{jk} C_j\right) \tag{3.6}$$

Thus replacing in the equation for the consumers we find:

$$\frac{dC_i}{dt} = C_i\left(\sum_k v_i p_{ik} \frac{r_k}{b_k} - \sum_j v_i \left(\sum_k p_{ik} \frac{p_{jk}}{b_k}\right) w_j C_j\right) \tag{3.7}$$

Letting $b_{ij} = \sum_k p_{ik} \frac{p^T_{kj}}{b_k}$ so $B = PD(b)^{-1}P^T$, we recover Eq. (3.2) with B positive definite for $m \geq n$. The procedure implicitly assumes a separation of timescales between consumers and resources (with resources equilibrating faster than consumers), and the positivity of all resources at the given equilibrium. It can be shown that the results of the next sections will hold even if we analyze the full consumer-resource model under the assumption that the all the resources are always present (see Appendix and [27, 80, 88]). Yet, relaxing the constraint

on the coexistence of resources leads to different effects.

Similarly, we can recover a symmetric interaction matrix A if we assume that the interactions are functions of a measure of similarity between species, $a_{ij} = f(i,j) = f(j,i) = a_{ji}$.

3.2.4 Assembly Graph

Each species in the regional pool is identified by its interactions and growth rates. Thus, the parameters (A, r) for Eq. (3.1) define the species pool. The pool in turn allows us to define the assembly graph G [25, 54, 69, 102]. By assuming $r > 0$, we are considering as members of the species pool only species which enough affinity with the local habitat such that they can sustain themselves [13].

Definition of G

G is a directed, simple graph, with vertex set $V(G)$. Elements of $V(G)$ are indexed by subsets $S \subseteq \{1, \ldots, n\}$ such that the S species can coexist, i.e., the system in Eq. (3.2) restricted to the S species has a globally stable interior equilibrium point x^S. The empty set $\{\emptyset\}$ trivially satisfies this constraint, and thus we include it in $V(G)$. The edges of the graph G are determined by invasion events. Let j be a species: j connects two vertices S and S' in G if: i) j is not in S, ii) j can invade x^S when rare ($x_j \approx 0$), and, iii) j is in S'—the set of species of the new community. We denote this transition by $S \xrightarrow{j} S'$.

The explicit separation of timescales between invasions and local dynamics allows the assembly graph G to completely capture the assembly process: any sequence of successful invasions will be represented as a walk within G, and all the possible states that the system can potentially arrive to are contained in G. As such, properties of the assembly process will be reflected in properties of G and we can forget about the particularities of the local dynamics.

Properties of G

Because of their relevance to the equivalence between top-down and bottom-up assembly, here we focus on three properties: accessibility, assembly endpoints, and assembly cycles.

Accessibility For any set S of coexisting species, we can ask whether S can arise through assembly. We can only observe a subset S if there is a sequence of invasions taking the system from the empty state to S. Thus, we are interested in determining whether, for any vertex S, there is a path in G starting at $\emptyset$ and ending at S. Such a path is called an assembly path for S. We call an assembly graph G accessible if all subsets S in G have at least one assembly path.

Assembly endpoints Historical contingencies arise if different sequences of invaders can drive the community towards different final states. We call the final states of the assembly process "assembly endpoints" ([102], also known as "permanent" or "persistent" states [38, 54, 69, 120]). As such, the existence of historical contingencies requires the existence of multiple assembly endpoints. Here we distinguish between two types of endpoints. In the simplest case, there is a set of species S which is resistant to invasions by any other species in the pool. Thus S is a vertex with no outgoing edges, called a "sink" in graph theory. In the more complex case, an endpoint is comprised of a set of feasible communities $\mathcal{U} = \{S_1, \ldots, S_k\}$ for which invasions only transition the system within the set $\mathcal{U}$, and any two subcommunities S_i, S_j in $\mathcal{U}$ are connected by a path in G. In this case, rather than having a single-vertex endpoint, we have a set of communities that function as a sink: once the assembly process reaches any of these nodes, the system can only move between the states in $\mathcal{U}$ (see Appendix for a more detailed discussion).

Cycles An assembly cycle represents a set of communities $\{S_1, \ldots, S_\ell\}$ such that we have a sequence of invasions taking $S_1 \to \ldots \to S_\ell \to S_1$ (for an example see [102]). Trivially,

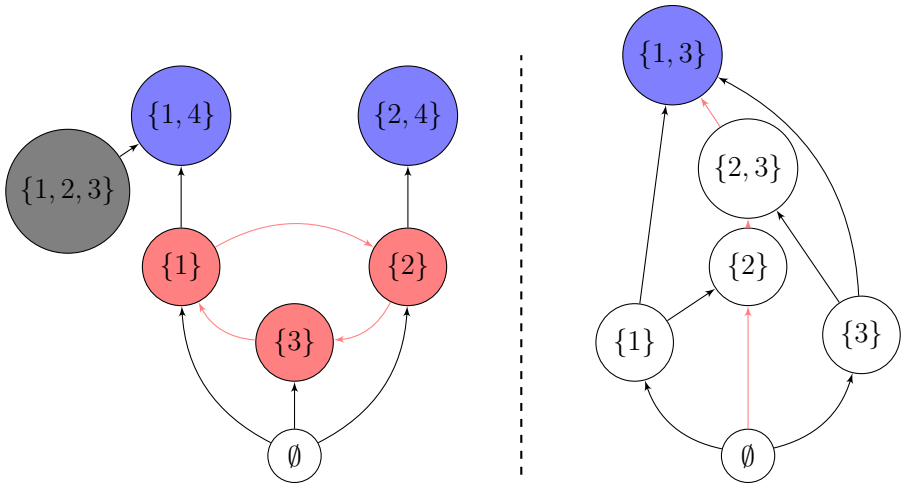

Figure 3.1: Example of assembly graphs G for a pool containing 4 (left) or 3 (right) species. Left: G is not accessible (i.e., the grey node is unreachable from the empty state), and possesses an assembly cycle (red nodes) and two assembly endpoints (blue nodes). Right: G is accessible, acyclic, and with a unique sink (enpoint) node for which a possible assembly path is highlighted in red.

assembly cycles map to directed cycles (in the graph-theoretical sense) within the assembly graph G. For a cycle to exist, it is necessary that at least one invasion event causes an extinction. Suppose that those types of transitions between $S_1 \xrightarrow{j} S_2$ are only possible if $S_1 \cup \{j\}$ cannot coexist (this will happen for example if feasibility of all the species in a community implies global stability of the attractor). Then, cycles are the natural generalization of intransitive competition (e.g., rock-paper-scissor communities [2]). Indeed, for the cycle to occur, we need a species j that sends some of the species in S_1 extinct, and conversely to go back to S_1 there must be a species in the total set of invaders such that makes j go extinct. Cycles are always present within an assembly endpoint $\mathcal{U}$ containing more than one community: for any two communities S_i and S_j in $\mathcal{U}$ we get a cycle (not necessarily simple) by joining the path going from $S_i \to S_j$ with the path from $S_j \to S_i$.

With these three definitions at hand, equivalence between top-down assembly and bottom-up assembly means: *Given a species pool, the bottom-up assembly endpoints are the same as the final attractors for the top-down assembly process.*

3.3 Results

First, we consider the case in which B is positive definite (stable). We have:

(a) G is accessible, i.e. there is a path leading from the empty set $\emptyset$ to any feasible subset S. This means that we can potentially observe any subset of the species of the pool which can coexist together. Moreover, for each S in G we can find an assembly path in which no species goes extinct. As such, to build S from the empty set we only need to choose the right order of invasions within S, guaranteeing that we do not observe "Humpty-Dumpty" communities [70, 96, 120].

(b) G has no (directed) cycles.

(c) G has unique sink vertex u and source vertex $\emptyset$. Coupled with (a) and (b), this implies that no historical contingency is possible—asymptotically, the assembly will always result in u.

For any graph with the above properties, there exists an ordering of the vertices such that two of them are connected only if one appears after the other (i.e., topological sorting [33]). This implies that we can find a quantity Q that increases monotonically as the assembly process unfolds. Thus, the assembly process "optimizes" Q. In the example of Figure 3.1, we can take Q to be the height of each node in the plot. For the assembly model, Q has a concrete meaning: it can be chosen as the Lyapunov function V in Eq. (3.10). In the case of $r_i = v_i$, we have that at each feasible subset S, V represents the total biomass at the equilibrium x^S. Thus, the assembly endpoint u is the feasible sub-community with the maximal total biomass, and at each step along an assembly path the total biomass present in the local community increases (Figure 3.2).

By the non-invasibility characterization of the attractors in Eq. (3.2), the unique sink vertex in G, u, is precisely the unique attractor of the dynamics for Eq. (3.2). That is, u

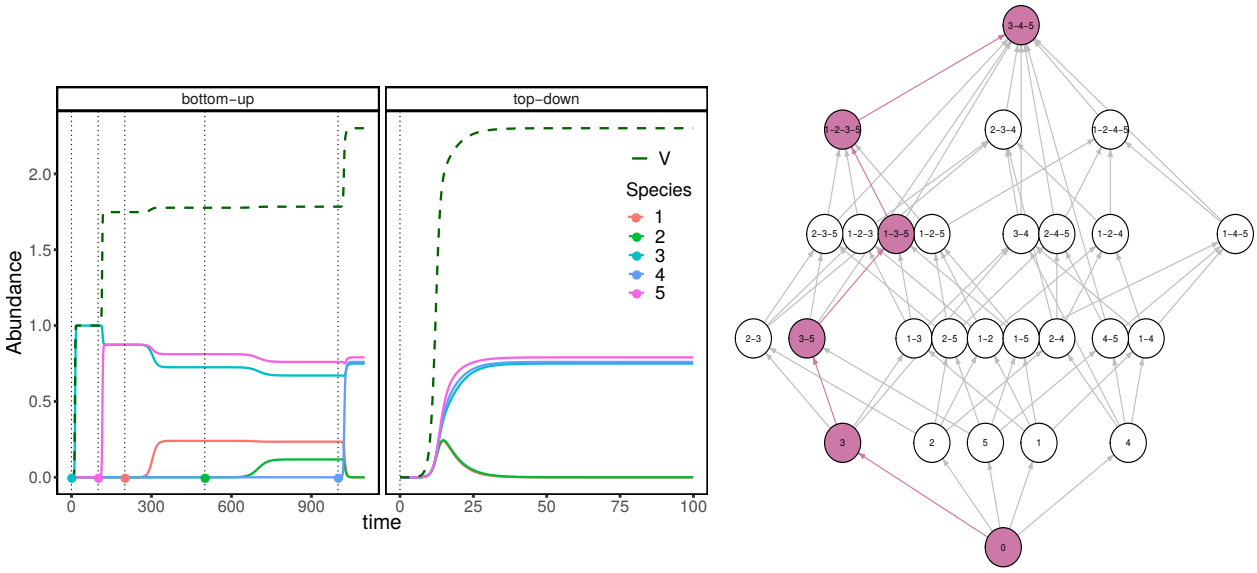

Figure 3.2: Assembly graph (right) and assembly trajectories (left) for a community of 5 species, with $r = v$. The assembly sequences are build assuming bottom-up assembly (left panel) and top-down assembly (right panel). The assembly sequence for bottom-up assembly is highlighted in the assembly graph. The quantity V is plotted in dashed dark green and shows a non-decreasing trajectory in both types of assembly processes.

corresponds to the state that would be reached when we initialize the system will all species present (top-down assembly). Because of the fact that G is accessible, u can be observed during assembly. Finally, acyclicity of G rules out the possibility of other types of assembly endpoints. Thus, it follows that when B is stable, the top-down and bottom-up assembly processes are equivalent.

When we relax the stability condition for B, there is no guarantee of a unique global attractor [14]. In this setting, there is the potential for priority effects and historical contingencies. Yet, a global Lyapunov function V for Eq. (3.2) exists regardless of the stability of the interaction matrix B [78]. As in the stable case, the existence of V implies that G is acyclic. Similarly, the attractors of the dynamics satisfy the same non-invasibility conditions, akin to the case of B stable (see Appendix). Thus, regardless of the stability of B, the assembly graph G is accessible and acyclic, with sink vertices in correspondence to attractors for the top-down assembly process. Hence, bottom-up and top-down assembly are still

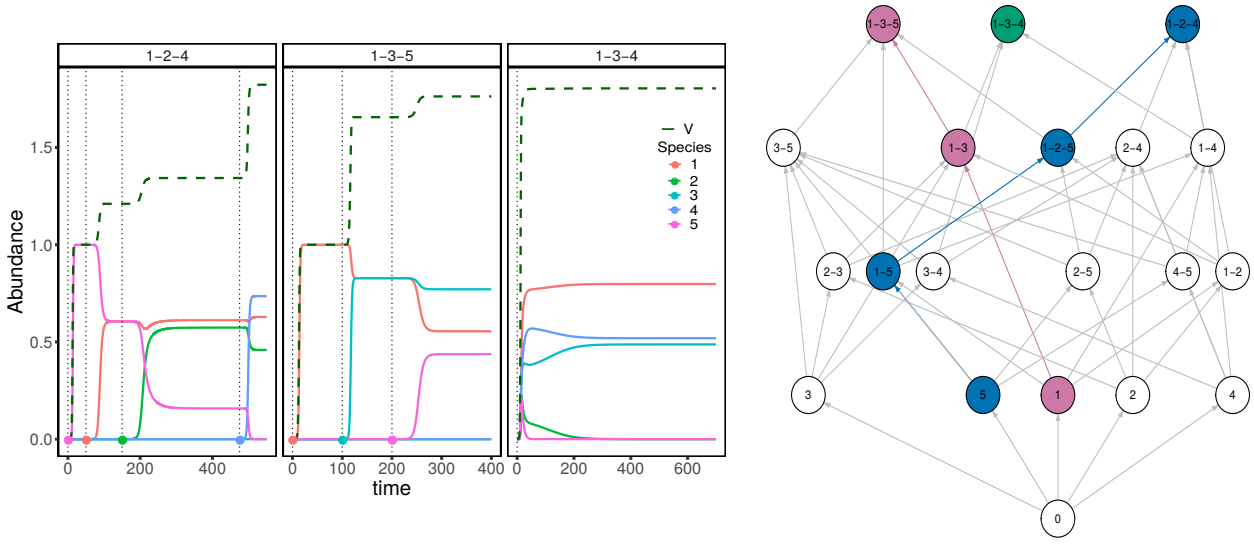

Figure 3.3: Assembly graph (right) and trajectories (left) for a community of 5 species, with $r = v$, and B unstable. The trajectories on the left are labeled by the assembly endpoint they reach. In the middle and left panel we proceed by bottom-up assembly, and the sequence of invasions is highlighted in the assembly graph on the right. In the rightmost panel we start with all the species at low density and converge to the subset $\{1, 3, 4\}$. The quantity V is plotted in dashed dark green and again shows a non-decreasing trajectory in all three different assembly processes.

equivalent: there are initial conditions for the top-down scenario leading to the same exact endpoints as found in the bottom-up scenario with different invasion sequences (Figure 3.3).

In summary, the equivalence between top-assembly and assembly is granted for competitive communities described by Eq. (3.2). Additionally, the assembly process described by Eq. (3.2) is indeed orderly and predictable, in the sense that the endpoints of assembly are the ones maximizing a given function (e.g., total biomass). In the Appendix, we show that optimization along assembly is a sufficient condition for the equivalence between top-down and bottom-up assembly. Nevertheless, if we want to recover only the accessibility of G (i.e., want to know any feasible subcommunity can be observed during bottom-up assembly), it is sufficient to impose optimization only locally. In this case, we require that for any feasible subset S, whenever we restrict assembly to the subcommunities of S, then assembly proceeds as an optimization process. Accessibility does not preclude the existence of cycles in G (see

Appendix for an example).

3.3.1 Extensions

Thus far, we have made an explicit separation of timescales between invasion events and local dynamics. This allowed us to only consider the asymptotic behavior of the local communities along an assembly path. If this condition is not satisfied, transient dynamics can play an important role (for example, see Schreiber and Rittenhouse [102], and recall the rock-paper-scissor example). To lift this condition, we model the number of invasion events as a Poisson process with rate ρ. At each invasion event, an invader j is sampled uniformly at random from the set of species in the pool, and is added to the community at low density, $x_j \approx 0$. Let S_1, S_2 be two feasible states, such that asymptotically we have $S_1 \xrightarrow{j} S_2$. Let t_{12} be the time it takes for the community to reach its asymptotic attractor. The probability that such a transition is observed under the new model is simply $e^{-\rho t_{12}}$. Thus, in the limit of $\rho \to 0$ we recover the bottom-up assembly with one invasion at a time we have studied above, with assembly graph denoted by G_{asym}. In the limit of $\rho \to \infty$ all the species are present in the local community and the model results in a Lotka-Volterra model with immigration [14]. To explore how the process changes along ρ, we take the assembly graph G to be a function of ρ.

Let $j_1, j_2, \ldots, j_\ell$ be a sequence of invaders arriving to the local community at times $t_1, \ldots, t_n$. Let $S_1 = \{j_1\}$, and τ_1 the time at which the local community reached an asymptotic configuration S_2. This means that all the invaders arriving between t_1 and τ_1 push the system from $S_1 \to S_2$. Proceeding in this fashion, we have transitions $S_1 \to S_2 \to \ldots \to S_m$ at times τ_i, and each transition is the product of all the invasions happening between τ_i and τ_{i+1}. For a fixed rate ρ we modify the assembly graph $G_\rho := G(\rho)$. Edges in G_ρ represent this new type of transition and are weighted by the probability that such a transition occurs (e.g., $e^{-\rho t_{12}}$ for our above example). The vertices remain the feasible subsets. Define

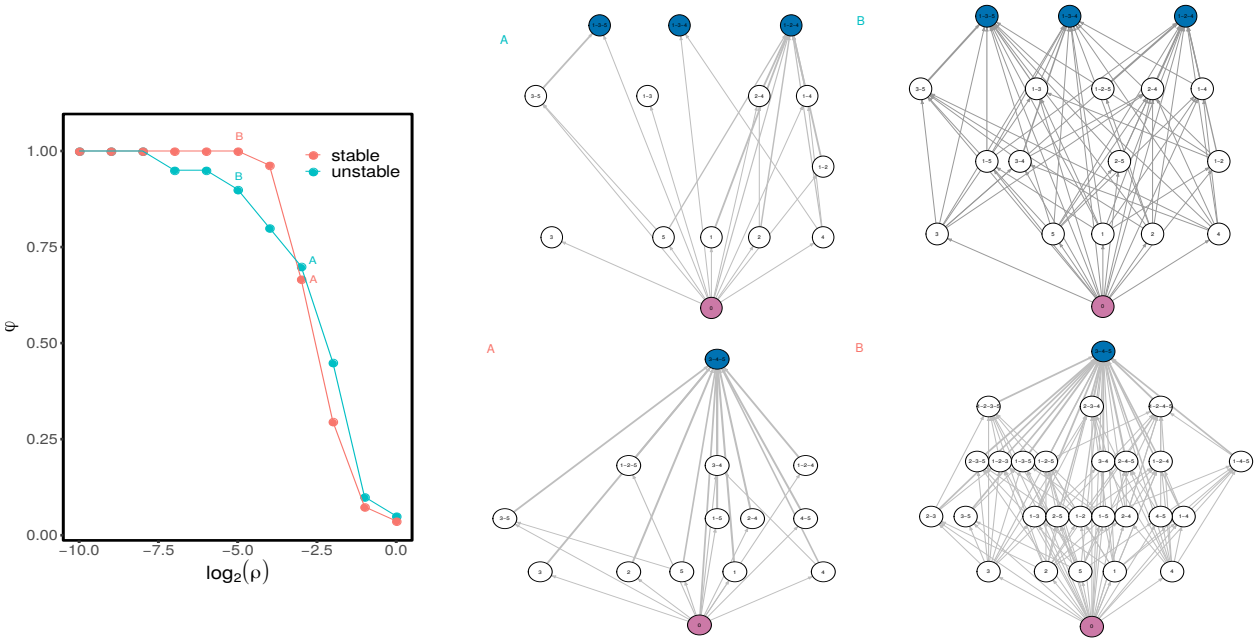

Figure 3.4: Fraction of accessible states as a function of the rate of invasion ρ for 500 random sequences of invasions of length 100, for the stable and unstable cases showed in the previous figures. On the right, we show the assembly graphs constructed under two particular rates. The width of the arrows represents the number of times this transition was observed during the simulations. Notice that the appearance of new sinks is due to the fact that we only consider as nodes states that reached their asymptotic equilibrium. Thus, the new sinks correspond to cases in which, although invasions are possible, a new equilibrium state is never reached due to the high invasion rate.

$G_0 = G_{\text{asym}}$, then all edges in G_0 appear for any G_ρ with non-zero probability. Increasing ρ will cause assembly to be able to access, with high enough probability to be observed, only feasible states with "fast" dynamics. Thus, at some value of ρ we would only be able to access, if any, only the fixed point attractors for the top-down assembly. Figure 3.4 shows how the fraction of feasible states observed along assembly changes as a function of ρ.

3.4 Conclusions

We have explored the equivalence between two different types of assembly processes, which we called top-down and bottom-up assembly. In top-down assembly, there is a single inva-

sion event in which all the species enter the local community at an arbitrary initial density and dynamics are fueled by ecological interactions, driving the system to a particular configuration of coexisting species. In contrast, in bottom-up assembly invaders arrive one at a time and at very low abundance, with invasion events spaced apart enough to allow for ecological dynamics to play out. More precisely, invasions are spread apart so that the local community reaches its asymptotic state between any two invasion events.

We showed that under a set of assumptions for the local dynamics—which are satisfied by symmetric competitive Lotka-Volterra models (see Appendix)—these processes are asymptotically equivalent: the final states the local community attains through either sequential invasions or a single, massive invasion event are the same. The condition for this equivalence is that assembly proceeds as an optimization process, a conclusion reminiscent of the early ideas about ecological succession [92]. However this equivalence, as we saw when relaxing the stability assumption for the local dynamics, does not preclude the existence of multiple assembly endpoints. Thus, while assembly may appear orderly and predictable along a particular assembly trajectory, the whole process is susceptible to priority effects and historical contingencies.

These conclusions are pertinent to the question of how to restore an ecological local community. Assuming that the pristine state was an assembly endpoint, even having access to all the members of the desired community will not be sufficient to restore the system, unless there is an assembly sequence comprised purely of members of the desired community [96]. The models studied here have the property that every feasible subsystem can be constructed by such a sequence. Contrast this situation with that in which to achieve a certain state through assembly, we need to go through "stepping stone" communities in which some invaders enter transiently in the system, and then disappear (as found experimentally by Amor et al. [6], Warren et al. [120]). In such cases, to bring the system to the desired state S we would need to be able to introduce the species in S as well as any other transient invader

in the right order—a much more challenging situation. Studying whether and when we can rebuild a community by introducing only the desired species in more complex models would be therefore be of both theoretical and applied interest [70, 102, 112, 120].

Our main theoretical device was the translation of the assembly process to a language centered around the assembly graph G (see also Capitán et al. [25], Hang-Kwang and Pimm [54], Law and Morton [69]). In the studied regime of assembly, G captures completely the assembly process so that features of the assembly dynamics are reflected in properties of G. We expect that further investigations of this graph will shed light on other features of assembly. For example, when multiple assembly endpoints exist, what is the probability that the process culminates at a particular endpoint? Studying random walks on G, where the structure of G determines the transition graph, may provide insights into this question [25].

The equivalence we have established is asymptotic in nature. For a large enough species pool, the number of invasions needed to converge to the final endpoint can be very large. Even when there is a unique assembly endpoint, looking at the community at a fixed number of steps along distinct assembly sequences may show markedly different community compositions. By computing the distribution of the length of the walks between the empty community and the sink nodes in G, we can estimate the speed of assembly and gauge the importance of transient assembly dynamics. Similarly, while the absence of cycles implies the equivalence between both assembly processes, it is not a necessary condition for the equivalence (see example in Appendix).

Our results suggest that we can approximate the outcomes of bottom-up assembly by analyzing the outcomes of the much more manageable process of top-down assembly. Moreover, we note that the structure of G can be studied empirically. Assuming Lotka-Volterra dynamics, the parameters needed to construct G are precisely those that can be inferred from coexistence-type experiments [85]. Importantly, the subset of experiments needed to perform this inference is much smaller than would be required by a naïve approach. These

improvements, in combination with the results for top-down assembly, enable the formulation of testable predictions. And, in guiding the design and interpretation of novel experiments, our results will further elucidate the structure of ecological assembly.

3.5 Appendix

Consider a community of $N = \{1, \ldots, n\}$ species whose dynamics are described by a set of autonomous ODEs:

$$\frac{dx_i}{dt} = x_i f_i(x) \tag{3.8}$$

In the following we explore sufficient conditions under which an assembly process, whose species pool and local dynamics are given by Eq. (3.8), satisfy the equivalence among bottom-up and top-down assembly. To do so, we turn to the study of the properties of the assembly graph G described in the main text. Equivalence in this case explicitly means that we can extend an assembly sequence such that we reach the same configurations attained when starting with all the species present. In other words, the assembly endpoints are the same for bottom-up assembly and top-down assembly.

3.5.1 Assembly Graph

Definition

For any subset of species $S \subseteq \{1, \ldots, n\}$ let Σ_S be an attractor for the for the dynamics restricted to the species in S. We will call Σ_S feasible if Σ_S contains all species in S. We will assume that the following conditions, which are satisfied by the competitive GLV system of the main text, are satisfied by the dynamics given by Eq. (3.8):

Condition 1.

(a) Any feasible Σ_S is globally attractive within S. Thus, there is no true multistability (i.e., any feasible state is uniquely defined by the species it contains).

(b) Let x^S be the time average of Σ_S, then an invasion for $j \notin S$ is successful if it can invade at low density ($x_j \approx 0$) the subsystem S at x^S.

Define the assembly graph G in an analogous manner as in the main text:

(a) $V(G)$ is indexed by the feasible sets S.

(b) An edge exists between S, S' if there is an species j that can invade S and the dynamics of the augmented system ends up in $\Sigma_{S'}$.

For any subset S, we denote by G_s the subgraph of G induced by vertices corresponding to subcommunities of S.

Assembly end-points

An assembly endpoint is the final set of configurations of the local community constructed during assembly [38, 54, 69, 102, 120]. By definition, we have transitions of states only if an state can be invaded by at least one other species. Thus, assembly endpoints are configurations such that invasions do not take the system out of them. Of course, then, the whole set of communities would be an assembly endpoint! In order to avoid that, we need to impose a minimality condition: we are only looking for sets that satisfy the previous property and which do *not* include any other such set. We can use the assembly graph G to make this statement precise:

A set of nodes $U = \{S_1, \ldots, S_n\}$ of G, is an assembly endpoint if and only if for any S_i its outgoing links points within U, and moreover that any such S_i can be reached from another S_j in U by a sequence of invasions. Thus U is a *strongly connected component* of G, and moreover when collapsing the graph G by the equivalence relation of $s \sim s'$ if there are directed paths $s \to s'$ and $s' \to s$ in G, U is a sink.

3.5.2 Equivalence

A necessary condition for the equivalence of the two types of assembly is that assembly endpoints are in correspondence with attractors for the dynamics of Eq. (3.8). Since no

assembly endpoint with more than one vertex satisfy such constraint we must have that all assembly endpoints are comprised of one vertex. This then forces the attractors to be characterized by a non-invasibility condition:

Condition 2. *Let S' be a sink vertex in G_S, then $\Sigma_{S'}$ is an attractor for the dynamics of* (3.8) *restricted to S. Furthermore, any attractor for any subsystem of* (3.8) *corresponds to such a vertex, i.e., attractors are characterized by being non-invasible by any other species (one at a time) not present in them.*

In conjunction with condition 1, condition 2 implies that G_S has a unique sink vertex (S) for any feasible S.

With the correct language now in place, we can state our main result:

Theorem 1.

(a) Assuming conditions 1 and 2, and G_S acyclic for any feasible subset S then G is accessible. Furthermore, for any feasible subset S there is an assembly sequence towards S with no extinctions and build purely from species present in S.

(b) If in addition G is acyclic then bottom-up assembly is equivalent to top-down assembly.

By theorem 1, to see the equivalence between the assembly scenarios when the local dynamics are described by the GLV model of the main text, it is enough to check that the assembly graph G is acyclic. As shown in the next section, this is provided by the existence of the Lyapunov function V.

Acyclicity

As stated in the main text, if a graph G is directed and acyclic there exists a quantity Q that we can associate to each node such that Q increases along paths. In the context of community dynamics, this condition is equivalent to:

Condition 3. *For any attractor Σ_S take x^S to be its time average. Then there is a (smooth) function $h : \mathbb{R}^n \to \mathbb{R}$, such that we can order x^S by $h(x^S)$ and an edge exist between $\Sigma_S \to \Sigma_{S'}$ only if $h(x^S) < h(x^{S'})$.*

Lemma 1. *G is a directed acyclic graph (DAG) iff we can find such an h.*

Proof. If G is a *DAG*, by topological sorting we find a quantity Q and we can order the nodes (attractors) by it, thus $Q : V(G) \to \mathcal{N}$ and increases along edges. The set of equilibria $\{x^S\}$ is finite thus discrete, so we can cover it by disjoint balls of radius ε. Define for each a bump function α_{x^S} supported at $B_\varepsilon(x^S)$, then define $h := \sum_{x^S} Q(\Sigma_S)\alpha_{x^S}$.

To prove the converse, suppose to the contrary that there is a loop $\Sigma_{S_1} \to \ldots \to \Sigma_{S_n} \to \Sigma_{S_1}$. Then by definition of h we must have $h(x^{S_1}) < \ldots < h(x^{S_n}) < h(x^{S_1})$, a clear contradiction. □

It is important to remark that h is *not necessarily* a global Lyapunov function for the dynamics of (3.8). The existence of a Lyapunov function provide such an h, but in general not the other way around. From the biological point of view, the hope would be that h can be given a more concrete meaning as done in the example of the main text (total biomass).

Accessibility

Accessibility of G imply that any node besides the empty state has at least one *incoming* edge, thus the only *source* node of G is $\emptyset$. If G is acyclic both conditions are equivalent:

Lemma 2. *Let G be a DAG, then G is accessible iff G has a unique source node.*

Proof. The preceding discussion shows that G accessible implies that G has a unique source node.

For the converse, let $\Sigma_S \neq \emptyset$ be a node in G, then by definition we can find an incoming edge $\Sigma_{S_1} \to \Sigma_S$, and if $\Sigma_{S_1} \neq \emptyset$ we can repeat the process. If we proceed for long enough

we must find $\emptyset$, for otherwise we have a cycle. Indeed, take a path of length larger than the number of vertices in G, then by the pigeonhole principle, two of the nodes must be the same hence we have a cycle. Since G is acyclic we are done. □

We distinguish between two types of incoming links:

(a) **Additions:** The invasion event does not cause the extinction of any of the species present in the system.

(b) **Extinctions:** The invasion event causes some of the species in the attractor to go extinct.

If G has a unique source node and at least one of the edges for each node is of the first type, then we can recover a *perfect addition ordering* for each node [109]. In particular, any attractor can be recovered by assembly *purely from its pieces*. Furthermore, this implies that for any feasible subset S, there is a species $i \in S$ such that the removal of i from S does not cause any further extinction.

If, on the contrary, we have a vertex S for which all the incoming edges are of the second type, then we must have *stepping-stone* species [6, 120], meaning that some species are necessary for the assembly process to lead to a community composition, but which are not part of it. And, contrary to the above, the removal of any species from a feasible set S will cause an extinction cascade.

The same proof as for lemma 2, but in reverse, shows that:

Proposition 1. *If for each feasible subset S, G_S is acyclic and has a unique sink node then G_S (and therefore G) is accessible. Moreover, we can find a perfect addition ordering for each S.*

The last proposition then says that we do not need *global* acyclicity for G to be accessible, we simply need the graph restricted to feasible sets to be acyclic which is a more *local*

condition, specially if the typical behavior of the community is that a considerable fraction of the species cannot coexist ([10, 14, 103]). It remains to show the last statement of theorem 1.

Proof of Theorem 1. The same type of argument as in lemma 2 shows that if G is acyclic, we can always extend an assembly sequence such that we reach a sink node, thus both processes are equivalent since any assembly endpoint with more than one vertex has a cycle.

□

3.5.3 Symmetric Competitive Lotka-Volterra

In this section we state the properties of the competitive GLV model studied in Section 3.2.3 which allow us to conclude that the system satisfies all the assumptions of Theorem 1. Recall the equations:

$$\frac{dx_i}{dt} = v_i x_i(\gamma_i - \sum_j b_{ij} x_j) \tag{3.9}$$

Assume, that all parameters are positive. Additionally, assume that for $B = (b_{ij})$ any principal submatrix B^S induced by a given set of species S is nonsingular. And that the matrix $(B^S|\gamma^S)$ for γ^S the subvector of growth rates induced by the species in S, has always full column rank.

In [78], it is shown that regardless of the stability properties of B:

$$V(x) = 2\sum_i \gamma_i x_i - \sum_{ij} x_i x_j b_{ij} \tag{3.10}$$

is maximized through the dynamics, thus is a global Lyapunov function for the system. More explicitly we find:

$$\frac{dV}{dt} = \sum_i \frac{\partial V}{\partial x_i}\frac{dx^i}{dt} = 2\sum_i v_i x_i(\gamma_i - \sum_j b_{ij}x_j)^2. \tag{3.11}$$

It then follows that either the system diverges, or the attractors are fixed points of the system which are local maxima for V constrained on $x_i \geq 0$. Because $b_{ij} > 0$ the first condition is impossible [58], thus all the attractors are constrained local maxima. We now show that the local maxima are given by fixed points x^S, with positive entries only for the species present in S and for which two additional properties hold:

(a) x^S is non-invasible, i.e. for any $k \notin S$ we have:

$$\gamma_k - \sum_k b_{kj}x_j^S < 0 \tag{3.12}$$

(b) Let $B^{(S)}$ be the submatrix of B containing the rows and columns only for the species present in S. Then $B^{(S)}$ is positive definite.

Let x^S be a constrained local maxima for V with $x_i^S > 0$ only for $i \in S$. Then for any $k \notin S$ we must have:

$$\frac{\partial V}{\partial x_k}(x^S) = 2(\gamma_k - \sum_{j\neq k} b_{kj}x_j^S) \leq 0 \tag{3.13}$$

By the condition on B and γ the inequality is strict.

Now, let V^S be V restricted to the species in S, and y^S be the restriction of x^S to S. Then y^S is an interior local maxima for V^S. It follows that the Hessian matrix of V^S, evaluated at y^S, must be negative definite. A direct computation shows:

$$\mathcal{H}(V^S)(y^S) = -2B^{(S)} \tag{3.14}$$

Thus $B^{(S)}$ must be positive definite.

Now consider a fixed point x^S as described above. Then for any other vector u so that $x^S + u \geq 0$, i.e. $u_k \geq 0$ for any $k \notin S$, we have:

$$V(x^S + u) - V(x^S) = 2(\sum_{k \notin S} u_k(\gamma_k - \sum_{j \neq k} b_{kj} x_j^S)) - \sum_{ij} u_i u_j b_{ij} \tag{3.15}$$

Because the second term is quadratic in u_i and the first sum is negative for any u with at least one $u_k > 0$, then taking $|u_i| < \varepsilon$ for some $\varepsilon \ll 1$ shows that $V(x^S + u) < V(x^S)$. For u with all $u_k = 0$ the result follows as above by the Hessian of V^S. Taking the intersection of both neighborhoods of x^S the claim follows.

3.5.4 Assembly of Consumer-Resource model

Recall the equations for the consumer resource model for n consumers and m resources:

$$\begin{aligned} \frac{dR_k}{dt} &= R_k \left(r_k - b_k R_k - \sum_j w_j p_{jk} C_j \right) \\ \frac{dC_i}{dt} &= C_i \left(-d_i + \sum_k v_i p_{ik} R_k \right) \end{aligned} \tag{3.16}$$

Assume that all parameters above are positive. Let $d = (d_i)$, $r = (r_k)$, $\gamma = (r, -d)$ and $P = (p_{ik})$. Define A as:

$$A = \begin{pmatrix} -D(b_k) & -P^t D(w) \\ D(v)P & 0 \end{pmatrix}$$

Let $N = (R, C)$. Then, we are back to the generalized Lotka-Volterra equations:

$$\frac{dN}{dt} = N \circ (\gamma + AN)$$

Where $\circ$ stands for the Hadamard (component-wise) product. By the special structure of A, it follows that it is a B-matrix [58], thus all solutions to the system are uniformly bounded

and there exists a saturated rest point. In [27] it is stated that besides some degenerate cases the saturated rest point is the unique attractor of the dynamics. To avoid the degenerate cases it is enough to consider the following:

(a) **Uniqueness of equilibrium solution** Any matrix submatrix of P has full column rank. Thus for any subsystem with number of resources at least as big as the number of consumers the equilibrium is unique.

(b) **Strict invasibility** Let $\tilde{A}$ be a principal submatrix of A induced by a subset of species S, and $\tilde{\gamma}$ the subset of γ induced by the same set S. Then the matrix $(\tilde{A}|\tilde{\gamma})$ has full column rank. A consequence of this is that we cannot have an equilibrium with more predators than resources: If that were the case, then the induced $\tilde{A}$ is singular and $\tilde{\gamma}$ is in the span of $\tilde{A}$ thus $(\tilde{A}|\tilde{\gamma})$ is not full column rank.

By performing a linear change of variables $C_i \to P_i = (w_i/v_i)C_i$, $R_k \to Y_k = b_k R_k$, and $a_{ij} = v_i p_{ik}/b_k$ we map the system to:

$$\begin{aligned} \frac{dP_i}{dt} &= P_i\left(-d_i + \sum_k a_{ik}Y_k\right) \\ \frac{dY_k}{dt} &= Y_k\left(r_k - Y_k - \sum_j a_{jk}P_j\right) \end{aligned} \tag{3.17}$$

In this form, the results of [88] say that the globally attractive fixed point minimizes the function:

$$\psi(Y) = \frac{1}{2}\sum_k (Y_k - r_k)^2$$

with the constraints:

$$-d_i + \sum_k a_{ik} Y_k \leq 0, i = 1, \ldots, n$$
$$Y_k \geq 0, k = 1, \ldots, m. \tag{3.18}$$

This follows by the application of the Karush-Kuhn-Tucker(KKT) conditions [88]: At the minimum points we have:

(a) $P_i(-d_i + \sum_k a_{ik} Y_k) = 0$, for $i = 1, \ldots, n$.

(b) $P_i \geq 0$, for $i = 1, \ldots, n$.

(c) $u_k Y_k = 0$, for $k = 1, \ldots, m$.

(d) $u_k \geq 0$, for $k = 1, \ldots, m$.

(e) $r_k - Y_k - \sum_j a_{jk} P_j = -u_k \leq 0$, for $k = 1, \ldots, m$.

It then follows that any set of points satisfying the KKT conditions is a saturated rest point, thus by uniqueness the global attractor is the unique one satisfying them. By convexity of ψ the KKT conditions are sufficient, thus the globally attractive solution is the unique minimum.

The above framework can be used to study an assembly process in which the initial state is the state with all resources present and the consumers never cause extinction of the resources. In this way, we study only the assembly of the consumers in a changing background of density, but not identity, of resources. The claim is simply that ψ strictly increases along assembly: Indeed, let $S = (S_P)$ be a set of consumers. Choose a successful invader j, taking the system $S \to S'$. By the non-invasibility conditions of the attractor it must be present in S'. Since j is a predator, the new equilibrium $Y^{S'}$ minimizes ψ with an additional constraint, therefore $\psi(Y^{S'}) \geq \psi(Y^S)$. By the uniqueness of the minimum and the invasibility condition we must have $\psi(Y^{S'}) > \psi(Y^S)$ (see also discussion in [80]). Thus by ψ we can order all the feasible equilibria of the system, thus by Lemma 1, the assembly

graph would be acyclic. By the non-invasibility characterization of the attractors, Theorem 1 applies.

It is important to remark the importance of the assumption on the coexistence of resources. If the consumers can send resources extinct, we can find assembly processes with cycles, thus there does not exist an ordering of the states of the system. This comes from the fact that ψ actually decreases under successful invasions of resources: We can view ψ of the recipient community as a minimization problem with the additional constraint of $Y_k = 0$ for some k, while when we add the resource $Y_k \geq 0$. By the invasibility condition of the new resource, the previous solution is no longer a minimum and thus ψ decreases. Figure 3.5 shows assembly graphs for consumer resource systems in which we can find cycles, and we have that either it is still possible to reach the attractor of top-down assembly or not. In case a cycle exists, there is always the possibility that even if there is a way out of it the sequence of invasions is such that assembly does not take the system out of the cycle. Yet, if we assume that each successful invasion happens with the same probability, the probability of this scenario vanishes asymptotically. Thus, we expect that asymptotically we only observe cycles if they happen to be the endpoints of assembly (right panel in fig Figure 3.5).

Construction of G

The construction of the assembly graph, even in this simple setting is a demanding computational task. For a set of parameters (A, r), in order to define the nodes of G we need to compute the set of all feasible subcommunities. The number of possible subcommunities is bounded by 2^n and the actual size depends on the parameters, nevertheless if the size of feasible sets is a non-vanishing fraction of the total number of possible communities we have exponential growth with n. This makes the construction of the assembly graph for large enough communities computationally intractable. Regardless of this constraint, we believe that G could be helpful to visualize assembly for small communities under experimental

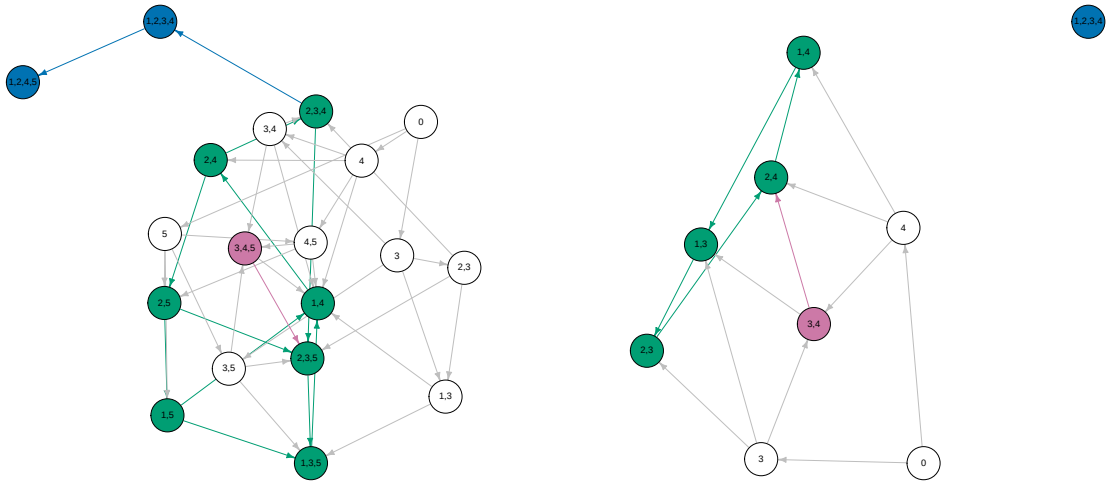

Figure 3.5: Assembly graph for consumer resource systems in which predators can make resources go extinct. In this case we see that cycles are not precluded: A strongly connected component for each graph is highlighted in green. Yet, we can have G accessible (left) or not (right). A path from the cycle to the top-down assembly attractor, if it exists, is highlighted in blue. The state with all resources present and an edge taking it to the respective cycle is highlighted in violet.

conditions, or as an arena where new theoretical devices can be applied to better understand assembly.

CHAPTER 4

PHYLOGENETIC EFFECTS ON COEXISTENCE IN LOTKA-VOLTERRA MODELS

Abstract

A species' traits influence the way in which it interacts with the environment. Thus, we expect traits to play a role in determining whether a given set of species coexists. Traits are, in turn, the outcome of an eco-evolutionary process summarized by a phylogenetic tree. Therefore, the phylogenetic tree associated with a set of species should encode information about the assembly properties of the community. Many studies have highlighted the potentially complex ways in which phylogenetic information is translated into species' ecological properties. However, much less emphasis has been placed on developing expectations for community properties under a particular hypothesis.

In this work, we couple a simple model of trait evolution on a phylogenetic tree with local community dynamics governed by Lotka-Volterra equations. This allows us to derive properties of surviving communities as a function of the number of traits, tree topology and the size of the species pool. Our results highlight how phylogenies and traits, in concert, affect the coexistence of a set of species.

In this way, our work provides new baseline expectations for the ways in which phylogenetic information is reflected in the structure of and coexistence within local communities.

4.1 Introduction

Gause's pioneering work [47] provided clear empirical evidence for the principle of competitive exclusion, which states that two species competing for a unique resource cannot coexist. In the context of niche theory, this principle resonates in the concept of limiting similarity—in a

purely interaction-driven community, species with similar niches are less likely to coexist [79]. Making stronger assumptions, one can draw a direct link between evolutionary relatedness among the members of an ecological community and their co-occurrence patterns. In particular, if one is willing to assume that species' traits are well-described by phylogeny, and that similarity in traits maps into strength of competition between species, one can connect the phylogenetic structure of an ecological community with coexistence [121]. While this hypothesis is not always supported by data [24], it has served as one of the cornerstones of the budding field of community phylogenetics [117, 122]. In recent years, several tools have been developed to test whether a given mechanism of community assembly (e.g., competitive exclusion or environmental filtering) has acted on the community, by analyzing the signal it leaves on its phylogenetic structure [44]. One criticism moved to this approach is that phylogenetic relatedness can affect the community in a variety of ways, meddling the link between phylogenetic and co-occurrence patterns [24, 82].

Here we take a step back and analyze a model in which we incorporate an explicit link between phylogenetic relatedness and ecological interactions. In particular, we connect phylogeny to species' traits, and then similarity in traits to the strength of interaction between any two species [12, 83]. We assume that a phylogenetic tree represents the relatedness among a regional pool of n species, and that species interactions are determined by a set of $\ell \geq n$ traits, which have evolved independently on the tree via Brownian motion [55]. Moreover, species are assumed to have a baseline competitive effect on each other, which is then modified according to their trait covariance. In this way, species that are more closely related tend to interact, on average, more strongly with each other than with distantly-related species. As we will show, the variance of the distribution of interaction strengths is controlled by the number of traits ℓ.

Clearly, phylogenetic relatedness would also influence intrinsic growth rates (species with similar traits would find similar environments to be hospitable). To more clearly separate

the effect of phylogeny on interactions from that on growth rates, we assume therefore that all species have the same intrinsic growth rate, thereby severing the connection between phylogeny and environmental filtering [13].

Having set up our model, we analyze the case in which all species in the pool are present in the local habitat at arbitrary initial conditions, and dynamics follow the Generalized Lotka-Volterra model. Contrary to previous simulation-based studies [44, 67] we develop an analytical framework to characterize many aspects of the resulting community of coexisting species, as a function of both the number of traits ℓ, and the tree structure. In particular, we show, that when the number of traits is large enough relative to the number of species in the pool, full coexistence is guaranteed by the tree-induced interaction structure. Furthermore, the abundance distribution of the community reflects the structure of the tree. On the other hand, while $\ell = n$ is a well-known necessary condition for coexistence [73, 127], full coexistence is almost never achieved in this case (see also [26]). Yet, even when the coexistence of the entire pool of species is precluded, one typically obtains communities of coexisting species of moderate size, as seen in the case of purely random interactions [10, 21, 103]. Differently from the purely random case, here we find that the probability that a particular species survives is determined by its position in the tree.

Our model shows that phylogenetic relatedness, modulated by the number of traits controlling the interactions between the species, affects multiple aspects of the local community. The explicit incorporation of community dynamics allows us to move from pairwise comparisons to global aspects of community structure. Furthermore, we advance the growing body of literature on random interaction models [10, 14, 21, 103] by analyzing a case in which the correlations between interaction strengths are controlled by phylogenetic relatedness.

4.2 Model

Consider a regional pool $\mathcal{R} = \{s_i\}$ of n species indexed by $1 \leq i \leq n$, and assume that a species' identity is defined by its $\ell \geq n$ trait values. For a given trait $1 \leq j \leq \ell$, collect the values of j for all members of the pool in the trait vector $\tau_j \in \mathbb{R}^n$. We sample each τ_j independently from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$. This choice implies that: (a) the trait values for a given species are independent, and thus we are not considering trade-offs between traits; (b) the processes leading to the appearance of the correlation structure Σ are statistically equivalent for distinct traits; (c) lastly, if $\Sigma_{ii} = \sigma$ for all i, then the distribution of trait values *within* a species is independent of species identity. For an example of an evolutionary process consistent with the assumptions above, consider the case in which $T_{\mathbb{R}}$ is the phylogenetic tree for the species in the regional pool, and each trait j starts at an ancestral mean value of 0, and evolves independently on the tree via Brownian motion. Then the value of trait j at the n tips, τ_j, follows the normal distribution $\mathcal{N}(0, \Sigma)$ with Σ induced by the tree structure, and called the variance-covariance matrix of $T_{\mathbb{R}}$ [55]. In Σ, the covariance between two species is given by the shared branch length on $T_{\mathbb{R}}$ [19]. As such, whenever $T_{\mathbb{R}}$ is ultrametric, then $\Sigma_{ii} = 1$ for all i. Unless otherwise specified, here Σ is always assumed to originate from an ultrametric, rooted phylogenetic tree (see Figures 4.1 and 4.2).

In this setting, each sampling of the ℓ traits defines a species pool $\mathcal{R}$. For a given pool, the following experiment is performed [103]: let all the species in the pool be present in the local habitat *at the same time* and at *arbitrary initial densities*. Population dynamics, fueled by the species' interactions and growth rates, will lead the community to an *asymptotic* state in which some of the species are extinct, while others coexist. Our aim is to characterize the resulting community of coexisting species in terms of the parameters ℓ, n and Σ.

To this end, we consider dynamics governed by the Generalized Lotka-Volterra (GLV) model. Species are assumed to differ only in their interactions, so that the growth of each

species in isolation is the same:

$$\frac{dx_i}{dt} = x_i(1 - \sum_j (\mu + A_{ij})x_j). \tag{4.1}$$

The interaction coefficients are modeled as deviations from a "mean-field" competition value $\mu > 0$. The deviations are controlled by trait similarity between species; more precisely, the deviations are modeled as the sample covariance matrix stemming from the trait sampling process, so that competition between two species is strengthened if their trait vectors are positively correlated and weakened otherwise:

$$\begin{aligned} A_{ij} &= \frac{1}{\ell} \sum_l \tau_l^i \tau_l^j, \\ G &= (\tau_j^i), \\ A &= \frac{1}{\ell} G G^T. \end{aligned} \tag{4.2}$$

In the supplementary information (section 4.5.1), we show how this model arises when performing a separation of time scales for consumer-resource models in which consumers have the same attack and death rates, but differ in the preference for the resources.

Under these assumptions, A is a symmetric and stable matrix, and a member of the *Wishart ensemble* [90, 125]:

$$A \sim \mathcal{W}_n(\frac{1}{\ell}\Sigma, \ell). \tag{4.3}$$

The Wishart distribution describes the probability with which a given *sample covariance matrix* is observed when sampling from a multivariate normal distribution. Given its many applications in statistics and other fields, the Wishart distribution has been studied in countless publications, allowing us to access a vast body of literature and results [16, 17, 66, 90].

Since A is stable, the community reaches a unique, globally-stable equilibrium, and the sub-community of coexisting species is characterized by a feasibility and non-invasibility

condition [58]. Importantly, in this case one can prove that the effect of the mean interaction strength μ on the resulting community is relatively straightforward: μ does not affect the identity of the survivors, and rescales species' biomasses by a constant (see supplementary information, section 4.5.6, for details). Thus, without loss of generality, we can assume $\mu = 0$, in which case that the regional pool is completely characterized by A.

To describe the statistical properties of the community of coexisting species, we need to condition the distribution of the variables of interest on the *unique* feasible and non-invasible sub-community for a given species pool $\mathcal{R}$. We focus on the following properties: distribution of the number of coexisting species, the total biomass of the community, and the relative abundance distribution of the coexisting species.

As an example of an empirical tree structure we take the phylogeny of the clade *Senna* (Fabales) [123]. The tree contains a total of 94 species and we use the outlier group comprising the species *Senna silvestris var guarantica, Senna siamea, Senna polyantha* and *Senna galeottiana* to root the subtree containing the remaining 90 species.

Notice that, as shown in chapter 3, the final community composition reached in each of our theoretical experiments is the same that would be reached under sequential, one-at-a-time species invasions. Thus, our results map directly to questions related to sequential, bottom-up assembly of ecological communities. In particular, our results can be used to infer properties of the assembly graph $\mathcal{G}$ associated with each regional pool, as studied in chapter 3 and Capitán et al. [25].

4.3 Results

Deterministic Limit. Let $\gamma = \frac{\ell}{n}$ be the ratio between the number of traits and the number of species in the pool. Then in the limit $\gamma \to \infty$ we find that $A \to \Sigma$ (i.e., the sample covariance matrix converges to the population covariance matrix). Thus, the properties of the community are given solely by Σ. The simplest case to study is that in which $\Sigma = I_n$

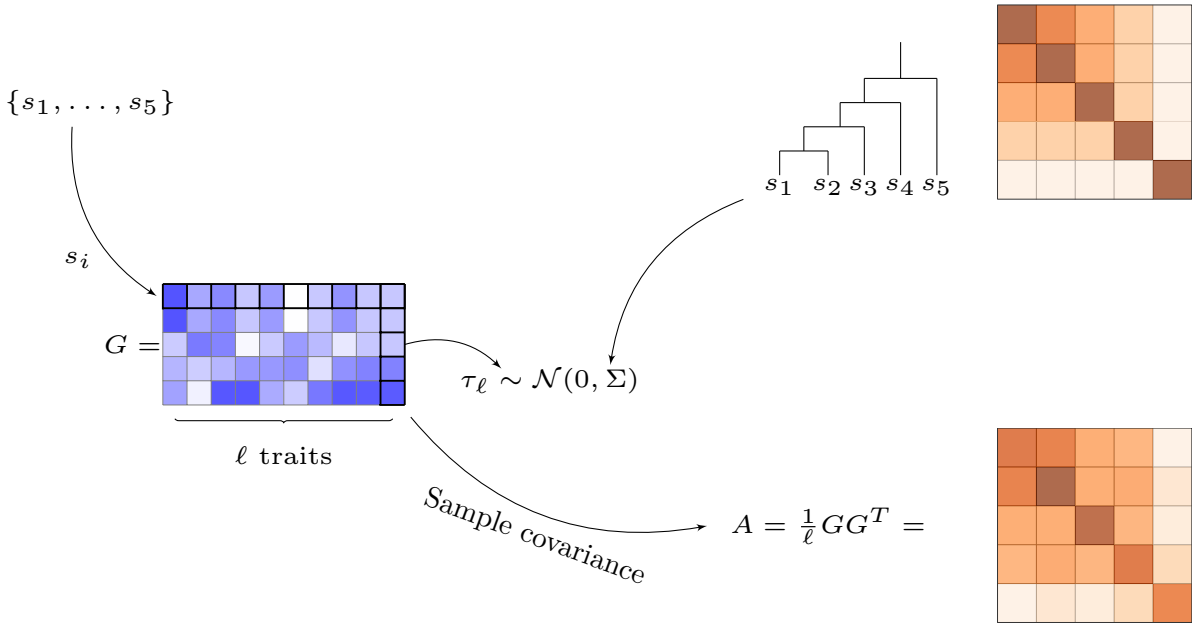

Figure 4.1: **Construction of the regional pool $\mathcal{R}$ and interaction matrix A.** Each species in the pool $\mathcal{R}$ is assigned ℓ trait values. The vector containing the values for trait j ($\tau_j \in \mathbb{R}^n$) of all members of the pool is sampled independently from $\mathcal{N}(0, \Sigma)$. This is equivalent to a neutral model of trait evolution for each j on a phylogenetic tree $T_\mathcal{R}$. The model relates the structure of $T_\mathcal{R}$ to the interactions between the species in the pool: the matrix Σ measures the shared evolutionary history between any two species s_i and s_j on $T_\mathcal{R}$ (in our example $\Sigma_{12} > \Sigma_{13} > \ldots > \Sigma_{15}$). In turn, the number of traits ℓ and Σ determine the interactions between species, stored in the matrix A.

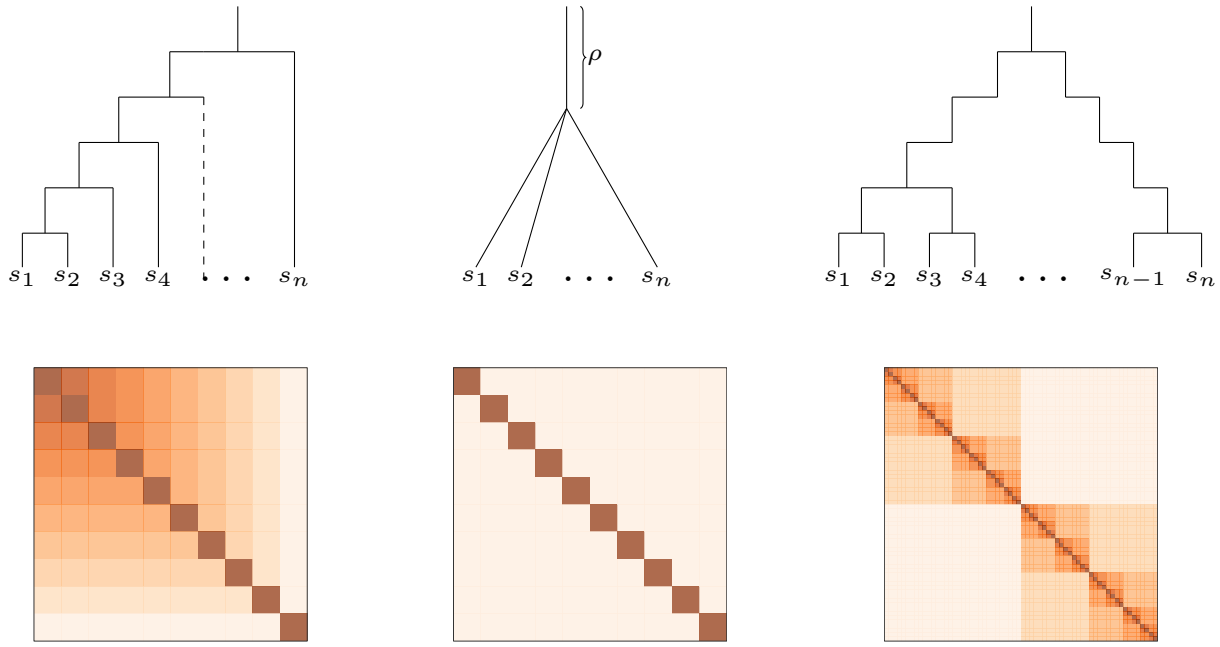

Figure 4.2: **Examples of ultrametric rooted phylogenies and its induced covariance matrices**. The perfectly hierarchical tree (left) has $n-1$ branching times $0 < t_1 < \ldots < t_{n-1}$ for a pool of n species, where each new branching happens to the "left" and creates a new pair of species. We call the time between branching events, $t_i - t_{i-1}$, *inter-branching times*. The star tree (middle) posses a unique branching event which generates all the n species. For the perfectly balanced tree (right) we have n branching *times* at each of which all the tips present up to that point generate two new species. Proceeding recursively, n branching times generate 2^n species and we have $n+1$ distinct inter-branching times. The covariance matrix associated with each tree is constructed as follows: For any s_i take γ_i to be the path "backwards" in time to the ancestral species at the root of the tree, then for any two s_i, s_j, let $t(i,j)$ be the time at with γ_i and γ_j merge, i.e. the coalescence time between s_i and s_j [119]. Then $\Sigma_{ij} = 1 - t(i,j)$. In particular Σ_{ij} is the total time for which the evolutionary processes for s_i and s_j are completely linked. For example, in the star tree $\Sigma_{ij} = \rho$ for any $i \neq j$ and $\Sigma_{ii} = 1$ given that each tree is ultrametric.

(the identity matrix), which represents the covariance matrix induced by the degenerate n-star tree with 0 shared branch length among all species (see Figure 4.2). In this case, complete coexistence of $\mathcal{R}$ and any of its sub-communities $S \subseteq \mathcal{R}$ follows trivially, since species are not interacting with each other. Interestingly, the same behavior is shared by any Σ induced by a tree T. The proof goes by induction on the following observation: let t_1 the time at which the first split happens in the phylogenetic tree; then, "cutting" the tree at this branching point generates non-interacting sub-trees $\tilde{T}_i$, which by induction have a fully coexisting equilibrium (i.e., the sub-community coexists, and each further subset also coexists). Pasting these sub-trees together at their roots gives us a degenerate tree for which the induced covariance matrix is a block diagonal matrix $\tilde{\Sigma}$. Going upwards, one recovers T by attaching a branch of length t_1 to the root. In particular Σ is a constant rank-one update of $\tilde{\Sigma}$. For a constant growth rate model, updates of this type do not affect the feasibility of the system, hence Σ has a feasible equilibrium (see Figure 4.8 and section 4.5.2 for a more detailed argument). Thus, for $\gamma \gg 1$, we are guaranteed full coexistence regardless of the shape of the tree and therefore $A \approx \Sigma$. Moreover, the assembly graph $\mathcal{G}$ for the species pool $\mathcal{R}$ contains all possible assembly histories (c.f. chapter 3 and [128]), i.e., any of the possible sub-communities can be built by starting with a single member of the community and adding the remaining members sequentially.

Total biomass and abundance distribution. Consider the two extreme tree topologies given by the "perfectly unbalanced" tree and the "perfectly balanced" tree (see Figure 4.9). Assuming equal inter-branching times, the total biomass of the system $W(n)$ for a pool of n species is given by $W(n) \approx \sqrt{n} - 1/4$ in the perfectly unbalanced case, and $W(n) = \frac{\log_2(n)+1}{2-1/n}$ in the perfectly balanced case (see Section 4.5.2 for details). Similarly, we are able to derive expressions for the individual biomass of each species s_i, where the index corresponds to the position in the ordered tips of the tree (see Figure 4.2). For the perfectly balanced case, the abundance distribution is trivial, since each species necessarily has the

same abundance. On the other hand, the hierarchical nature of the perfectly unbalanced tree is reflected in the individual biomasses, with species that split from the rest early on having higher abundances. Figures 4.3 and 4.9 show that the results are qualitatively unchanged if we sample the inter-branching times from exponential or uniform distributions which are then renormalized to the simplex (to keep with the case of an ultrametric tree). The uneven distribution of abundances for the unbalanced tree helps explain the difference the total biomass: in the perfectly unbalanced case, as n grows there is a fraction of species (outliers) that interact less and less strongly with the rest of the community, so that their abundance approaches the limit 1 (obtained for non-interacting species). In contrast, in the perfectly balanced case the abundance of all the species is the same, and approximately $\frac{\log_2(n)}{n}$.

To compare these results with those stemming from a more complicated tree structure, we repeated the calculation using the phylogenetic tree of the *Senna* clade (Fabales) [123]. The average total biomass $W(n)$ for sub-communities of different sizes shows that at small enough sizes, $W(n)$ behaves as predicted by the perfectly hierarchical model—which reflects the hierarchical low-level structure of the tree (Figure 4.6); as the size of the sub-community increases, $W(n)$ reaches values even smaller than the perfectly unbalanced tree. The species' abundance distribution, as for the perfectly hierarchical tree, reflects the tree structure: the abundance profile shows peaks at each of the outliers within clades, and an overall decreasing trend towards the nested parts of the tree (see also Figure 4.6).

Star phylogenies. Classical results in theoretical ecology have extended the principle of competitive exclusion to the case of multiple resources/regulating factors, showing that a necessary condition to observe a non-degenerate coexisting community of n species in our model is $\ell \geq n$ [73, 127]. We have shown above that, for a fixed size of the pool n in the $\ell \to \infty$ limit full coexistence is guaranteed. To characterize the cases in between $\ell = n$ and $\ell \to \infty$, we exploit the fact that A follows the Wishart distribution; as such we can make use of tools developed in statistics and economics to fully explore how the limit of

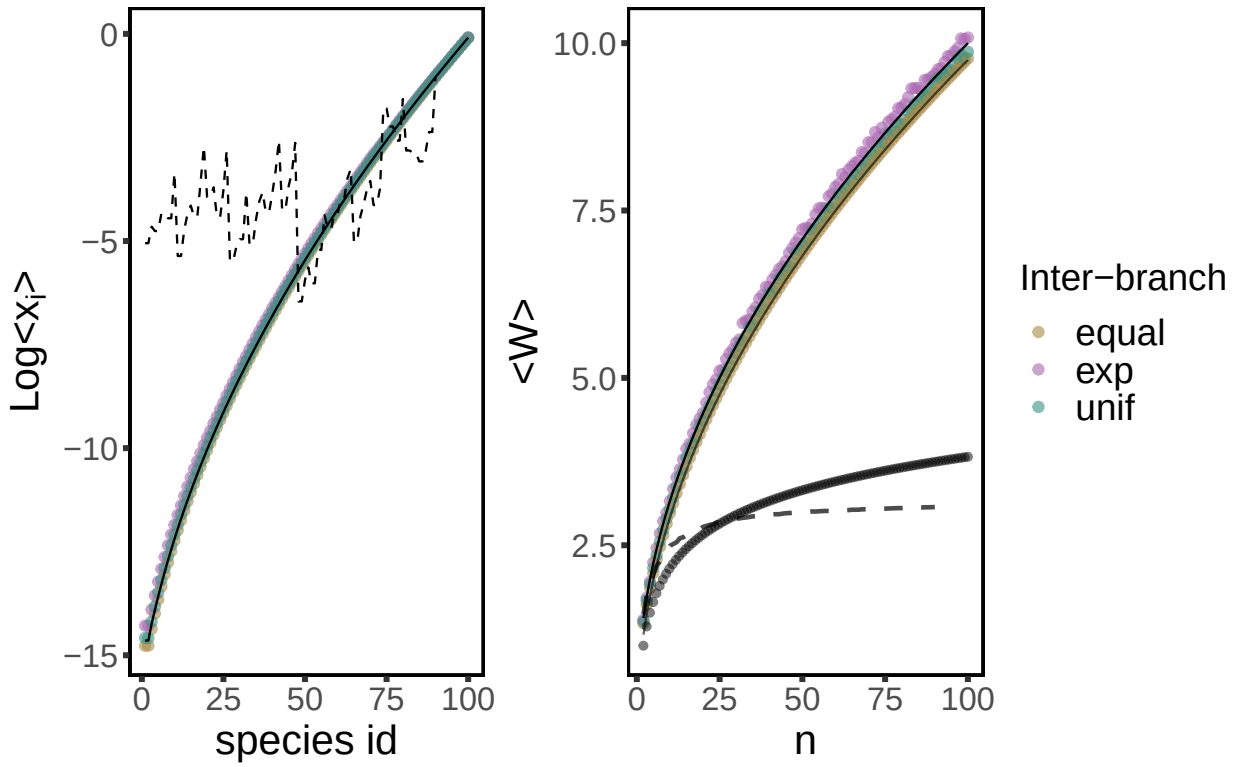

Figure 4.3: **Individual and total abundance for the deterministic limit**. Log individual abundance (left) and total abundance (right) for the communities in the deterministic limit of a perfectly unbalanced tree. Dots mark the average values when sampling the branch lengths from an exponential distribution with rate 1, a uniform $[0, 1]$ distribution, and the case of equal branch lengths. The total branch length is renormalized to 1 in all cases. Solid lines are the analytic predictions under equal branch length $1/n$. In the right panel, the two solid lines are given by $\sqrt{n}$ and $\sqrt{n} - 1/4$, the black dots represent the analytic formula for a perfectly balanced tree which shows logarithmic growth $\sim \frac{\log_2(n)+1}{2}$. The black dashed line on the right panel is the scaling with size of sub-trees of the *Senna* phylogenetic tree, and the dashed line on the left plot is the abundance distribution for the full tree (compare with Figure 4.6).

full coexistence is approached (see Section 4.5.3). To start, let Σ be induced by a star-tree with shared root of length ρ (see Figure 4.1). In this setting there is a constant correlation ρ among the species in the pool. We find that for $\gamma \approx 1$, full coexistence is never achieved for large enough communities (Figure 4.10). Nevertheless, the community does not collapse completely and a non-vanishing fraction of species is observed to coexist at the attractor (Figure 4.4). The effect of increasing correlation among species is, as expected, to reduce the mean fraction of survivors $\wp$. In particular, to have an attractor with at least half of the species, we need:

$$2\gamma \geq 1 + \frac{n\rho}{\pi(1-\rho)} \tag{4.4}$$

The quantity $\zeta = \frac{\rho}{1-\rho}$ could be interpreted in the framework of population genetics as the ratio of shared to private mutations for each species. It is a key quantity, in the sense that two distinct pools $\mathcal{R}$ and $\mathcal{R}'$, of sizes n and n' will yield the same mean fraction of survivors for a given γ whenever $n\zeta = n'\zeta$.

The distribution of total biomass, W, for the community of coexisting species is influenced by γ and ρ in two different ways: the parameters affect both the distribution of the number of survivors, and the conditional distribution of W for a given number of survivors. Assuming that the distribution of survivors is highly peaked at the mode, we derive an approximation for the mean of W that match results from simulations (see fig. 4.12 and section 4.5.4 for exact results and full distribution). For small enough γ the variance of the interactions allows the possibility of positive interactions that enhance W, as γ increases the interaction matrix converge to the purely competitive interaction matrix given by Σ. This convergence explains the decrease of W with γ depicted in Figure 4.5.

Using the same strategy, we are able to derive approximations (see section 4.5.5 for exact formula) for the survival function of the relative abundance distribution under distinct values of ρ and γ. In particular, the distribution becomes very peaked as γ increases, while

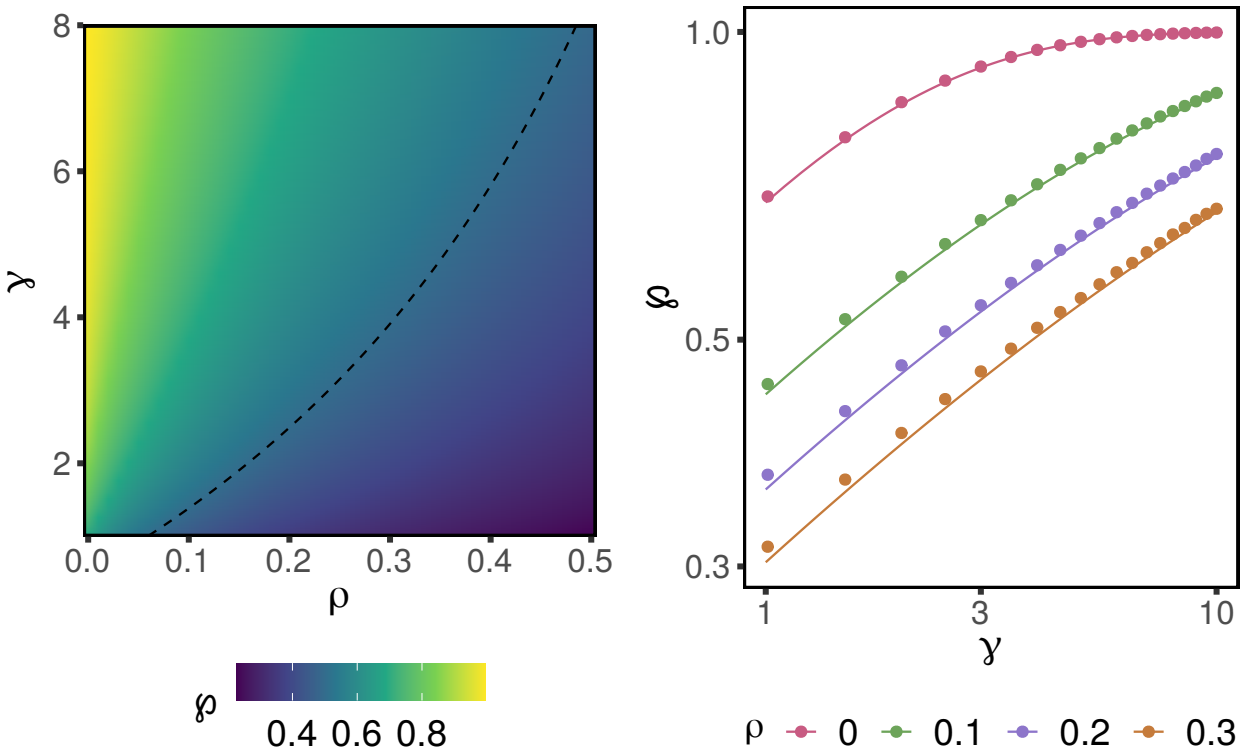

Figure 4.4: **Mean fraction of survivors $\wp$ as a function of ρ and γ.** In the right panel, we compare our analytical approximations (solid lines) with simulations (dots) for a regional pool of 50 species (log-log scale). The left panel explores in more detail the parameter space (γ, ρ). At the dashed line we have that the mean number of survivors is half of the species. As suggested by eq. (4.4), the value of γ giving a fixed $\wp$ increases sharply with ρ.

increasing ρ tends to make the distribution flatter (Figure 4.5).

Beyond constant correlation. Imposing a more general covariance structure Σ is challenging from a mathematical standpoint due to the breaking of the statistical equivalence among species—species in distinct parts of the tree have now different statistical properties. Because in this case the identity of the species does matter, instead of looking at the total number of survivors we focus on how the probability that a particular species survives (p_s) changes with its position in the tree. Simulations for the phylogentic tree of the *Senna* clade show that the model recreates the phenomenon of phylogenetic over-dispersion: for a group of closely related species p_s peaks at the outliers of the clade. Furthermore p_s reflects the tree structure in the same way that the total abundance distribution does (compare Figures 4.1 and 4.6).

To further explore this issue, we are able to analytically compute the distribution of survivals for distinct sub-communities for a 3-species tree with equal inter-branching times (Figure 4.7): sub-communities containing the outlier species s_3 always have higher probability to be observed than those in which s_3 is absent. More generally, the formulas in section 4.5.3 can be evaluated numerically to find the probability of survival of a particular sub-community under an arbitrary, not necessarily ultrametric, tree structure.

4.4 Discussion

By considering local community dynamics of a trait-based interaction model, our results provide a clear link between phylogenetic information at the regional species pool level, and many aspects of species coexistence. Importantly, while the tree structure is reflected in the local community patterns, the number of traits controlling interspecific interactions modulates the outcomes.

We found that, when phylogenetic relatedness completely controls interactions, i.e., when the number of traits is sufficiently high compared to the number of species, full coexistence

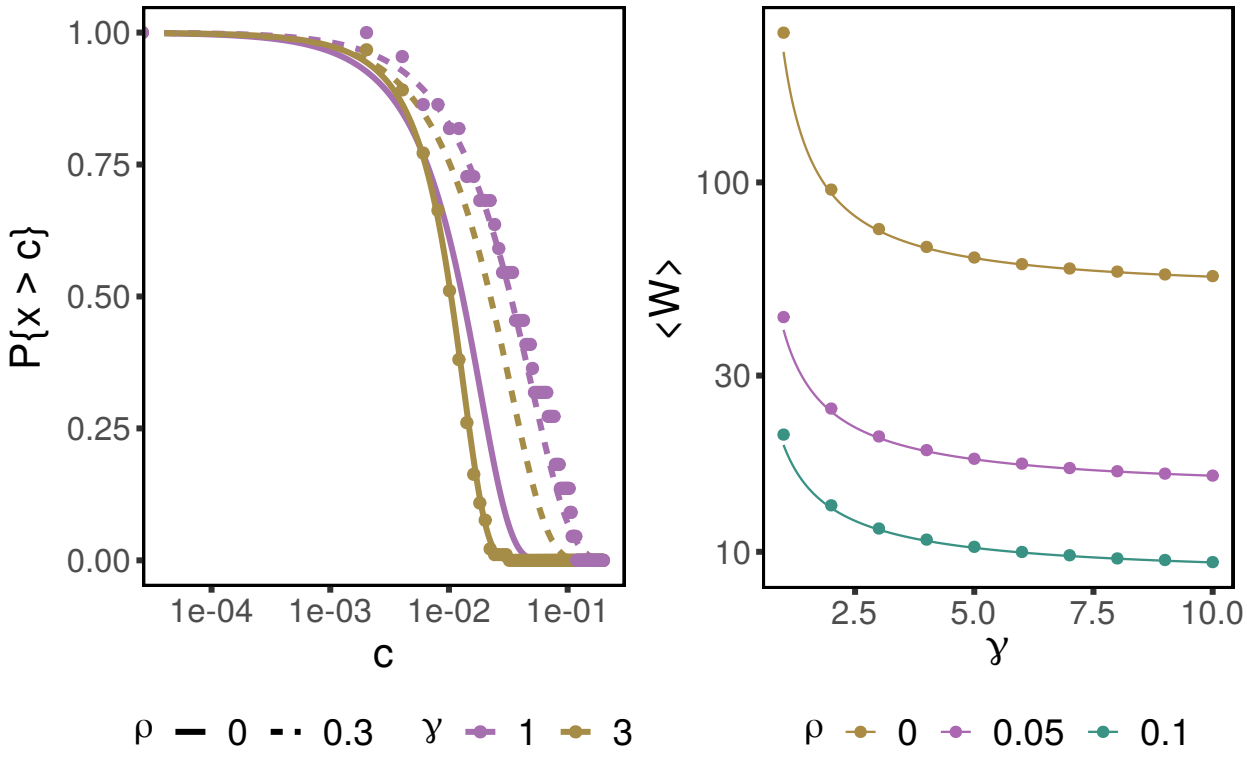

Figure 4.5: **Mean total biomass and relative abundance distribution**. The panel on the right shows (note the log-transformation for the y-axis) the mean total biomass for the community of coexisting species; the points represent simulations, and solid lines the corresponding analytical approximations for a pool of 50 species (see Section 4.5.6 for the effect of changing μ). The total biomass decreases as γ grows, because the overall strength of interaction between species decreases. The survival function for the relative abundance values of the community is plotted on the left panel (note the log x-axis), where again points stand for simulations and lines for analytical predictions for distinct γ and ρ values, and a pool of 100 species. For clarity, we just show simulations for the parameters $(\rho, \gamma) \in \{(0, 3), (0.3, 1)\}$. In particular we have that as γ increases the distribution becomes more and more peaked (as expected) while increasing ρ flattens the distribution.

.

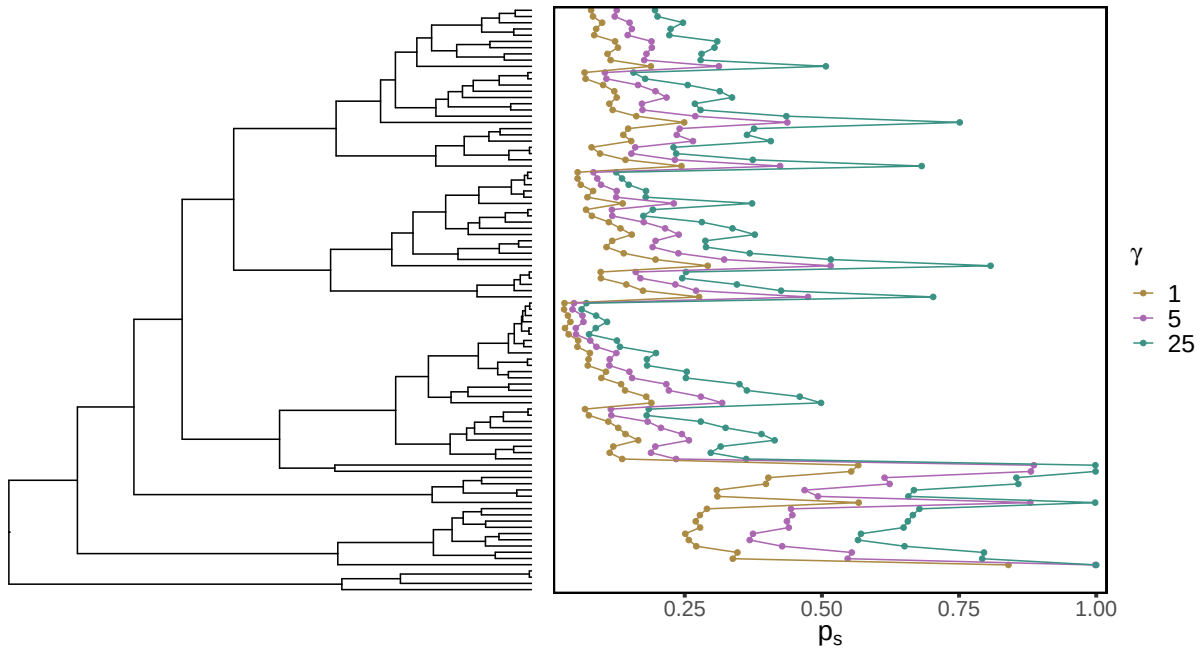

Figure 4.6: **Probability of individual species survival for an empirical tree**. The probability that a species is observed in the survival community p_s, after 5000 simulations, is shown alongside the phylogenetic tree (*Senna* clade) where the outermost group is used to set the root. The values p_s reflects the tree structure and the abundance distribution showed in Figure 4.1: The peaks in p_s correspond to outliers within a group of closely related species and p_s has a decreasing trend towards the most nested parts of the tree (upward direction). In particular, the model produces phylogenetic over-dispersion.

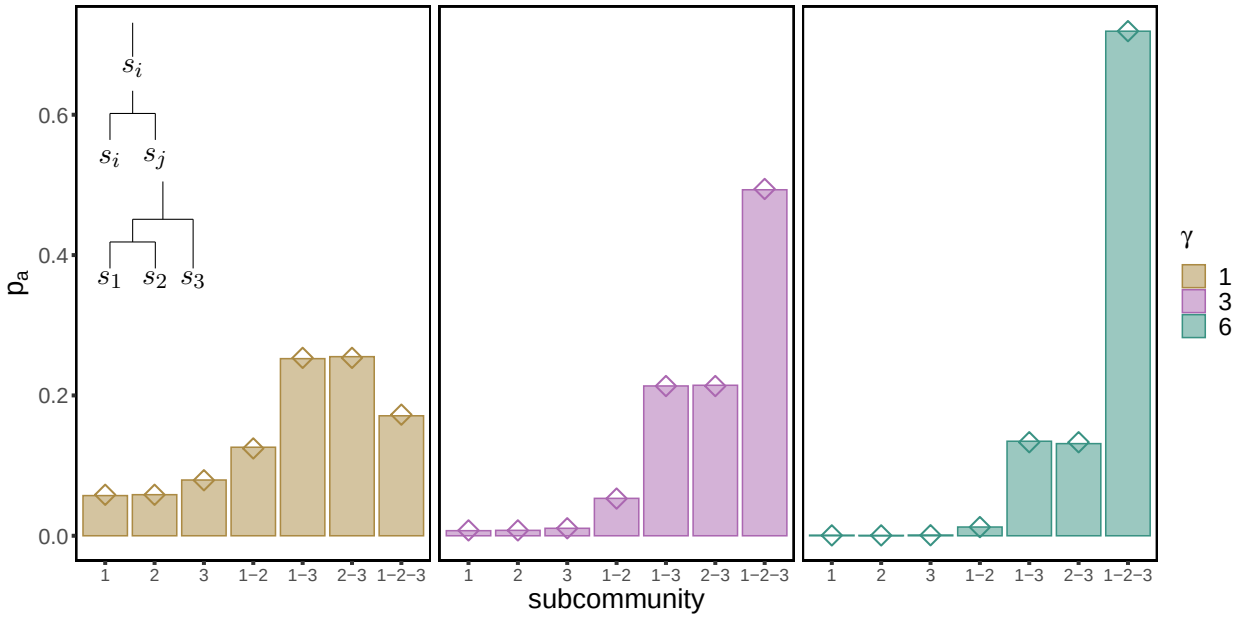

Figure 4.7: **Sub-communities of perfectly hierarchical tree** Probability that a sub-community of the three-species tree with equal branch lengths is the attractor of the dynamics. The inset shows the tree sub-structures for each of the sub-communities. Bars represent frequencies over 50000 simulations, and dots the analytical predictions.

of any sub-community is guaranteed. This result requires both the tree structure (inducing a particular interaction matrix) and the assumption that all species have equal growth rates. While we expect this result to qualitatively hold for small deviations from the assumption of identical growth rates (see Section 4.5.7), it is false in general when these requirements are not satisfied. Additionally, under the same assumptions, the abundance distribution of the community reflects the tree structure at distinct levels: high biomass is recorded for the outliers within each clade (local tree structure), and one expects an overall decreasing trend towards the nested parts of the tree (coarser structure).

When the number of traits is comparable to the number of species, our model is an instance of a Lotka-Volterra model with random interactions. The analysis of models considering random interactions between species has a long tradition in ecology [4, 52, 81], and in recent years the field has moved beyond questions concerning the stability and feasibility of the whole system, focusing instead on the study of the properties of the sub-communities

of coexisting species resulting from dynamics [10, 14, 21, 103]. In these models, one usually assumes that species interactions are independent of species identity (but see [5, 51]). The star-tree case we analyzed above satisfies this assumption, but with a stronger correlation structure than previously considered. Yet the analysis of this case shares the same flavor as previous results: while full coexistence is almost always precluded, the community settles at attractors that can contain a moderate number of species. We have derived an approximation for the mean number of survivors, and found that it depends on the ratio of traits to species, γ, and on the total ratio of shared to private mutations for each species $n\zeta$. As long as this two quantities are the same, pools of distinct sizes will yield the same distribution for the number of survivors. Contrary to previous studies [103], the analytic tractability of the model allows us to be able to derive exact expressions for the total biomass and relative abundance distribution of the system.

The general case of an arbitrary, tree-induced correlation structure provides a biologically-meaningful way in which one can relax the statistical equivalence between species. Taking advantage of the vast literature on the Wishart ensemble in fields ranging from economics to statistics [17, 66, 90], we are able to derive exact integral formulas to compute the probability of survival for a sub-community under arbitrary correlations among species. In this way, one can measure properties of the system (conditioning on a final sub-community) by numerically evaluating the integral expression. For small enough communities and simple enough phylogenies, this approach can be replicated on each sub-system to compute the marginal distribution of the properties of interest. Yet, as the number of species grows, calculations become too burdensome. As such, devising new analytical techniques to tackle the general case would not only add to the growing body of literature on random interaction models, but also advance our understanding of the effects of phylogenies on communities.

While our approach can be extended in a variety of ways, we discuss some of the most promising avenues.

First, instead of assuming that the same tree structure controls the evolution of all the traits ℓ, we can partition the traits into m classes and assume that the evolution of each class is determined by a distinct phylogenetic tree. These type of processes are studied in population genetics when either admixture or incomplete lineage sorting lead to traits that cannot be explained by a single tree [91]. In such cases, A would no longer follow the Wishart distribution but rather will be a sum of, possibly degenerate, Wishart matrices.

Second, our assumption of constant growth rates among species allowed us to examine how phylogenetic relatedness influences coexistence in a purely interaction-driven model. Indeed, we explicitly severed any link between phylogeny and fitness differences among species. When we include variation in the growth rates, we expect our results to hold only for small enough variances. Furthermore, in this case the restriction on $\ell \geq n$ can be lifted, provided that there is a background competitive effect μ strong enough to prevent divergence of the dynamics (See SI). A potential follow-up is therefore to tie also the growth rates to phylogenetic information—varying how strongly relatedness influences growth rates and interactions, one could investigate the duality of "competition" vs. "filtering" usually discussed in the literature [44, 82, 122].

Lastly, our approach assumes an explicit separation between the evolutionary processes at the regional level (which give rise to the phylogenetic structure) and the ecological interactions (that happen at the local level). To break this separation, one could model the tree generation process and ecological dynamics concurrently. For example, as done by Maynard et al. [83], one could "run" the dynamics after each new speciation event, thereby pruning the community to a coexisting sub-community. One would then take the sub-tree of that community as the starting point for the new speciation event. In this setting, in a similar manner to chapter three and the framework of adaptive dynamics [61], we have a separation of time-scales between the speciation events and the local community dynamics. Traits evolve on the tree between each pruning event. In this regard, our results provide baseline

comparisons, and even suggests patterns that would emerge from the process: assuming that the number of traits is a constant ℓ, the community cannot reach more than ℓ species, yet at the early steps of the process the ratio of traits to species could be extremely high—hence we expect that most speciation events occurring early on would not cause extinctions; in this case, the bulk of the phylogenetic structure would be built at the beginning of the process. Perturbing the growth rates slightly, one could compare the structure of this tree with the structure of the tree found by simply letting the tree generation process run, and after having the same number of speciation events let the species interact and get a coexisting sub-community.

While there has been an extensive discussion over the potential ways in which phylogenies could affect ecological differences between species, and thus influence the strength of their interactions [24], much less has been said about the patterns one would observe under a particular hypothesis. In this work, by linking phylogenies to a simple model of trait evolution and local community dynamics, we were able to fully characterize many global aspects of the community. We showed that the phylogenetic structure of the species pool and the the number of traits determining competition affect the results in concert. Our results provide an additional baseline prediction about the effect of phylogenies and traits on local communities.

4.5 Supplementary Information

4.5.1 Motivation

From consumer-resource dynamics to covariances

We start with a model of consumer-resource dynamics in which the consumers differ only in the relative preference of each resource and the resources have an homogeneous growth rate. Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^\ell$ be vectors denoting the density of predators and resources. We model the dynamics as the MacArthur's consumer-resource model [76]:

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \boldsymbol{x} \circ (-d\mathbf{1}_n + \alpha\tilde{G}\boldsymbol{y}), \\ \frac{d\boldsymbol{y}}{dt} &= \boldsymbol{y} \circ (r\mathbf{1}_\ell - \boldsymbol{y} - \beta\tilde{G}^T\boldsymbol{x}), \end{aligned} \tag{4.5}$$

where $\circ$ stands for the Hadamard (component-wise) matrix product, and $\mathbf{1}_k = (1, \ldots, 1)^T \in \mathbb{R}^k$ is a notation for a column vector whose entries are exactly k ones.

By our assumptions, matrix $\tilde{G} \in \mathbb{R}_+^{n\times\ell}$ encodes the preference distribution (alternatively, the time allocation distribution) of the predators over the resources, so that $\tilde{G}\mathbf{1}_\ell = \mathbf{1}_n$. Then by a separation of time scales, which implies that resource densities remain at equilibrium, we can model the competition between the consumers as following competitive Lotka-Volterra dynamics [76]:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{x} \circ (\alpha r\tilde{G}\mathbf{1}_\ell - d\mathbf{1}_n - \alpha\beta\tilde{G}\tilde{G}^T\boldsymbol{x}) = \boldsymbol{x} \circ ((\alpha r - d)\mathbf{1}_n - \alpha\beta\tilde{G}\tilde{G}^T\boldsymbol{x}). \tag{4.6}$$

As long as $n \leq \ell$ (besides measure zero sets) we have that matrix $\tilde{A} := \tilde{G}\tilde{G}^T$ is positive definite. This property of $\tilde{A}$ allows one to further transform the system (4.6) without affecting the set of coexisting species. In particular we can perform the following operations (see section 4.5.6 for a more detailed discussion):

(a) Rescale the growth rate, $\boldsymbol{v} = (\alpha r - d)\mathbf{1}_n$, by any positive constant.

(b) Multiply $\tilde{A}$ by a positive, constant diagonal matrix.

(c) Multiply both $\tilde{A}$ and $\boldsymbol{v}$ by a positive diagonal matrix.

Following this operations we reduce the system to

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{x} \circ (\mathbf{1}_n - \tilde{G}\tilde{G}^T\boldsymbol{x}). \tag{4.7}$$

To distinguish the effect of the mean of $\tilde{G}$, write $\tilde{G} = G + \frac{1}{n}\mathbf{1}_n\mathbf{1}_\ell^T$. Notice that this decomposition, together with the restriction $\tilde{G}\mathbf{1}_\ell = \mathbf{1}_n$, implies that $G\mathbf{1}_\ell = \mathbf{0}_n$, which means that the entries of G have zero mean —here $\mathbf{0}_k = (0, \ldots, 0)^T$ stands for a column vector formed by k zeros. Then matrix $\tilde{A}$ can be decomposed as $\tilde{A} = GG^T + \mathbf{1}_n\mathbf{1}_n^T$. Because the system in (4.7) has constant growth rates then one can show (section 4.5.6) that, as long as $\ell > n$ (the strict inequality arising due to G having rank $\ell - 1$), the set of coexisting species for (4.7) is invariant to the shift $\mathbf{1}_n\mathbf{1}_n^T$. Therefore the system reduces to:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{x} \circ (\mathbf{1}_n - GG^T\boldsymbol{x}) = \boldsymbol{x} \circ (\mathbf{1}_n - A\boldsymbol{x}), \tag{4.8}$$

where we have defined $A := GG^T$. This is the competitive, deterministic dynamics that we have assumed for the consumers throughout this chapter. Observe that the set of coexisting species remains unchanged if we define interaction matrix $A = \frac{1}{\ell}GG^T$, as in the main text, because of the aforementioned invariant operations.

Modelling the covariance matrix

From (4.8) we see that the interactions between species A_{ij} are fully determined by the row vectors $\boldsymbol{G}_i$. Because each row $\tilde{\boldsymbol{G}}_i$ of matrix $\tilde{G}$ is a preference vector, then it lies on the standard $\ell - 1$ dimensional simplex $\Delta^{\ell-1} = \{\tilde{\boldsymbol{G}}_i \in \mathbb{R}^\ell | \sum_{j=1}^{\ell} \tilde{G}_{ij} = 1, \text{ for } i = 1, \ldots, n\}$,

which implies that $\boldsymbol{G}_i$ lies on a bounded subset of a linear subspace of $\mathbb{R}^\ell$ defined by the restrictions $\sum_{j=1}^{\ell} G_{ij} = 0$ for $i = 1, \ldots, n$. By choosing a suitable (linear) coordinate system $\{\boldsymbol{w}_j\}_{j=1}^{\ell}$ we can express

$$\begin{aligned} \boldsymbol{G}_i &= \sum_{j=1}^{\ell} c_i^j \boldsymbol{w}_j, \\ A_{ij} &= \boldsymbol{G}_i \boldsymbol{G}_j^T = \sum_{k=1}^{\ell} c_i^k c_j^k. \end{aligned} \tag{4.9}$$

Therefore, the entries of A are fully determined by the coordinates of row vectors $\boldsymbol{G}_i$ on the basis $\{\boldsymbol{w}_j\}_{j=1}^{\ell}$.

To model coordinates c_i^j we assume that each (rescaled) preference vector $\boldsymbol{G}_i$ is the result of a diffusion process starting at the origin of this space (this maps back to our $\tilde{\boldsymbol{G}}$ matrix as saying that every consumer has an *homogeneous* preference for any resource). Assuming that each coordinate is independent and letting the diffusion time be small enough, then coefficients c_i^j are normally distributed, $c_i^j \sim \mathcal{N}(0, \sigma)$. The invariant properties of the model allow us to forget about the deviation σ and simply model $c_i^j \sim \mathcal{N}(0, 1)$. This shows that A satisfies the assumptions of model (4.8) up to a change of number of traits from ℓ to $\ell - 1$.

4.5.2 Deterministic Limit

Full coexistence

We provide more details for the proof that, in the deterministic limit, every subcommunity of the pool is feasible. Since every subcommunity has an interaction matrix induced by a tree, it is enough to show that feasibility is guaranteed whenever this is the case.

We proceed by induction on the number of species. For $n = 1$ the claim holds trivially. Let T be a phylogenetic tree (not necessarily ultrametric) for n species, and Σ its respective covariance matrix. Let t_1 be the time at which the first split happens, so that at t_1 the

ancestral branch splits into $m \geq 2$ lineages (L_i, with $i = 1, \ldots, m$) where each L_i contains at most $n-1$ species. Lineages are defined by the condition that species $j, k \in L_i$ if and only if the shared branch length between both species $t(j,k)$ satisfies $t(j,k) > t_1$. That is, each lineage contains the subset of species whose shared evolutionary time is strictly greater than t_1. For each L_i, take T_i to be the subtree induced by L_i. Consider $\tilde{T}$, the tree obtained by shrinking the segment between the root and t_1 to a point (see Figure 4.8), then $\tilde{T}$ is a phylogenetic tree, for which the covariance matrix $\tilde{\Sigma}$ is block diagonal and given by diagonal blocks $\tilde{\Sigma}_i$. Each $\tilde{\Sigma}_i$ is the covariance matrix of the tree $\tilde{T}_i$ which is obtained from T_i by shrinking the root branch by t_1. By induction it follows that each block $\tilde{\Sigma}_i$ is feasible, hence $\tilde{\Sigma}$ is also feasible. Observe that, going backwards, T is obtained from $\tilde{T}$ by adding a root segment of length t_1. In particular this says that the shared evolutionary times of all species increases by t_1, i.e. $\Sigma = \tilde{\Sigma} + t_1 \mathbf{1}_n \mathbf{1}_n^T$, so that Σ is a constant rank-one update of $\tilde{\Sigma}$. Then by section 4.5.6, the equilibrium associated to Σ is feasible.

Perfectly hierarchical trees

Consider a perfectly hierarchical tree T_n with n tips and branching times $t_0 = 0 < t_1 < \ldots < t_n < 1$ (see Figure 4.2), and let Σ_n be its covariance matrix. Then it follows trivially that

$$\Sigma_n = \begin{pmatrix} \tilde{\Sigma}_{n-1} & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1}^T & s_1 \end{pmatrix} + t_1 \mathbf{1}_n \mathbf{1}_n^T, \tag{4.10}$$

where $s_i := \sum_{j=i+1}^{n} \Delta t_j$, for $\Delta t_j = t_j - t_{j-1}$ the time between two branching events— the *inter-branching time*. In this subsection we find accurate bounds for the total biomass and analyze the expected abundance distribution.

Define the vector of abundances $\boldsymbol{x}_n = (x_n^i)$ for a hierarchical tree T_n with n tips. In the

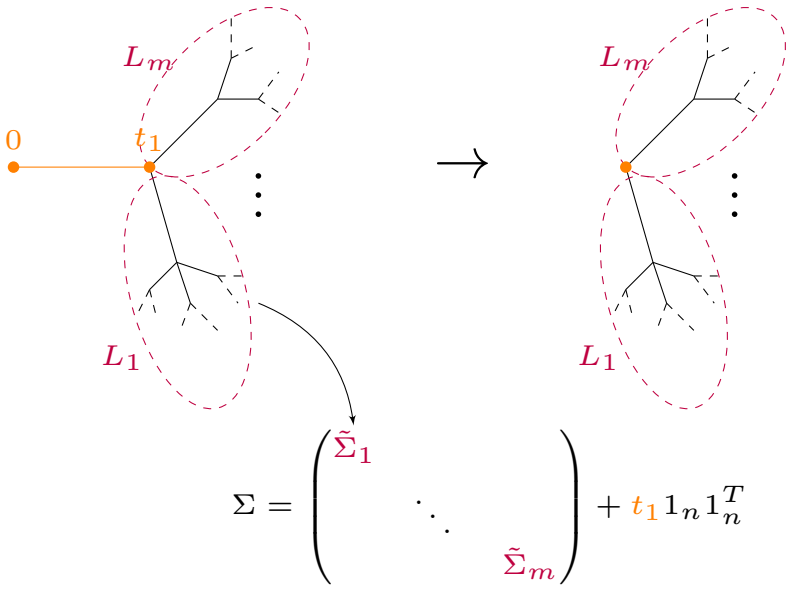

Figure 4.8: **Schematic representation of the inductive step on the proof of full coexistence**. Starting with the tree T (left), we shrink the ancestral branch up to the first splitting time t_1 to have a degenerate tree $\tilde{T}$ (on the right). $\tilde{T}$ splits at time 0 into m distinct subtrees induced by the lineages L_i for $i = 1, \ldots, m$. The covariance matrix for T, Σ is obtained from the covariance matrix $\tilde{\Sigma}$ of $\tilde{T}$ by "adding back" the ancestral branch. This amounts to a constant rank-one update of $\tilde{\Sigma}$ which preserves feasibility.

deterministic limit, this vector satisfies the linear system

$$\Sigma_n \boldsymbol{x}_n = \mathbf{1}_n. \tag{4.11}$$

As in the proof of feasibility, $\boldsymbol{x}_n$ is given recursively by the updated equilibrium abundances $\tilde{\boldsymbol{x}}_{n-1}$and s_1^{-1} of the non-interacting subtrees $\tilde{T}_{n-1}$ and the one formed by the first species, respectively. Indeed, if we look for solutions of the form $\boldsymbol{x}_n = \begin{pmatrix} a\tilde{\boldsymbol{x}}_{n-1} \\ x_n^n \end{pmatrix}$, where the vector of abundances $\tilde{\boldsymbol{x}}_{n-1}$ satisfies $\tilde{\Sigma}_{n-1}\tilde{\boldsymbol{x}}_{n-1} = \mathbf{1}_{n-1}$, $\tilde{\Sigma}_{n-1}$ being the covariance matrix of the subtree $\tilde{T}_{n-1}$, the equilibrium condition (4.11) for $\boldsymbol{x}_n$ reduces to a linear system for a and x_n^n:

$$\begin{cases} a + at_1 \mathbf{1}_{n-1}^T \tilde{\boldsymbol{x}}_{n-1} + t_1 x_n^n = 1, \\ at_1 \mathbf{1}_{n-1}^T \tilde{\boldsymbol{x}}_{n-1} + (s_1 + t_1) x_n^n = 1. \end{cases} \tag{4.12}$$

The solution is $a = s_1 x_n^n$, with $x_n^n = (s_1 + t_1 + s_1 t_1 \mathbf{1}_{n-1}^T \tilde{\boldsymbol{x}}_{n-1})^{-1}$. Let $\tilde{W}_{n-1} := \sum_{i=1}^{n-1} \tilde{x}_{n-1}^i = \mathbf{1}_{n-1}^T \tilde{\boldsymbol{x}}_{n-1}$. Then $\boldsymbol{x}_n$ can be written in terms of $\tilde{W}_{n-1}$, $\tilde{\boldsymbol{x}}_{n-1}$, $s_0 = s_1 + t_1$, and s_1 as

$$\begin{aligned} x_n^n &= \frac{1}{s_0 + t_1 \tilde{W}_{n-1} s_1}, \\ x_n^i &= \frac{s_1 \tilde{x}_{n-1}^i}{s_0 + t_1 \tilde{W}_{n-1} s_1}, \quad 1 \le i < n. \end{aligned} \tag{4.13}$$

In particular, this implies the following recurrence for the total biomass, W_n:

$$W_n = \frac{1 + \tilde{W}_{n-1} s_1}{s_0 + t_1 \tilde{W}_{n-1} s_1}. \tag{4.14}$$

In the case of equal inter-branching times, $\Delta t_i = \frac{1}{n}$ for all $i = 1, 2, \ldots, n$, observe that $s_0 = 1$, $s_1 = \frac{n-1}{n}$ and $\tilde{\Sigma}_{n-1} = \frac{n-1}{n}\Sigma_{n-1}$. Hence $\boldsymbol{x}_{n-1} = s_1 \tilde{\boldsymbol{x}}_{n-1}$ and $W_{n-1} = s_1 \tilde{W}_{n-1}$, so

Eqs. (4.13) and (4.14) above reduce to:

$$\begin{aligned} x_n^n &= \frac{n}{n+W_{n-1}}, \\ x_n^i &= \frac{n x_{n-1}^i}{n+W_{n-1}}, \quad 1 \le i < n, \\ W_n &= \frac{n(1+W_{n-1})}{n+W_{n-1}}. \end{aligned} \tag{4.15}$$

The following proposition provides accurate upper and lower bounds for total biomass in the limit of large number of species.

Proposition 2. *Let*

$$\varphi(n) := \frac{4n-1-\sqrt{16n^2+1-8n\sqrt{n-1}}}{4\sqrt{n-1}}. \tag{4.16}$$

Then, for equal branching times, it holds that $\sqrt{n}-\varphi(n) > W_n > \sqrt{n}-1/4$ *for* $n \ge 2$ *and* $\varphi(n) \to 1/4$ *in the limit* $n \to \infty$.

Proof. Direct computation shows that the inequality holds at $n = 2$ so we proceed by induction on n.

Consider first the lower bound. Suppose it holds at $n-1$, then:

$$W_n = \frac{n(1+W_{n-1})}{n+W_{n-1}} = n\left(1-\frac{n-1}{n+W_{n-1}}\right) > \frac{n(\sqrt{n-1}+3/4)}{n+\sqrt{n-1}-1/4}.$$

If the claim were not satisfied at n we would have

$$\sqrt{n}-1/4 \ge \frac{n(\sqrt{n-1}+3/4)}{n+\sqrt{n-1}-1/4}.$$

Rearranging terms, this gives the following chain of equivalent inequalities:

$$\begin{aligned} n\sqrt{n}+\sqrt{n-1}\sqrt{n}+\frac{1}{16} &\geq n\sqrt{n-1}+n+\frac{1}{4}(\sqrt{n-1}+\sqrt{n}), \\ n(\sqrt{n}-1)+\sqrt{n-1}\sqrt{n}(1-\sqrt{n})+\frac{1}{16} &\geq \frac{1}{4}(\sqrt{n-1}+\sqrt{n}), \\ \sqrt{n}(\sqrt{n}-1)(\sqrt{n}-\sqrt{n-1})+\frac{1}{16} &\geq \frac{1}{4}(\sqrt{n-1}+\sqrt{n}). \end{aligned} \tag{4.17}$$

Multiplying both sides by $\sqrt{n-1}+\sqrt{n}$ we get

$$\sqrt{n}(\sqrt{n}-1)+\frac{1}{16}(\sqrt{n-1}+\sqrt{n}) \geq \frac{1}{4}(\sqrt{n-1}+\sqrt{n})^2 = \frac{1}{4}(2n-1+2\sqrt{n-1}\sqrt{n}). \tag{4.18}$$

The last inequality implies

$$\frac{3}{4} \geq \frac{7}{8}\sqrt{n},$$

which says $n \leq 1$. This is a contradiction and we are done.

We proceed in the similar way for the upper bound. By induction hypothesis at $n-1$ we have

$$W_n < \frac{n(\sqrt{n-1}+1-\varphi(n))}{n+\sqrt{n-1}-\varphi(n)}.$$

If the inequality is not satisfied at n then, a similar chain of inequalities yields

$$n-\sqrt{n}+\varphi(n)^2(\sqrt{n}+\sqrt{n-1}) \leq \varphi(n)(2n-1+2\sqrt{n-1}\sqrt{n}). \tag{4.19}$$

Note that the above restriction is exactly the same as (4.18) with the inequality reversed and changing $\varphi(n)$ instead of $1/4$. Using that $\sqrt{n} > \sqrt{n-1}$, the last inequality implies

$$n-\sqrt{n}+2\sqrt{n-1}\varphi(n)^2-(4n-1)\varphi(n) \leq 0.$$

In particular, this means that $\varphi(n) \leq u$ for u the smaller root of the above quadratic equation,

$$u := \frac{4n - 1 - \sqrt{16n^2 - 8n + 1 - 8n\sqrt{n-1} + 8\sqrt{n-1}\sqrt{n}}}{4\sqrt{n-1}},$$

but with this definition and (4.16) it is easy to see that

$$u > \frac{4n - 1 - \sqrt{16n^2 + 1 - 8n\sqrt{n-1}}}{4\sqrt{n-1}} = \varphi(n),$$

which is again a contradiction and this completes the proof for the upper bound.

We have just proved that $\sqrt{n} - \varphi(n) > W_n > \sqrt{n} - 1/4$. In particular, this implies that $\varphi(n) < 1/4$. Taking the limit in the numerator of expression (4.16) it is easy to see that the leading order is

$$\begin{aligned}\lim_{n\to\infty} 4n - 1 - \sqrt{16n^2 + 1 - 8n\sqrt{n-1}} &= \lim_{n\to\infty} \frac{(4n-1)^2 - (16n^2 + 1 - 8n\sqrt{n-1})}{4n - 1 + \sqrt{16n^2 + 1 - 8n\sqrt{n-1}}} \\ &= \lim_{n\to\infty} \sqrt{n-1},\end{aligned}$$

which shows that

$$\lim_{n\to\infty} \varphi(n) = \frac{1}{4} \tag{4.20}$$

and the proof is complete. □

Note that, for large communities, a very good approximation for the total biomass in a perfectly hierarchical tree is given by the formula $W_n = \sqrt{n} - \frac{1}{4}$.

The recursions in (4.15) for individual abundances can be easily solved in terms of total biomass W_n as

$$x_n^i = \prod_{j=i}^{n} \frac{j}{j + W_{j-1}}. \tag{4.21}$$

This formula gives the abundance of the i-th species (in increasing order of the tips) for $i \geq 2$

(observe that the first two species have the same abundance). Alternatively,

$$\log(x_n^i) = \sum_{j=i}^{n} \log\left(\frac{j}{j + W_{j-1}}\right) = -\sum_{j=i}^{n} \log\left(1 + \frac{W_{j-1}}{j}\right).$$

Approximating W_{j-1} by its lower bound, $W_{j-1} \approx \sqrt{j-1} - 1/4$, we find

$$\log(x_n^i) \approx -\sum_{j=i}^{n} \log\left(1 + \frac{\sqrt{j-1} - 1/4}{j}\right). \tag{4.22}$$

Cutting the series for $\log(1+x)$ at second order and considering only the leading term, with respect to j for the quadratic term, we get:

$$\log(x_n^k) \approx -\sum_{j=k}^{n} \frac{\sqrt{j-1}}{j} - \frac{1}{4j} - \frac{1}{2}\frac{j-1}{j^2} \approx -\sum_{j=k}^{n} \frac{1}{\sqrt{j}} - \frac{3}{4j}. \tag{4.23}$$

By the Euler-Maclaurin formula we obatin:

$$\log(x_n^k) \approx 2(\sqrt{n} - \sqrt{j-1}) + \frac{3}{4}(\log(n) - \log(j-1)). \tag{4.24}$$

and we can further refine the first terms x_n^k for k small by replacing the actual value W_j.

Perfectly balanced tree

The total biomass for perfectly balanced trees is easier to derive because the covariance matrix has constant row sums in that case. To show this statement, order tree splits by the time they happen ($t_1 < \ldots < t_q$). At each time t_i, the number of lineages doubles, so we get a total of $n = 2^q$ species. As species interact by their shared evolutionary time, in this case each species shares the time with 2^{q-k} other species. Now let $s_k = \sum_{i=1}^{k} \Delta t_i$, Δt_i being the inter-branching time —compare the different notation for s_k here and in the previous subsection. Summing over all possible split times we get the sum over any row of A (observe

that $A_{ii} = 1$),

$$r_q = \sum_{j=1}^{n} A_{ij} = 1 + \sum_{k=1}^{q} 2^{q-k} s_k, \tag{4.25}$$

which is independent of i. Because row sums are constant, the vector or equilibrium abundances can be written as $\boldsymbol{x}_n = x\mathbf{1}_n$, and substitution into $\Sigma_n \boldsymbol{x}_n = \mathbf{1}_n$ yields $r_q x = 1$. Therefore, individual abundances at equilibrium are constant and given by $x = r_q^{-1}$. Consequently, the total biomass at equilibrium, W_q, is simply given by

$$W_q = \frac{2^q}{1 + \sum_{k=1}^{q} 2^{q-k} s_k}. \tag{4.26}$$

By our assumption of ultrametric trees, we have $s_k < 1$ (we need to add the tip lengths to sum up to one). In the particular case of equal inter-branching times, $\Delta t_i = \frac{1}{q+1}$, then $s_k = \frac{k}{q+1}$ and

$$r_q = 1 + \frac{2^{q-1}}{q+1} \sum_{k=1}^{q} \frac{k}{2^{k-1}}. \tag{4.27}$$

Observe that

$$\sum_{k=1}^{q} \frac{k}{2^{k-1}} = \frac{\partial}{\partial x} \left(\frac{1 - x^{q+1}}{1 - x} \right) \Big|_{x=\frac{1}{2}} = 4\left(1 - \frac{1}{2^q}\left(q + 1 - \frac{q}{2}\right)\right). \tag{4.28}$$

Thus,

$$r_q = 1 + \frac{2^{q+1} - q - 2}{q+1} = \frac{2^{q+1} - 1}{q+1}, \tag{4.29}$$

and the total biomass reads

$$W_q = \frac{q+1}{2 - 2^{-q}}. \tag{4.30}$$

Let $n = 2^q$ be the number of species, then the number of tree splits is $q = \log_2(n)$. In terms of the number of species, the formula is given by

$$W_n = \frac{\log_2(n) + 1}{2 - 1/n}, \tag{4.31}$$

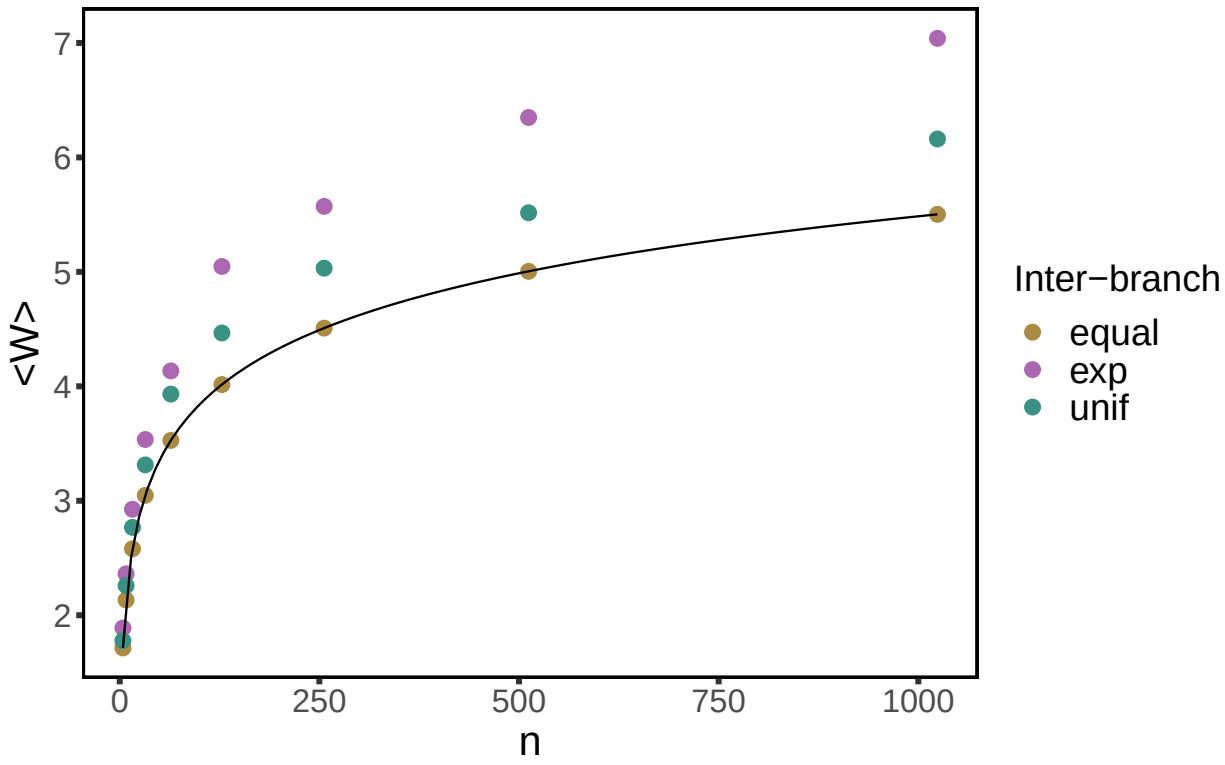

Figure 4.9: **Total abundance for the perfectly balanced tree**

which grows logarithmically with n. Figure 4.9 compares the case of perfectly balanced tress for equal branching times with two cases, in which sampling times are drawn from exponential and uniform distributions.

4.5.3 Number of species that survive

We have shown above that, in the $\ell \to \infty$ limit, full coexistence is guaranteed. To study species coexistence for finite $\ell \geq n$ we use the fact that A follows the Wishart distribution. As in [103], first we will compute the probability of the equilibrium point being feasible, i.e., where all species survive. Second, since the attractor is unique (it is the only saturated equilibrium point that appears), we can calculate the probability that the equilibrium point cannot be invaded by the remaining species in the pool. Then we will show that the probability of feasibility and non-invasibility factors into the corresponding product, which yields

the distribution of the number of species that coexist, as well as the expected number of species that survive.

Because matrix $A = GG^T$ is symmetric and positive definite, it is diagonally-stable [58], which implies that generalized Lotka-Volterra dynamics exhibits a single, globally stable fixed point [58], so there is a unique endpoint for the dynamics. Let us write the equilibrium abundances of the attractor, formed by m survivors, as

$$\boldsymbol{x}_n = \begin{pmatrix} \boldsymbol{x}_m \\ \\ \boldsymbol{0}_{n-m} \end{pmatrix}, \tag{4.32}$$

where, without loss of generality, we have located the survivors as the first m species. Let $\{S\}_m$ denote the set of species that survive (i.e., the support of the endpoint). Therefore, the attractor can be fully characterized by two conditions [103]:

- Define the vector $\boldsymbol{z}_n = \boldsymbol{1}_n - A\boldsymbol{x}_n = (x_n^i)$ with components z_n^i. Then it holds: first, $z_n^i = 0$ for all species $i \in \{S\}_m$, which simply states that equilibrium abundances of survivors satisfy the linear system $A_m \boldsymbol{x}_m = \boldsymbol{1}_m$, for A_m the submatrix of A restricted to the support $\{S\}_m$. Second, it also holds that $z_n^i < 0$ for all species $i \notin \{S\}_m$, i.e., the fixed point *cannot be invaded* by the remaining species outside the endpoint. We have, therefore, a fixed point that cannot be invaded.

- The equilibrium point hast to be *feasible*, i.e., $\boldsymbol{x}_m > \boldsymbol{0}_m$ —here we use the notation that vectors $\boldsymbol{a} > \boldsymbol{b}$ if all inequalities are satisfied component-wise.

Since matrix A belongs to the Wishart ensemble, these two conditions are to be understood in statistical terms. In the following subsections we are going to compute exact formulae for the probability that all the species in the pool form a *feasible* attractor, and the probability that an endpoint formed by m species remains *non-invasible*. Using the properties of the Wishart ensemble [90], we will calculate separately the probabilities of feasibility and non-

invasibility, and with them we will obtain the distribution of the number of species that survive.

Probability of feasibility

Let n be the number of species in the community and ℓ the number of traits, and define $\gamma := \ell/n$ as the ratio between the number of traits and the size of the pool. An equilibrium point for the system such that all species coexist satisfies:

$$A\boldsymbol{x}_n = \mathbf{1}_n, \text{ with } x_n^i > 0 \text{ for all } i = 1, \ldots, n. \tag{4.33}$$

The probability of feasibility is then the probability that $A^{-1}\mathbf{1}_n$ has all entries greater than 0. Observe that interaction matrix is defined as $A = \frac{1}{\ell}GG^T$ in the main text. Since rescaling by a positive constant in A does not affect the condition for feasibility, we can forget about the rescaling by the number of traits ℓ.

Let $A \sim \mathcal{W}_n(\Sigma, \ell)$ and $L_{n-1} = (I_{n-1}, \mathbf{0}_{n-1})$ be a rectangular $(n-1) \times n$ matrix with 0 in its last column, I_k being the $k \times k$ identity matrix. Then equation (2) of [66] (similarly stated in the proof of Theorem 1 in [17]) implies that

$$\widetilde{\boldsymbol{x}} := \frac{L_{n-1}A^{-1}\mathbf{1}_n}{\mathbf{1}_n^T A^{-1}\mathbf{1}_n} \sim t_{n-1}\left(\ell - n + 2, \frac{L_\ell \Sigma^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n}, \frac{L_{n-1}R_1L_{n-1}^T}{(\ell - n + 2)\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n}\right), \tag{4.34}$$

where $t_p(\nu, \boldsymbol{\mu}, \Lambda)$ is a multivariate, p-dimensional t distribution with ν degrees of freedom, localization vector $\boldsymbol{\mu}$ and dispersion matrix Λ [115]. Matrix R_1 is given by

$$R_1 = \Sigma^{-1} - \frac{\Sigma^{-1}\mathbf{1}_n\mathbf{1}_n^T\Sigma^{-1}}{\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n}. \tag{4.35}$$

Up to a normalization by a positive constant (which is precisely the total biomass, $\mathbf{1}_n^T A^{-1}\mathbf{1}_n$, given that A is positive definite), vector $\widetilde{\boldsymbol{x}} = (\tilde{x}_i)$ precisely gives the abundances of the *first*

$n-1$ species. Moreover, the last (normalized) abundance is expressed as $1-\mathbf{1}_{n-1}^T\widetilde{\boldsymbol{x}}$, so the probability of feasibility turns out to be

$$P_{\rm f}(n)=\int d^{n-1}\widetilde{\boldsymbol{x}}f(\widetilde{\boldsymbol{x}})\Theta(1-\mathbf{1}^T\widetilde{\boldsymbol{x}})\prod_{i=1}^{n-1}\Theta(\tilde{x}_i), \tag{4.36}$$

for $f(\widetilde{\boldsymbol{x}})$ the probability density function of the multivariate t distribution defined in (4.34).

Because a multivariate t distribution is the ratio between a multivariate Gaussian and the square root of a chi-square distribution, it holds that if $\widetilde{\boldsymbol{x}}\sim t_p(\nu,\boldsymbol{\mu},\Lambda)$, then we have that $\widetilde{\boldsymbol{x}}=\boldsymbol{y}/\sqrt{u/\nu}+\boldsymbol{\mu}$, where $\boldsymbol{y}\sim\mathcal{N}(\mathbf{0},\Lambda)$ is a multivariate Gaussian and $u\sim\chi^2_\nu$, which is independent of $\boldsymbol{y}$. Therefore, conditioning on u, we find that $\boldsymbol{y}_u:=\widetilde{\boldsymbol{x}}|u\sim\mathcal{N}(\boldsymbol{\mu},\nu\Lambda/u)$ and we can transform the integral above to get

$$P_{\rm f}(n)=\int_0^\infty du\,g(\nu,u){\rm Pr}(\boldsymbol{y}_u>\mathbf{0}_{n-1},\mathbf{1}_{n-1}^T\boldsymbol{y}_u<1), \tag{4.37}$$

where $u\sim\chi^2_\nu$, $g(\nu,u)$ is the corresponding pdf with $\nu=\ell-n+2$, and the random variable $\boldsymbol{y}_u$ is distributed as a multivariate normal,

$$\boldsymbol{y}_u\sim\mathcal{N}\left(\frac{L_{n-1}\Sigma^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n},\frac{L_{n-1}R_1L_{n-1}^T}{u\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n}\right). \tag{4.38}$$

In this way, all the dependence in the number of traits ℓ remains included in the chi-square distribution. Eqs. (4.37) and (4.38) yield the probability of feasibility for an arbitrary covariance matrix Σ. An explicit calculation of the probability of feasibility amount to evaluating the probability ${\rm Pr}(\boldsymbol{y}_u>\mathbf{0}_{n-1},\mathbf{1}_{n-1}^T\boldsymbol{y}_u<1)$. This can be done explicitly for the case of constant, non-negative correlation.

Constant, non-negative correlation

Consider the covariance matrix $\Sigma = (1-\rho)I_n + \rho \mathbf{1}_n \mathbf{1}_n^T$ with $\rho \geq 0$. Then (4.38) simplifies to:

$$y_u \sim \mathcal{N}\left(\frac{1}{n}\mathbf{1}_{n-1}, \frac{1-\rho+n\rho}{un(1-\rho)}\left(I_{n-1} - \frac{1}{n}\mathbf{1}_{n-1}\mathbf{1}_{n-1}^T\right)\right). \tag{4.39}$$

Let us define

$$\alpha_u := \frac{1-\rho+n\rho}{un(1-\rho)} \text{ and } \beta_u := \frac{\alpha_u}{n}. \tag{4.40}$$

In this way, the covariance matrix Σ_u in (4.39) can be expressed as $\Sigma_u = \alpha_u I_{n-1} - \beta_u \mathbf{1}_{n-1}\mathbf{1}_{n-1}^T$. Σ_u has two eigenvalues, α_u and $\alpha_u + (n-1)\beta_u$. The first has multiplicity $n-1$, and the second 1. Hence the determinant follows immediately,

$$|\Sigma_u| = \alpha_u^{n-2}(\alpha_u - (n-1)\beta_u). \tag{4.41}$$

The inverse can be easily calculated:

$$\Sigma_u^{-1} = \frac{1}{\alpha_u}\left(I + \frac{\beta_u}{\alpha_u - (n-1)\beta_u}\mathbf{1}_{n-1}\mathbf{1}_{n-1}^T\right). \tag{4.42}$$

Therefore we can write the pdf for the random variable y_u as

$$\begin{aligned} f_u(\boldsymbol{y}) &= Ke^{-\frac{1}{2}\left(\boldsymbol{y}-\frac{1}{n}\mathbf{1}_{n-1}\right)^T \Sigma_u^{-1}\left(\boldsymbol{y}-\frac{1}{n}\mathbf{1}_{n-1}\right)} \\ &= Ke^{-\frac{1}{2\alpha_u}\left(\left\|\boldsymbol{y}-\frac{1}{n}\mathbf{1}_{n-1}\right\|^2 + \frac{\beta_u}{\alpha_u-(n-1)\beta_u}\left(\mathbf{1}_{n-1}^T\left(\boldsymbol{y}-\frac{1}{n}\mathbf{1}_{n-1}\right)\right)^2\right)} \end{aligned} \tag{4.43}$$

for $K = (2\pi)^{-(n-1)/2}|\Sigma_u|^{-1/2}$. First we have to compute the probability

$$p(u) := \Pr(\boldsymbol{y}_u > \mathbf{0}_{n-1}, \mathbf{1}_{n-1}^T \boldsymbol{y}_u < 1) = \int_{\mathbb{R}^{n-1}} d^{n-1}\boldsymbol{y} f_u(\boldsymbol{y}) \Theta(1 - \mathbf{1}_{n-1}^T \boldsymbol{y}) \prod_{i=1}^{n-1} \Theta(y_i), \tag{4.44}$$

with $\Theta(x)$ the Heaviside step function, defined as $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ if $x < 0$. Thus after a change of variables $\boldsymbol{y}' = \boldsymbol{y} - \frac{1}{n}\mathbf{1}_{n-1}$, we have

$$p(u) = K \int_{\mathbb{R}^{n-1}} d^{n-1}\boldsymbol{y} e^{-\frac{1}{2\alpha_u}\left(\|\boldsymbol{y}\|^2 + (\mathbf{1}_{n-1}^T \boldsymbol{y})^2\right)} \Theta\left(\frac{1}{n} - \mathbf{1}_{n-1}^T \boldsymbol{y}\right) \prod_{i=1}^{n-1} \Theta\left(y_i + \frac{1}{n}\right), \tag{4.45}$$

where we have omitted primes to ease notation and we have used (4.40) to see that

$$\frac{\beta_u}{\alpha_u - (n-1)\beta_u} = 1. \tag{4.46}$$

To simplify the term $(\mathbf{1}_{n-1}^T \boldsymbol{y})^2$ in the exponential, we introduce a Dirac's delta function,

$$p(u) = K \int_{\mathbb{R}^{n-1}} d^{n-1}\boldsymbol{y} \int_{\mathbb{R}} d\omega e^{-\frac{1}{2\alpha_u}\left(\|\boldsymbol{y}\|^2 + \omega^2\right)} \delta(\omega - \mathbf{1}_{n-1}^T \boldsymbol{y}) \Theta\left(\frac{1}{n} - \omega\right) \prod_{i=1}^{n-1} \Theta\left(y_i + \frac{1}{n}\right), \tag{4.47}$$

and use its integral representation,

$$\delta(\omega - \mathbf{1}_{n-1}^T \boldsymbol{y}) = \frac{1}{2\pi} \int_{\mathbb{R}} d\xi e^{-i\xi(\omega - \mathbf{1}_{n-1}^T \boldsymbol{y})}. \tag{4.48}$$

This transformation, together with an interchange in the order of integration, yields the following expression for $p(u)$:

$$p(u) = \frac{K}{2\pi} \int_{\mathbb{R}} d\omega \int_{\mathbb{R}} d\xi \int_{\mathbb{R}^{n-1}} d^{n-1}\boldsymbol{y} e^{-\frac{1}{2\alpha_u}\left(\|\boldsymbol{y}\|^2 + \omega^2\right) + i(\mathbf{1}_{n-1}^T \boldsymbol{y} - \omega)\xi} \Theta\left(\frac{1}{n} - \omega\right) \times \prod_{i=1}^{n-1} \Theta\left(y_i + \frac{1}{n}\right). \tag{4.49}$$

Apparently we are increasing the complexity of the integral, but rearranging terms we observe that

$$p(u) = \frac{K}{2\pi}\int_{\mathbb{R}} d\xi \int_{\mathbb{R}} d\omega e^{-\frac{\omega^2}{2\alpha_u} - i\omega\xi}\Theta\Big(\frac{1}{n} - \omega\Big)\int_{\mathbb{R}^{n-1}} d^{n-1}\boldsymbol{y} e^{-\frac{\|\boldsymbol{y}\|^2}{2\alpha_u} + i\xi \mathbf{1}_{n-1}^T \boldsymbol{y}} \prod_{i=1}^{n-1}\Theta\Big(y_i + \frac{1}{n}\Big), \tag{4.50}$$

and the integral over $\boldsymbol{y}$ factorizes,

$$p(u) = \frac{K}{2\pi}\int_{\mathbb{R}} d\xi \int_{-\infty}^{1/n} d\omega e^{-\frac{\omega^2}{2\alpha_u} - i\omega\xi}\left(\int_{-1/n}^{\infty} dy e^{-\frac{y^2}{2\alpha_u} + iy\xi}\right)^{n-1}. \tag{4.51}$$

Now, in the integral over ω, change to the variable $\omega' = -\omega$ to get

$$\begin{aligned} p(u) &= \frac{K}{2\pi}\int_{\mathbb{R}} d\xi \int_{-1/n}^{\infty} d\omega e^{-\frac{\omega^2}{2\alpha_u} + i\omega\xi}\left(\int_{-1/n}^{\infty} dy e^{-\frac{y^2}{2\alpha_u} + iy\xi}\right)^{n-1} \\ &= \frac{K}{2\pi}\int_{\mathbb{R}} d\xi \left(\int_{-1/n}^{\infty} dy e^{-\frac{y^2}{2\alpha_u} + iy\xi}\right)^{n}. \end{aligned} \tag{4.52}$$

Let

$$\Phi(x) := \frac{1}{2}\Big(1 + \operatorname{erf}(x/\sqrt{2})\Big) \tag{4.53}$$

be the cdf of the standard Gaussian distribution, which can be extended to the complex plane. Then it holds that

$$\int_{-1/n}^{\infty} dy e^{-\frac{y^2}{2\alpha_u} + iy\xi} = \sqrt{2\pi\alpha_u}\, e^{-\frac{\alpha_u\xi^2}{2}}\Phi\left(\frac{1/n + i\alpha_u\xi}{\sqrt{\alpha_u}}\right). \tag{4.54}$$

Therefore, the sought probability can be written as

$$p(u) = \frac{K(2\pi\alpha_u)^{n/2}}{2\pi}\int_{\mathbb{R}} d\xi e^{-\frac{n\alpha_u\xi^2}{2}}\Phi\left(\frac{1/n + i\alpha_u\xi}{\sqrt{\alpha_u}}\right)^n. \tag{4.55}$$

An alternative way to express the integral over ξ it is to consider a path Γ in the complex plane such that $\Gamma = \{z \in \mathbb{C} | z = x_0 + i\xi\}$ and then reducing the result to the limit $x_0 \to 0$, so that the integral over the imaginary axis is well defined. In practice, this amounts to change to the variable $\zeta = i\xi$. Consequently, an equivalent form of writing this equation is

$$p(u) = -i\sqrt{\frac{n\alpha_u}{2\pi}} \int_\Gamma d\zeta e^{\frac{n\alpha_u \zeta^2}{2}} \Phi\left(\frac{1/n + \alpha_u \zeta}{\sqrt{\alpha_u}}\right)^n, \tag{4.56}$$

where we have used that $K = \sqrt{n}(2\pi\alpha_u)^{-(n-1)/2}$ in this case. Finally, according to (4.37), in the case of constant, positive correlation the probability of feasibility is given by a two dimensional integral,

$$P_{\mathrm{f}}(n) = -i\sqrt{\frac{n}{2\pi}} \int_0^\infty du\, g(\nu, u)\sqrt{\alpha_u} \int_\Gamma d\zeta e^{\frac{n\alpha_u \zeta^2}{2}} \Phi\left(\frac{1/n + \alpha_u \zeta}{\sqrt{\alpha_u}}\right)^n, \tag{4.57}$$

where $g(\nu, u)$ is the pdf of the chi-square distribution with $\nu = \ell - n + 2$ degrees of freedom. Figure 4.10 compares this exact formula with numerical simulation for different values of the correlation.

Probability of non-invasibility

In this subsection we compute the probability that an attractor formed by $m \leq n$ species cannot be invaded by the remaining $n - m$ species. Let $A \sim W_n(\Sigma, \ell)$. Observe that for invasibility the rescaling of interaction matrix as $A = \frac{1}{\ell} GG^T$ does not matter. Partition matrices A and Σ in four blocks as follows:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{4.58}$$

where Σ_{11} refers to the species that belong to the support $\{S\}_m$ of the attractor, Σ_{22} is related to those species outside the attractor, and off-diagonal matrices are formed by the

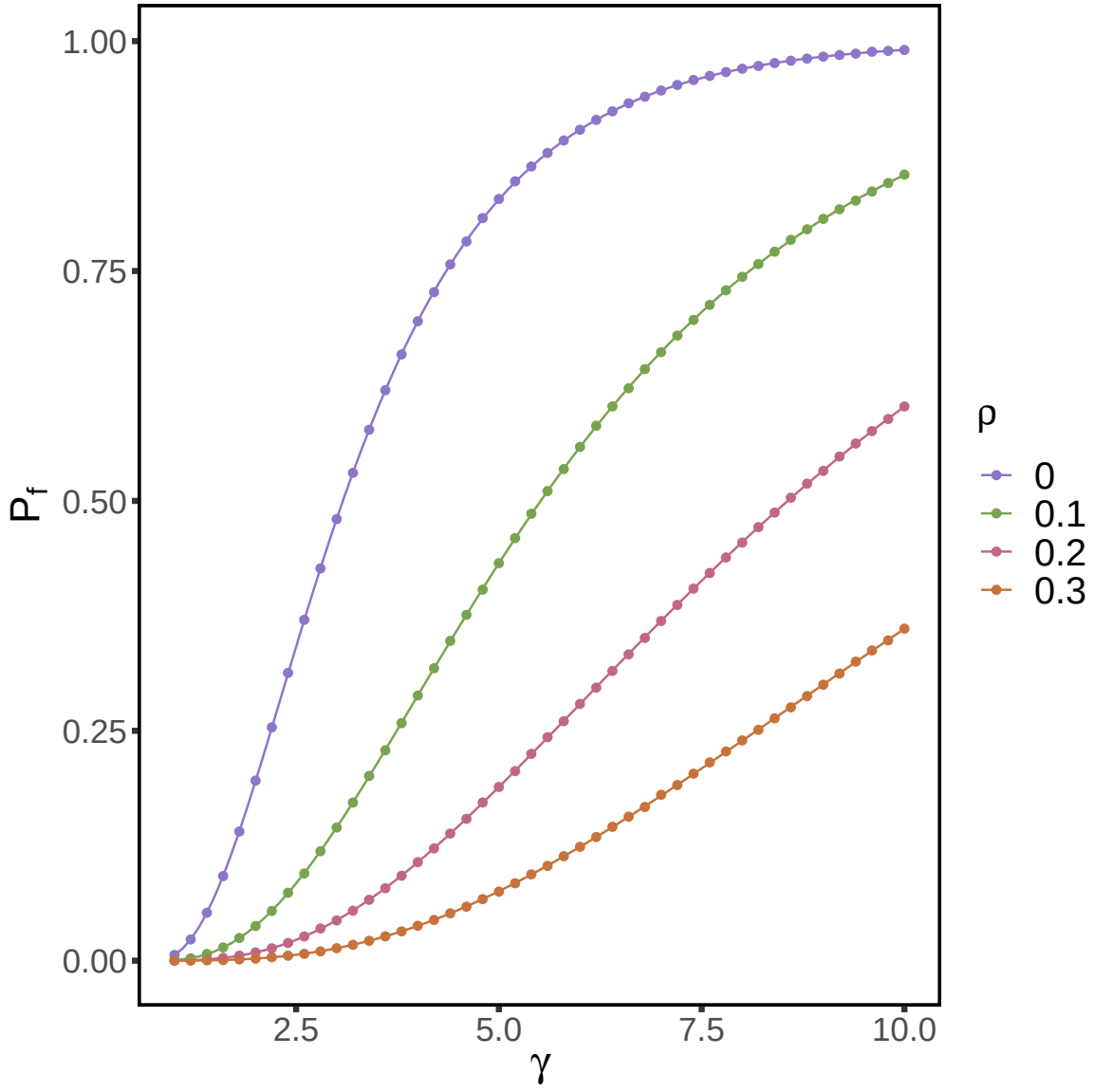

Figure 4.10: **Probability of feasibility as a function of the ratio γ of number of traits to number of species for different *constant* correlation matrices.** The simulations were done with $n = 10$ species. Dots are simulations, solid lines are numerical evaluations of the exact formula (4.57). The larger the correlation, the slower curves approach to one in the deterministic limit $\gamma \to \infty$.

corresponding rows and columns in $\{S\}_m$ and $\{S\}_n \setminus \{S\}_m$, and *vice versa*. The exact same notation applies to blocks in A.

Then by theorem 3.2.10 of [90] we have that

$$A_{21}|A_{11} \sim \mathcal{N}(\Sigma_{21}\Sigma_{11}^{-1}A_{11}, \Sigma_{22.1} \otimes A_{11}), \tag{4.59}$$

where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is the Schur complement of Σ_{22}, $\otimes$ is the tensor product of matrices, and the normal distribution appearing is meant to be understood as the distribution of the *flatten* matrix A_{21}. By the properties of the normal distribution it follows that

$$\begin{aligned} A_{21}A_{11}^{-1}|A_{11} &\sim \mathcal{N}(\Sigma_{21}\Sigma_{11}^{-1}, \Sigma_{22.1} \otimes A_{11}^{-1}), \\ A_{21}A_{11}^{-1}\mathbf{1}_m|A_{11} &\sim \mathcal{N}(\Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, \mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m \Sigma_{22.1}). \end{aligned} \tag{4.60}$$

In order to get the last line, we first transpose the matrix, then notice that the $\mathbf{1}_m^T$ operator acts on the vector of elements of the matrix as $I_m \otimes \mathbf{1}^T$. Hence by the property $(A \otimes B)(C \otimes D) = AC \otimes BD$ of the tensor product the second statement above follows.

As mentioned at the beginning of Sec. 4.5.3, the probability that the attractor cannot be invaded by any species in $\{S\}_n \setminus \{S\}_m$ coincides with the probability that $\boldsymbol{z} = \mathbf{1}_{n-m} - A_{21}A_{11}^{-1}\mathbf{1}_m < \mathbf{0}_{n-m}$. Define $W := \mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m$ and $f_W(w)$ as the pdf of the random variable W, which is non-negative. Then

$$\begin{aligned} P_{\text{ni}}(m,n) &= \int_0^\infty dw f_W(w) \Pr(\boldsymbol{z} < \mathbf{0}|W = w) \\ &= \int_0^\infty dw f_W(w) \int_{\mathcal{V}_w^+} dA_{11} \Pr(A_{11}|W = w) \Pr(\boldsymbol{z} < \mathbf{0}|A_{11}, W = w), \end{aligned} \tag{4.61}$$

where $\mathcal{V}^+$ is the set of positive definite symmetric matrices and $\mathcal{V}_w^+$ the set conditional to $W = \mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m = w$. Using that $\boldsymbol{z} = \mathbf{1}_{n-m} - A_{21}A_{11}^{-1}\mathbf{1}_m$ and (4.60), the conditional

variable $\boldsymbol{z}|A_{11}, W = w$ is distributed as

$$\boldsymbol{z}|A_{11}, W = w \sim \mathcal{N}\Big(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, w\Sigma_{22.1}\Big), \tag{4.62}$$

which does not depend explicitly on A_{11}. Neither does $\Pr(\boldsymbol{z} < \mathbf{0}|A_{11}, W = w)$, so we can factor this probability out of the integration over A_{11}. In this way, we can write

$$P_{\text{ni}}(m, n) = \int_0^\infty dw f_W(w) Q_{n-m}^-\big(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, w\Sigma_{22.1}\big), \tag{4.63}$$

because $\int_{\mathcal{V}_w^+} dA_{11}\Pr(A_{11}|W = w) = 1$. In (4.63) we have defined Q_p^- as the probability that a multivariate Gaussian variable with the specified parameters is contained in the fully negative orthant,

$$Q_p^-\left(\boldsymbol{\mu}, \Lambda\right) := (2\pi)^{-p/2}|\Lambda|^{-1/2}\int_{\mathbb{R}_-^n} d\boldsymbol{y} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Lambda^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}. \tag{4.64}$$

Corollary 3.2.6 in [90] implies that $A_{11} \sim \mathcal{W}_m(\Sigma, \ell)$. Therefore, theorem 3.2.12 in the same reference holds, which ensures that

$$W^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m = \frac{\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m}{\mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m} \sim \chi^2_{\ell-m+1}. \tag{4.65}$$

This means that

$$g(\nu', w) = -w^{-2}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m f_W\big(w^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m\big), \tag{4.66}$$

for $g(\nu, w)$ the pdf of a $\chi^2_{\nu'}$ distribution with $\nu' = \ell - m + 1$ degrees of freedom. Now, making the change of variable $w' = w^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m$ in (4.63) we finally get

$$P_{\text{ni}}(m, n) = \int_0^\infty dw g(\nu', w) Q_{n-m}^-\big(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, w^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m\Sigma_{22.1}\big). \tag{4.67}$$

As for the case of feasibility, (4.67) is an exact formula for the probability that an endpoint

composed by m species cannot be invaded by the remaining $n-m$ species. Similarly, the multidimensional integral associated to Q^-_{n-m} can be reduced to a single integral in the case of constant, non-negative correlation, as we show in the following subsection. Thus, in that particular case, the probability of non-invasibility is expressed as a double integral.

Constant, non-negative correlation

In the case of constant, non-negative correlation, (4.67) simplifies to:

$$P_{\text{ni}}(m) = \int_0^\infty dw g(\nu', w) Q^-_{n-m}(\boldsymbol{\mu}, \Sigma_w) \tag{4.68}$$

with

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1-\rho}{1-\rho+m\rho}\mathbf{1}_{n-m}, \\ \Sigma_w &= \frac{m(1-\rho)}{w(1-\rho+m\rho)}\left(I_{n-m} + \frac{\rho}{1-\rho+m\rho}\mathbf{1}_{n-m}\mathbf{1}^T_{n-m}\right). \end{aligned} \tag{4.69}$$

Now focus on the probability Q^-_{n-m}. Making the substitution $\boldsymbol{y'} = k\boldsymbol{y}$ in (4.64) it is easy to show that

$$Q^-_p(\boldsymbol{\mu}, \Lambda) = Q^-_p(\boldsymbol{\mu}/k, \Lambda/k^2). \tag{4.70}$$

Therefore, for $k = \frac{m(1-\rho)}{1-\rho+m\rho}$ we recover Eq. (4.84) with $\boldsymbol{\mu}$ and Λ given by

$$\boldsymbol{\mu} = \frac{1}{m}\mathbf{1}_{n-m}, \quad \Sigma_w = \frac{1-\rho+m\rho}{mw(1-\rho)}\left(I_{n-m} + \frac{\rho}{1-\rho+m\rho}\mathbf{1}_{n-m}\mathbf{1}^T_{n-m}\right). \tag{4.71}$$

Now let us write $\Sigma_w := \alpha_w I_{n-m} + \beta_w \mathbf{1}_{n-m}\mathbf{1}^T_{n-m}$, with $\alpha_w := \frac{1-\rho+m\rho}{mw(1-\rho)}$, $\beta_w := \frac{\rho\alpha_w}{1-\rho+m\rho}$. As we did for the probability of feasibility, the probability Q^-_{n-m} can be written as a one-dimensional integral. For that is crucial that, contrary to what happened in the case of feasibility, correlations given by Σ_w are positive —notice the plus sign in (4.71). This is due to the special structure of Σ_w, which implies that the correlation between any two distinct y_i, y_j in (4.64) is constant and given by $\lambda = \frac{\rho}{1+m\rho} \geq 0$. Hence, the following result of Tong [115]

(section 8.2.5) applies:

Proposition 3. *Let $\boldsymbol{x}$ be distributed according to $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ such that covariance matrix entries satisfy $\Sigma_{ii} = \sigma_i^2$ and $\Sigma_{ij} = \sigma_i \sigma_j \lambda$. Then, the joint probability that $\boldsymbol{x} \in C := \{\boldsymbol{x} \in \mathbb{R}^n | b_i \leq x_i \leq a_i, i = 1, \ldots, n\}$, where $-\infty \leq b_i < a_i \leq \infty$ for i=1,..., n, is expressed as*

$$\Pr(\boldsymbol{x} \in C) = \int_{-\infty}^{\infty} dy \phi(y) \prod_{i=1}^{n} \left[\Phi\left(\frac{(a_i - \mu_i)/\sigma_i + \sqrt{\lambda} y}{\sqrt{1-\lambda}} \right) - \Phi\left(\frac{(b_i - \mu_i)/\sigma_i + \sqrt{\lambda} y}{\sqrt{1-\lambda}} \right) \right] \tag{4.72}$$

for $\phi(z)$ and $\Phi(z)$ the pdf and cdf, respectively, of a univariate standard normal distribution.

In our particular case $\sigma_i^2 = \frac{1+m\rho}{wm(1-\rho)}$, $\lambda = \frac{\rho}{1+m\rho}$, $b_i = -\infty$, $a_i = 0$ and, according to (4.71), $\mu_i = \frac{1}{m}$ for $i = 1, \ldots, n-m$. Therefore, putting all the pieces together, we can write

$$P_{\text{ni}}(m, n) = \int_0^{\infty} dw g(\nu', w) \int_{-\infty}^{\infty} dy \phi(y) \Phi\left(\frac{-1/m + y\sqrt{\beta_w}}{\sqrt{\alpha_w}} \right)^{n-m}. \tag{4.73}$$

As for the probability of feasibility, in the case of constant, non-negative correlation we can reduce it to a two-dimensional integral.

Notice the resemblance between the expressions for feasibility and non-invasibility — Eqs. (4.57) and (4.73). In the case of $\rho > 0$, by a changing $y \to y' \frac{\alpha_w}{\sqrt{\beta_w}}$, we can make the resemblance stronger:

$$P_{\text{ni}}(m, n) = \sqrt{\frac{1-\rho+m\rho}{2\pi\rho}} \int_0^{\infty} dw g(\nu', w) \sqrt{\alpha_w} \int_{-\infty}^{\infty} dy e^{-\frac{(1-\rho+m\rho)\alpha_w y^2}{2\rho}} \times \Phi\left(\frac{-1/m + y\alpha_w}{\sqrt{\alpha_w}} \right)^{n-m}. \tag{4.74}$$

Observe that the number of degrees of freedom of the $\chi^2_{\nu'}$ distribution here is $\nu' = \ell - m + 1$. Notice also that the change of variables leading to (4.74) does not apply for $\rho = 0$. This case is trivial, however, and will not be discussed explicitly.

Sign independence of Feasibility and Invasibility

In this section we show that the joint probability of feasibility and non-invasibility factors into the product of the two probabilities calculated above. For that purpose, it suffices to show that

$$\Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m\big) = \Pr(\boldsymbol{z} < \mathbf{0}_{n-m}). \tag{4.75}$$

For that purpose we can calculate

$$\begin{aligned}\Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m\big) &= \int_0^\infty dw\, g_W(w) \Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m, W = w\big) \\ &= \int_0^\infty dw\, g_W(w) \int_{\mathcal{G}_w^+} dA_{11} \Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}, W = w\big) \Pr\big(A_{11} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m, W = w\big),\end{aligned} \tag{4.76}$$

where $W = \mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m$ as for the calculation of P_{ni}, and g_W is the pdf of the random variable $W|A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m$. In the second line we have introduced an integral over the set $\mathcal{G}_w^+$ of symmetric matrices and positive definite that verify the conditions $A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m$ and $W = \mathbf{1}_m^T A_{11}^{-1}\mathbf{1}_m = w$. As before, by (4.62) we can factor the probability $\Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}, W = w\big)$ out, so we get

$$\Pr\big(\boldsymbol{z} < \mathbf{0}_{n-m} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m\big) = \int_0^\infty dw g_W(w) Q_{n-m}^-\big(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, w\Sigma_{22.1}\big), \tag{4.77}$$

which coincides with (4.67) except for the probability density g_W. In the last step we have used the normalization condition $\int_{\mathcal{G}_w^+} dA_{11} \Pr(A_{11} | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m, W = w) = 1$.

Observe that the condition $A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m$ is equivalent to the conditions $\mathbf{1}_{m-1}^T \widetilde{\boldsymbol{x}} < 1$ and $\widetilde{\boldsymbol{x}} > \mathbf{0}_{m-1}$, for $\widetilde{\boldsymbol{x}}$ the vector of the first $m-1$ relative abundances defined in (4.34). Let $R := \{\boldsymbol{v} \in \mathbb{R}^{m-1} | \mathbf{1}_{m-1}^T \boldsymbol{v} < 1, \boldsymbol{v} > \mathbf{0}_{m-1}\}$ the set of vectors satisfying the two last

conditions. Then it is easy to see that

$$g_W(w) = \frac{d}{dw}\Pr\big(W < w | A_{11}^{-1}\mathbf{1}_m > \mathbf{0}_m\big)$$
$$= \frac{d}{dw}\Pr\big(W < w | \widetilde{\boldsymbol{x}} \in R\big) = \frac{d}{dw}\Pr(W < z) = f_W(w). \quad (4.78)$$

The last equality in the chain above follows because W and $\widetilde{\boldsymbol{x}}$ are independent random variables —see the proof of theorem 1 in [17].

This shows that the probability of observing and endpoint with m survivors can be factored as the probability of feasibility (4.37) times the probability (4.67) that the attractor cannot be invaded by the remaining $n - m$ species in the pool.

Distribution of the number of coexisting species

Due to the independence shown in the previous section, the probability that the system settles in a subset $\{S\}_m \subset \{1, \ldots, n\}$ formed by m species is simply

$$\Pr(\{S\}_m | n, \ell, \Sigma) = \binom{n}{m} P_{\mathrm{a}}(m, n) = \binom{n}{m} P_{\mathrm{f}}(m) P_{\mathrm{ni}}(m, n), \quad (4.79)$$

because all subsets with cardinality m are statistically equivalent.

Assuming constant and non-negative correlation, in Figure 4.11 we compare numerical integration of Eqs. (4.57) and (4.73) appearing in (4.79) with simulations.

Average number of species

In this section we will focus on the case of constant correlation. Our aim is to approximate the integrals for feasibility and invasibility in the large number of species limit by a saddle point technique. With these approximations, we provide an analytical way to compute the probability of coexistence $\Pr(\{S\}_m | n, \ell, \rho)$ —cf. Eq. (4.79)— as well as an approximation

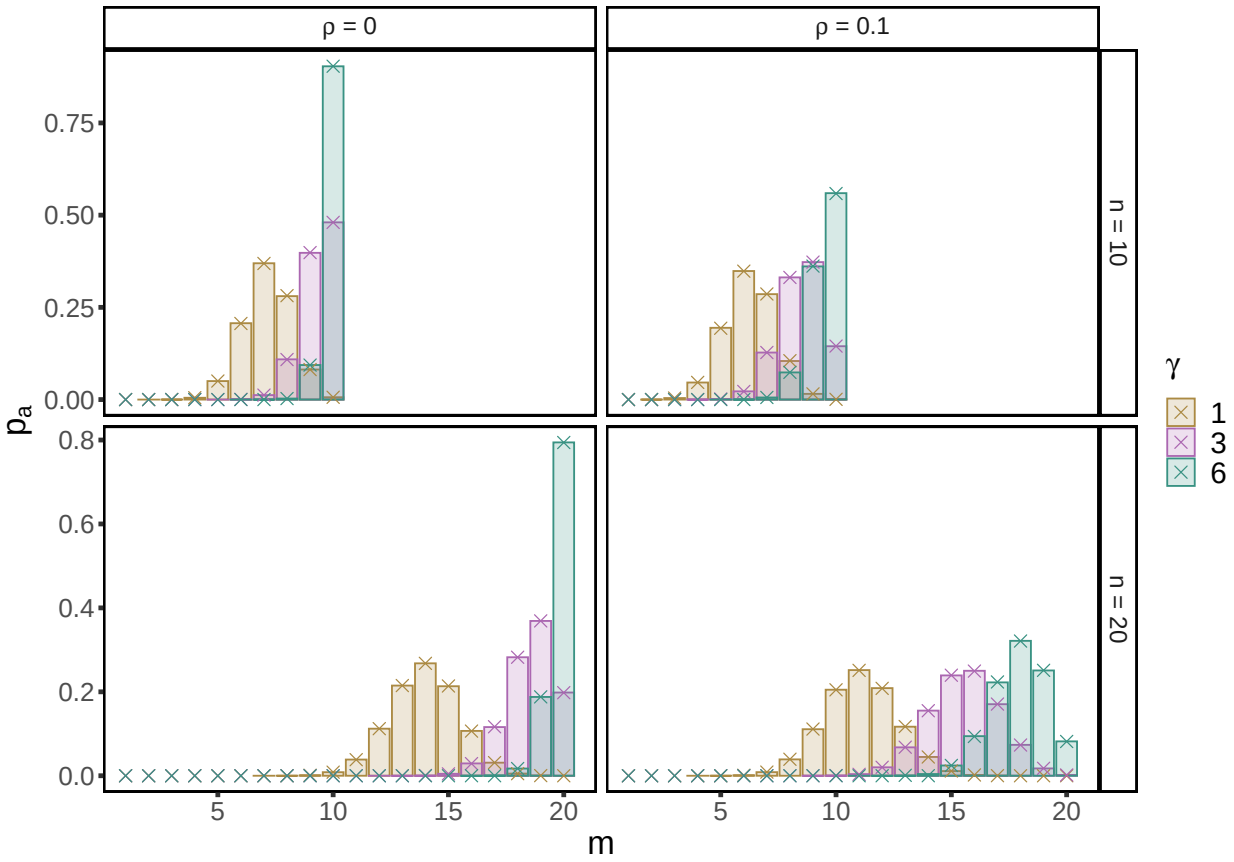

Figure 4.11: **Distribution of the set of coexisting species as a function of the ratio γ of number of traits to number of species for different *constant* correlation matrices.** The simulations were done with $n = 10, 20$ species. Bar are simulations, crosses are numerical evaluations of formula (4.79).

for the average fraction of species

$$\wp(n,\ell,\rho) := \frac{1}{n}\sum_{m=0}^{n}\binom{n}{m} m P_{\rm a}(m,n). \tag{4.80}$$

We distinguish the cases $\rho > 0$ and $\rho = 0$ for invasibility. For $\rho > 0$ we use expression (4.74). Let us define $q := m/n$ as the fraction of survivors, and recall that $\ell = n\gamma$. Also let

$$\lambda_q := mw\alpha_w = 1 + \frac{m\rho}{1-\rho} = 1 + \frac{nq\rho}{1-\rho}. \tag{4.81}$$

In terms of λ_q, the probability of non-invasibility reads

$$P_{\rm ni}(m,n) = \frac{\lambda_q}{\sqrt{2\pi(\lambda_q-1)}}\int_0^\infty dw g(\nu,w) w^{-1/2}\int_{-\infty}^{\infty} dy e^{-\frac{y^2\lambda_q^2}{2w(\lambda_q-1)}}\times \\ \Phi\left(-\sqrt{\frac{w}{m\lambda_q}} + y\sqrt{\frac{\lambda_q}{mw}}\right)^{n-m}. \tag{4.82}$$

Now we make a change of variables,

$$\begin{aligned} w' &= \sqrt{\frac{w}{m}}, \\ \frac{y'}{w'} &= \frac{y}{\sqrt{wm}}. \end{aligned} \tag{4.83}$$

Then the integral becomes

$$P_{\rm ni}(m,n) = \frac{2\lambda_q}{\sqrt{2\pi(\lambda_q-1)}}\int_0^\infty dw m^{3/2} g(\nu', mw^2)\int_{-\infty}^{\infty} dy e^{-\frac{my^2\lambda_q^2}{2w^2(\lambda_q-1)}}\times \\ \Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)^{n-m}. \tag{4.84}$$

Recall that the probability density function $g(\nu', x)$, for $\nu' = \ell - m + 1$, is:

$$g(\nu, x) = \frac{x^{(\ell-m-1)/2} e^{-x/2}}{2^{(\ell-m+1)/2}\Gamma((\ell-m+1)/2)} \tag{4.85}$$

Hence the integral (4.84) is

$$\begin{aligned} P_{\text{ni}}(m,n) =& \frac{\lambda_q m}{\sqrt{\pi(\lambda_q-1)}} \frac{(m/2)^{(\ell-m)/2}}{\Gamma((\ell-m+1)/2)} \int_0^\infty dw w^{\ell-m-1} e^{-mw^2/2} \\ & \times \int_{-\infty}^{\infty} dy e^{-\frac{my^2\lambda_q^2}{2w^2(\lambda_q-1)}} \Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)^{n-m} \\ =& \frac{\lambda_q m}{\sqrt{\pi(\lambda_q-1)}} \frac{(m/2)^{(\ell-m)/2}}{\Gamma((\ell-m+1)/2)} \int_0^\infty dw w^{-1} \int_{-\infty}^{\infty} dy e^{nF_{\text{ni}}(w,y)}, \end{aligned} \tag{4.86}$$

where the exponent $F_{\text{ni}}(w, y)$ has been defined as

$$F_{\text{ni}}(w,y) := (\gamma - q)\log(w) - \frac{qw^2}{2} - \frac{qy^2\lambda_q^2}{2w^2(\lambda_q-1)} + (1-q)\log\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right). \tag{4.87}$$

Now we evaluate the double integral in the limit $n \to \infty$ via a saddle-point technique. For that purpose, since the exponential becomes peaked around the maximum of the exponent, we calculate the equations to be satisfied by the critical point. Taking derivatives of the exponent we get

$$\begin{aligned} \frac{\partial F_{\text{ni}}}{\partial y} &= -\frac{qy\lambda_q^2}{w^2(\lambda_q-1)} + \frac{(1-q)\sqrt{\lambda_q}}{w} \frac{\phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)}{\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)}, \\ \frac{\partial F_{\text{ni}}}{\partial w} &= \frac{\gamma - q}{w} - qw + \frac{qy^2\lambda_q^2}{w^3(\lambda_q-1)} - (1-q)\left(\frac{1}{\sqrt{\lambda_q}} + \frac{y\sqrt{\lambda_q}}{w^2}\right) \frac{\phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)}{\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)}. \end{aligned} \tag{4.88}$$

Therefore at a critical point $(w^\star, y^\star)$ we have the following conditions:

$$\begin{aligned} -\frac{qy\lambda_q^{3/2}}{w(\lambda_q-1)} + (1-q)\frac{\phi\Big(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\Big)}{\Phi\Big(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\Big)} &= 0, \\ \gamma - q - qw^2 - \frac{qy\lambda_q}{\lambda_q - 1} &= 0. \end{aligned} \tag{4.89}$$

Similarly we can rewrite the integral for the probability that an endpoint formed by m species is feasible, see Eq. (4.57), as

$$P_{\mathrm{f}}(m) = -i\sqrt{\frac{\lambda_q}{2\pi}}\int_0^\infty du g(\nu,u)u^{-1/2}\int_\Gamma d\zeta e^{\frac{\lambda_q\zeta^2}{2u}}\Phi\bigg(\sqrt{\frac{u}{m\lambda_q}} + \zeta\sqrt{\frac{\lambda_q}{mu}}\bigg)^m, \tag{4.90}$$

where now the number of degrees of freedom is $\nu = \ell - m + 2$.

Following essentially the same procedure as before, i.e. making a change of variables and replacing the density function for the χ^2_ν distribution we get

$$P_{\mathrm{f}}(m) = -im^{3/2}\sqrt{\frac{\lambda_q}{2\pi}}\frac{(m/2)^{(\ell-m)/2}}{\Gamma((\ell-m)/2+1)}\int_{-\infty}^\infty du\int_\Gamma d\zeta e^{nF_{\mathrm{f}}(u,\zeta)}, \tag{4.91}$$

with the exponent

$$F_{\mathrm{f}}(u,\zeta) := (\gamma - q)\log(u) - \frac{qu^2}{2} + \frac{q\lambda_q\zeta^2}{2u^2} + q\log\Phi\bigg(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\bigg). \tag{4.92}$$

Similarly, the conditions satisfied by the critical point $(u^\star, \zeta^\star)$ are

$$\begin{aligned} \frac{\zeta\sqrt{\lambda_q}}{u} + \frac{\phi\Big(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\Big)}{\Phi\Big(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\Big)} &= 0, \\ \gamma - q - qu^2 - q\zeta &= 0. \end{aligned} \tag{4.93}$$

Notice that the product of the densities of the χ^2 distributions in each integral —

Eqs. (4.86) and (4.91)— introduce an extra term which scales exponentially with $m = nq$, namely

$$\frac{m^{\ell-m}}{2^{\ell-m}\Gamma((\ell-m)/2+1)\Gamma((\ell-m)/2+1/2)} = \frac{m^{\ell-m}}{\Gamma(\ell-m+1)}. \tag{4.94}$$

Using the Stirling's asymptotic form of the gamma function we get

$$\frac{m^{\ell-m}}{\Gamma(\ell-m+1)} \sim \frac{e^{n(\gamma-q)(1+\log q-\log(\gamma-q))}}{\sqrt{2\pi n(\gamma-q)}}. \tag{4.95}$$

Let

$$F_{\mathrm{e}}(q) := (\gamma-q)(1+\log q-\log(\gamma-q)) \tag{4.96}$$

and

$$F_{\mathrm{c}}(q) := -q\log q-(1-q)\log(1-q), \tag{4.97}$$

$F_{\mathrm{c}}(q)$ being the exponent appearing in Stirling's asymptotic formula for the binomial coefficient $\binom{n}{nq}$. Consequently the probability that the system settles in an endpoint with $m = nq$ species is given, up to a normalization factor, by:

$$\Pr(\{S\}_m|n,\ell,\rho) = \binom{n}{m} P_{\mathrm{a}}(m,n) \sim \exp\{n(F_{\mathrm{f}}(u^\star,\zeta^\star,q) + F_{\mathrm{ni}}(w^\star,y^\star,q) + F_{\mathrm{e}}(q) + F_{\mathrm{c}}(q))\}. \tag{4.98}$$

Observe that critical point coordinates $u^\star$, $\zeta^\star$, $w^\star$ and $y^\star$ depend implicitly on q through (4.89) and (4.93). Observe that one can use the asymptotic expansion (4.98) to obtain numerically the distribution of the number of survivors, $\Pr(\{S\}_m|n,\ell,\rho)$, up to a normalization factor. The calculation amounts to solve numerically the non-linear systems (4.89) and (4.93).

We are now ready to provide an analytical approximation for the mean fraction of survivors $\wp$, cf. Eq. (4.80). In the limit of large pool size n, we can approximate the mean of the distribution $\Pr(\{S\}_m|n,\ell,\rho)$ by its mode, which is easier to compute. In fact, to calculate the mode of the distribution q in the large n limit we need to find the $q^\star$ value that maximizes the exponent in (4.98). Due to the critical point conditions for $(u^\star,\zeta^\star)$ and

$(w^\star, y^\star)$, $q^\star$ satisfies

$$\frac{\partial F_{\mathrm{f}}}{\partial q} + \frac{\partial F_{\mathrm{ni}}}{\partial q} + \frac{\partial F_{\mathrm{e}}}{\partial q} + \frac{\partial F_{\mathrm{c}}}{\partial q} = 0. \tag{4.99}$$

Evaluated at the critical points $(u^\star, \zeta^\star)$ and $(w^\star, y^\star)$, the derivatives read

$$\begin{aligned}
\frac{\partial F_{\mathrm{ni}}}{\partial q} &= -\log(w) - \frac{w^2}{2} - \frac{y^2\lambda_q}{2w^2} + \frac{y}{2} - \log\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right), \\
\frac{\partial F_{\mathrm{f}}}{\partial q} &= -\log(u) - \frac{u^2}{2} + \lambda_q\frac{\zeta^2}{2u^2} + \frac{\zeta(\lambda_q - 1)}{2\lambda_q} + \log\Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right), \\
\frac{\partial F_{\mathrm{e}}}{\partial q} &= \log\left(\frac{\gamma - q}{q}\right) + \frac{\gamma - q}{q} = \log\left(\frac{\gamma - q}{q}\right) + \frac{u^2}{2} + \frac{w^2}{2} + \frac{q\zeta}{2} + \frac{qy\lambda_q}{2(\lambda_q - 1)}, \\
\frac{\partial F_{\mathrm{c}}}{\partial q} &= \log(1 - q) - \log q.
\end{aligned} \tag{4.100}$$

Therefore the condition for $q^\star$ reduces to

$$-\log\left(\frac{qwu}{\gamma - q}\right) + \frac{\lambda_q}{2}\left(\frac{\zeta^2}{u^2} - \frac{y^2}{w^2}\right) + \frac{2\lambda_q - 1}{2}\left(\frac{y}{\lambda_q - 1} + \frac{\zeta}{\lambda_q}\right) + \log\frac{(1-q)\Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}{q\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)} = 0. \tag{4.101}$$

A direct calculation shows that, at $wu = \frac{\gamma - q}{q}$, the terms up to the last logarithm vanish. We now show that the last term can be written as $\left(wu - \frac{\gamma - q}{q}\right)h$ for some function h.

Indeed, using conditions (4.93) and (4.89) we have

$$\frac{(1-q)\phi(-w, -y, q)}{q\Phi(-w, -y, q)} - \frac{\phi(u, \zeta, q)}{\Phi(u, \zeta, q)} = \frac{(u + w)\sqrt{\lambda_q}}{uw}\left(\frac{\gamma - q}{q} - uw\right), \tag{4.102}$$

where we have used the abbreviations $\Phi(u, \zeta, q) := \Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)$ and $\phi(u, \zeta, q) := \phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)$ to simplify notation. Therefore,

$$\frac{(1-q)\Phi(u, \zeta, q)}{q\Phi(-w, -y, q)} = \frac{\phi(u, \zeta, q)}{\phi(-w, -y, q)} + \frac{(u + w)\Phi(u, \zeta, q)\sqrt{\lambda_q}}{uw\phi(-w, -y, q)}\left(\frac{\gamma - q}{q} - uw\right). \tag{4.103}$$

Letting $\mu_q := (\gamma - q)/q$, it holds that

$$\frac{\phi(u,\zeta,q)}{\phi(-w,-y,q)} = e^{(\mu_q^2-(uw)^2)((\lambda_q-1)^2u^2-\lambda_q^2w^2)/(2\lambda_q u^2w^2)}. \tag{4.104}$$

Now, due to the series representation of the exponential function we have

$$\frac{\phi(u,\zeta,q)}{\phi(-w,-y,q)} = 1 + (\mu_q - uw)h(u,w), \tag{4.105}$$

where

$$h(u,w) := \frac{q(u+w)\Phi(u,\zeta,q)\sqrt{\lambda_q}}{uw\phi(-w,-y,q)} + \sum_{j=1}^{\infty}\frac{1}{j!}(\mu_q - uw)^{j-1}\left((\mu_q+uw)\frac{(\lambda_q-1)^2u^2-\lambda_q^2w^2}{2\lambda_q u^2w^2}\right)^j. \tag{4.106}$$

Thus, the claim follows by using the series expansion of $\log(1+x)$. Therefore, all the terms in (4.101) vanish at $uw = \mu_q$.

We have just shown that the last logarithm in (4.101) is equal to zero. Consequently $q^\star$ satisfies

$$\frac{(1-q)\Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}{q\Phi\left(-\frac{w}{\sqrt{\lambda_q}} + \frac{y}{w}\sqrt{\lambda_q}\right)} = 1. \tag{4.107}$$

At the point $uw = \mu_q$ we can write

$$\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q} = \frac{\lambda_q w - (\lambda_q - 1)u}{\sqrt{\lambda_q}} = \frac{w}{\sqrt{\lambda_q}} - \frac{y}{w}\sqrt{\lambda_q}, \tag{4.108}$$

which in turn implies that

$$\Phi\left(\frac{\lambda_q w - (\lambda_q - 1)u}{\sqrt{\lambda_q}}\right) = q^\star. \tag{4.109}$$

Let $\hat{q} := \Phi^{-1}(q^\star) = \sqrt{2}\mathrm{erf}^{-1}(2q^\star - 1)$, for erf^{-1} the inverse error function. Then it holds

that $(\lambda_q w - (\lambda_q - 1)u)/\sqrt{\lambda_q} = \hat{q}$ and using eq. (4.93) we can solve for $u^\star, w^\star$ in terms of $\hat{q}$, yielding

$$\begin{aligned} u^\star &= \sqrt{\lambda_q}\left(\frac{\phi(\hat{q})}{q^\star} + \hat{q}\right), \\ w^\star &= \frac{1}{\sqrt{\lambda_q}}\left((\lambda_q - 1)\frac{\phi(\hat{q})}{q^\star} + \lambda_q \hat{q}\right). \end{aligned} \tag{4.110}$$

The final condition for $q^\star$ at the saddle point reduces to substitute the expressions above into the condition $uw = \mu_q$, which finally reads

$$\frac{\gamma}{q^\star} = 1 + \left(\frac{\phi(\Phi^{-1}(q^\star))}{q^\star} + \Phi^{-1}(q^\star)\right)\left(\frac{\phi(\Phi^{-1}(q^\star))}{q^\star}(\lambda_{q^\star} - 1) + \Phi^{-1}(q^\star)\lambda_{q^\star}\right). \tag{4.111}$$

The case $\rho = 0$ for invasibility is similar, and simpler.

Level Curves

Eq. (4.111) gives a very good approximation to the level curves on the (ρ, γ) plane mapping to constant mean fraction of survivors $q = m/n$. This implicit condition can be rewritten equivalently as

$$\gamma = q + \Phi^{-1}(q)H(q) + \frac{n\rho H(q)^2}{1 - \rho}, \tag{4.112}$$

where $H(q) := \phi(\Phi^{-1}(q)) + q\Phi^{-1}(q)$. This condition is compared with simulation results in Figure 4.4 of the main text (right panel).

4.5.4 Total biomass distribution at endpoints

The proof of independence of invasibility and feasibility (section 4.5.3) also shows that, for any fixed size m of a subset of species and total biomass w, we have that $\Pr(\boldsymbol{z}_{n-m} < \mathbf{0}_{n-m} | \boldsymbol{x}_m > \mathbf{0}_m, W = w) = \Pr(\boldsymbol{z}_{n-m} < \mathbf{0}_{n-m} | W = w)$. This remark, together with the independence of W and $\boldsymbol{x}_m > \mathbf{0}_m$ (feasibility), helps us derive the distribution of total biomass. To simplify notation we do not rescale the interaction matrix by ℓ (as shown in

section 4.5.6 this would amount to a rescaling of total biomass $w \to \ell w$). The cdf for the random variable W is precisely

$$\Pr(W < w) = \sum_{m=0}^{n} \binom{n}{m} P_{\rm a}(m,n) \Pr(W < w|m), \tag{4.113}$$

where $\Pr(W < w|m)$ is the probability that $W < w$ conditional on the m-species endpoint is feasible and non-invasible. Thus,

$$\begin{aligned} \Pr(W < w|m) &= \frac{\Pr(W < w, \boldsymbol{x}_m > \mathbf{0}_m, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m})}{P_{\rm a}(m,n)} \\ &= \frac{\Pr(W < w, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m}|\boldsymbol{x}_m > \mathbf{0}_m) P_{\rm f}(m)}{P_{\rm a}(m,n)} \\ &= \frac{\Pr(W < w, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m}) P_{\rm f}(m)}{P_{\rm a}(m,n)}, \end{aligned} \tag{4.114}$$

the last equality following from the statement in the paragraph above. Now, using the notations introduced in the last section, it holds that

$$\begin{aligned} \Pr(W < w, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m}) = \int_0^\infty & du g(\nu', u) \Theta(u - w^{-1} \mathbf{1}_m^T \Sigma_{11}^{-1} \mathbf{1}_m) \\ & \times Q_{n-m}^-(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, u^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m\Sigma_{22.1}). \end{aligned} \tag{4.115}$$

Hence, using (4.114) and $P_{\rm a}(m,n) = P_{\rm f}(m) P_{\rm ni}(m,n)$, the probability density function of the biomass distribution can be expressed as

$$\begin{aligned} g_{\rm a}(w) &= \sum_{m=0}^{n} \binom{n}{m} P_{\rm f}(m) \frac{\partial \Pr(W < w, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m})}{\partial w} \\ &= \sum_{m=0}^{n} \binom{n}{m} \frac{\tilde{w}}{w} P_{\rm f}(m) g(\nu', \tilde{w}) Q_{n-m}^-(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, \tilde{w}^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m\Sigma_{22.1}), \end{aligned} \tag{4.116}$$

where $\tilde{w} := w^{-1}\mathbf{1}_m^T\Sigma_{11}^{-1}\mathbf{1}_m$. Figure 4.12 shows the comparison of (4.116) with simulations for the constant correlation case in the case in which the interaction matrix is rescaled by

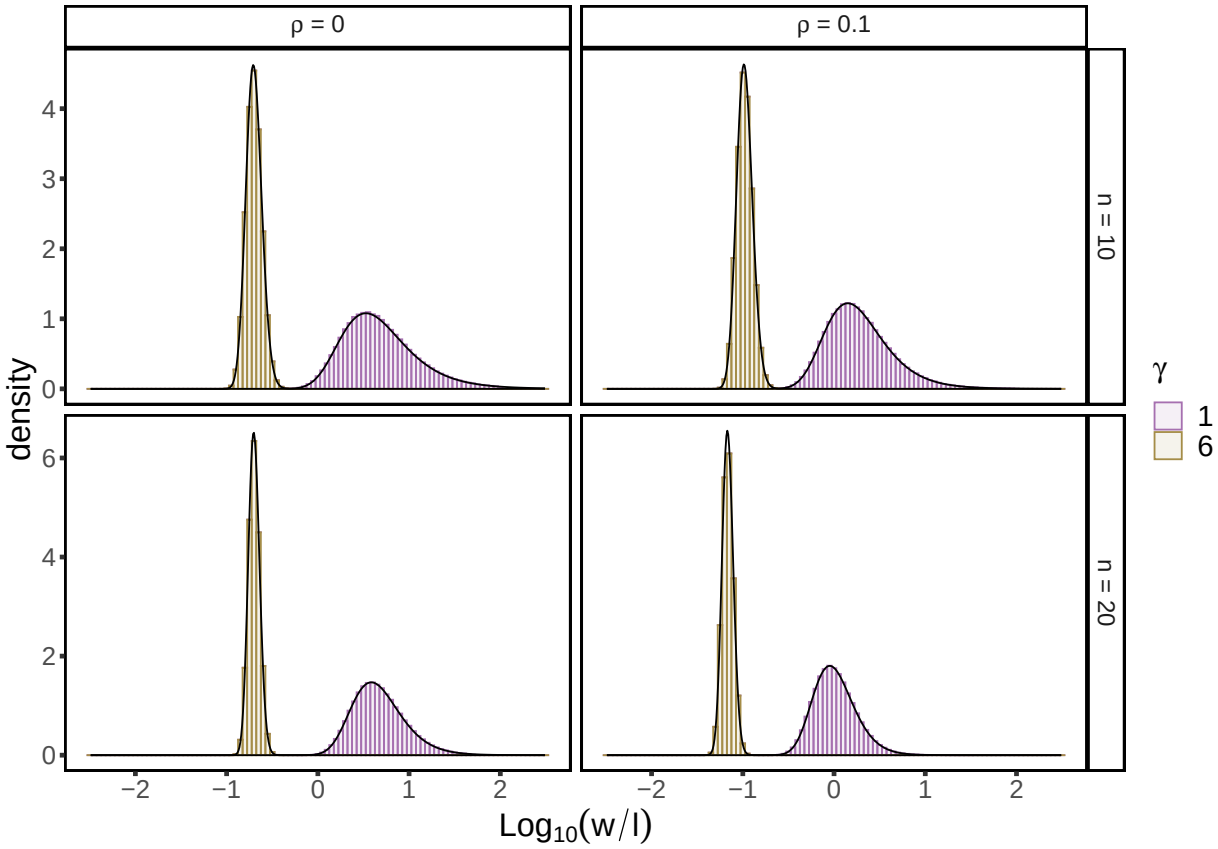

Figure 4.12: **Distribution of the total biomass w of the survival community as a function of the ratio γ of number of traits k to number of species n for different *constant* correlation matrices.** The simulations were done with $n = 10, 20$ species. Histograms are simulations and black lines are the numerical integration of (4.116).

the number of traits.

Going back to re-scaling the interaction matrix by ℓ, total biomass transforms as $w \to \ell w$. By the above calculation and a change of variables $w \to \tilde{w}$, the moments of the distribution of ℓW conditional to m coexisting species are given by

$$\mathbb{E}[(\ell W)^k | m] = \int_0^\infty dw (\ell w)^k g_{\mathrm{a}}(w|m) = \frac{1}{P_{\mathrm{ni}}(m,n)} \int_0^\infty dw g(\nu', w) \times (\ell w^{-1} \mathbf{1}_m^T \Sigma_{11}^{-1} \mathbf{1}_m)^k Q_{n-m}^-(\mathbf{1}_{n-m} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_m, w^{-1}\mathbf{1}_m^T \Sigma_{11}^{-1}\mathbf{1}_m \Sigma_{22.1}). \quad (4.117)$$

By the saddle point calculation done while computing the expected number of survivors we

can approximate the mean of $\ell W|m$ for $\rho \geq 0$, $m = nq$ and $\ell = \gamma n$ as follows: the above integral satisfies (4.86) up to a multiplication by $\frac{\gamma}{w^2 q}\mathbf{1}_m^T \Sigma_{11}^{-1}\mathbf{1}_m$ —observe the rescaling in (4.83). Hence the exponent in the integral does not change so, at the solution (y_0, w_0) of (4.89), neglecting all but the leading order terms we can approximate

$$\mathbb{E}[\ell W|m] \approx \frac{\ell}{(1-\rho+\rho m)w_0^2}. \tag{4.118}$$

Assuming that the distribution of survivors is highly peaked at the mode, we can approximate the mean of W by the mean conditional at the mode, which we get from Eq. (4.111):

$$\mathbb{E}[\ell W] \approx \frac{\ell}{(1-\rho+\rho q^\star n)w_0(q^\star)^2}. \tag{4.119}$$

4.5.5 Relative abundances

For an equilibrium attractor $\boldsymbol{x}_m$ with m species, let $\boldsymbol{v} := \boldsymbol{x}_m / \sum_{i=1}^m x_m^i$ be the relative abundance vector. In particular, $v_m = 1 - \sum_{i=1}^{m-1} v_i =: 1 - \tilde{v}$. By section 4.5.3, Eq. (4.34), we know that $\tilde{v}$ follows a multivariate t distribution, so we can write the distribution function for v_m conditional on $\boldsymbol{x}$ being feasible as

$$\begin{aligned}\Pr(v_m < c|\boldsymbol{x}_m > \mathbf{0}_m) &= 1 - \Pr(v_m > c|\boldsymbol{x} > \mathbf{0}_m) \\ &= 1 - \frac{1}{P_{\mathrm{f}}(m)}\int_0^\infty du g(\nu, u)\Pr(\boldsymbol{y}_u > \mathbf{0}_{m-1}, \mathbf{1}_{m-1}^T \boldsymbol{y}_u < 1-c)\end{aligned} \tag{4.120}$$

with $\nu = \ell - m + 2$. The independence of $\tilde{v}$ and invasibility gives us the distribution of v conditional to $\boldsymbol{x}_m$ being an attractor of the system with m out of n survivors. Let $\boldsymbol{z}_{n-m}$ be

defined as in section 4.5.3. Then

$$\Pr(v_m < c|m) = \frac{\Pr(v_m < c, \boldsymbol{x} > \mathbf{0}_m, \boldsymbol{z}_{n-m} < \mathbf{0}_{n-m})}{P_{\mathrm{a}}(m,n)}$$
$$= \frac{\Pr(\boldsymbol{z}_{n-m} < \mathbf{0}_{n-m}|\boldsymbol{x} > \mathbf{0}_m, v_m < c)\Pr(v_m < c|\boldsymbol{x} > \mathbf{0}_m)}{P_{\mathrm{ni}}(m,n)} = \Pr(v_m < c|\boldsymbol{x} > \mathbf{0}_m), \quad (4.121)$$

where we have used the independence of feasibility and invasibility, $P_{\mathrm{a}}(m,n) = P_{\mathrm{f}}(m)P_{\mathrm{ni}}(m,n)$.

In case of a constant correlation $\rho \geq 0$, all species are equivalent so any surviving species i has the same distribution as x_m. Applying the same derivation as for the feasibility case, and using the notation of the saddle point calculation with $m = qn$ (see Eq. (4.90)), we get

$$\Pr(v_m < c|m) = 1 - \frac{i\sqrt{\lambda_q}}{\sqrt{2\pi}P_{\mathrm{f}}(m)} \int_0^\infty du g(\nu,u) u^{-1/2} \int_\Gamma d\zeta e^{\frac{\lambda_q \zeta^2}{2u}}$$
$$\times \Phi\left(\sqrt{\frac{u}{n\lambda_q}} + \zeta\sqrt{\frac{\lambda_q}{nu}}\right)^{m-1} \Phi\left(\sqrt{\frac{u}{n\lambda_q}} - c\sqrt{\frac{nu}{\lambda_q}} + \zeta\sqrt{\frac{\lambda_q}{nu}}\right). \quad (4.122)$$

Letting $\tilde{c} = cn$, the integral above can be approximated by the same saddle point calculation we did for feasibility (section 4.5.3) up to a multiplication factor given by

$$\frac{\Phi\left(\frac{u}{\sqrt{\lambda_q}}(1-\tilde{c}q) + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}{\Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}. \quad (4.123)$$

Thus, for (u,ζ) satisfying the system of equations (4.93) with ζ real, we get an approximation for the distribution function by neglecting all but the leading terms:

$$\Pr(v_m < c|m) = 1 - \frac{\Phi\left(\frac{u}{\sqrt{\lambda_q}}(1-\tilde{c}q) + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}{\Phi\left(\frac{u}{\sqrt{\lambda_q}} + \frac{\zeta}{u}\sqrt{\lambda_q}\right)}. \quad (4.124)$$

This distribution was compared to simulations in Figure 4.5 (left panel).

4.5.6 Invariant Lotka-Volterra operations

In this section we detail the operations that can be performed in a symmetric stable GLV system without changing the subset of coexisting species.

Let $\boldsymbol{r} \in \mathbb{R}^n$ be the vector of growth rates, and $A \in \mathbb{R}^n$ a symmetric and positive definite interaction matrix. Let $\{S\}_m \subset \{1, \ldots, n\}$ be the *unique* subset of m species that form the attractor, with vector of densities $\boldsymbol{x} = (x_i)$. Then $\boldsymbol{x}$ satisfies:

$$\begin{cases} x_i > 0, & i \in \{S\}_m, \\ x_i(A\boldsymbol{x} + \boldsymbol{r})_i = 0, & \text{for all i}, \\ (A\boldsymbol{x} + \boldsymbol{r})_i < 0, & i \notin \{S\}_m. \end{cases} \tag{4.125}$$

Then we can easily see the effect of the following operations on A and $\boldsymbol{r}$ on the attractor $\boldsymbol{x}$. Let $\kappa > 0$ and D a positive diagonal matrix. The operations that maintain the identity of the species in the endpoint are:

(a) $A \to \kappa A$: then $\boldsymbol{x} \to \kappa^{-1}\boldsymbol{x}$.

(b) $\boldsymbol{r} \to \kappa \boldsymbol{r}$: Then $\boldsymbol{x} \to \kappa \boldsymbol{x}$.

(c) $A \to DAD, \boldsymbol{r} \to D\boldsymbol{r}$: Then $\boldsymbol{x} \to D^{-1}\boldsymbol{x}$.

After any of these operations, the set of coexisting species remains *unchanged*.

Additionally, in the case of $\boldsymbol{r} = \kappa \mathbf{1}_n$, for $\kappa > 0$, we can perform an additional operation:

$$A \to B = A + \mu \mathbf{1}_n \mathbf{1}_n^T. \tag{4.126}$$

Then shifting

$$\boldsymbol{x} \to \boldsymbol{y} = \frac{\kappa \boldsymbol{x}}{1 + \mu \mathbf{1}_n^T \boldsymbol{x}}, \tag{4.127}$$

by direct computation of conditions (4.125) we see that $\boldsymbol{y}$ is a non-invasible equilibrium. If

we additionally restrict $\mu > 0$, $\boldsymbol{y}$ satisfies the feasibility property and B is positive definite so again the support $\{S\}_m$ of the attractor is unchanged.

4.5.7 Varying growth rates

In this section we analyze the effect that growth rates are not equal for all species. By continuity, we expect our results to hold when $\boldsymbol{r} = \mathbf{1}_n + \boldsymbol{\epsilon}_n$ and $\|\boldsymbol{\epsilon}_n\| \ll 1$ if $\ell \geq n$. In case $\ell < n$, the matrix A is singular and the solutions of the system can be unbounded. To correct for that, assume that $A = A + \mu \mathbf{1}_n \mathbf{1}_n^T$ where μ is a sufficiently large enough perturbation so that $A_{ij} + \mu > 0$. In this case $-(A + \mu \mathbf{1}_n \mathbf{1}_n^T)$ is negative semidefinite and dissipative [58], so the solutions are always bounded. Still, the solutions can be degenerate in the sense that there is a hyperplane of non-invasible equilibria towards which the system converges. By perturbing the growth rates we can correct for that. Assume now that $\boldsymbol{r} = \mathbf{1}_n + \mathcal{N}(0, \sigma^2)$, where $\sigma \ll 1$ and that $\hat{\boldsymbol{x}}$ is a saturated rest point of the system (which exists because $A_{ij} + \mu > 0$). Without lost of generality, we can assume that the first m species survive. Then, we have

$$A\hat{\boldsymbol{x}} + \boldsymbol{r} = \begin{pmatrix} \mathbf{0}_m \\ \boldsymbol{z} \end{pmatrix}. \tag{4.128}$$

For $\boldsymbol{z} \in \mathbb{R}_-^{n-m}$, if any $z_i = 0$, then for the system considering only the species $\{1, \ldots, m\} \cup \{i\}$ we have that the restriction of $\boldsymbol{r}$ to this subsystem is contained on a plane of dimension $m < m+1$. Since the distribution of $\boldsymbol{r}$ is continuous, the probability of this event is 0 almost surely. Hence $z_i < 0$ for any i so that invasibility is *strict*. Furthermore, the same argument shows that the rank of A restricted to the survivor subset must be m, i.e. the restriction of matrix A to the set of coexisting species is *full rank*.

Apply the usual Lyapunov function for the system [58],

$$V(\boldsymbol{x}) = -\sum_{i=1}^{n} (\hat{x}_i \log x_i - x_i). \tag{4.129}$$

Defined for any $\boldsymbol{x} \in \mathbb{R}_+^n$, with a global minimum at $\boldsymbol{x} = \hat{\boldsymbol{x}}$ and radially unbounded, then we have

$$\dot{V}(\boldsymbol{x}) = -\sum_{ij} A_{ij}(x_i - \hat{x}_i)(x_j - \hat{x}_j) + \sum_i (x_i - \hat{x}_i)\Big(r_i + \sum_j a_{ij}\hat{x}_j\Big). \tag{4.130}$$

The first sum is non-negative since the matrix is negative semidefinite, and the second is non-positive and is negative unless $x_i = 0$ for any $i > m$. Given that the restriction of A to the survivors subset is full rank then $\dot{V} = 0$ only at $\hat{\boldsymbol{x}}$, which implies that $\hat{\boldsymbol{x}}$ is globally stable and, in particular, is unique [58].

In these cases, while our previous analyses are not exact because of the perturbations introduced in the vector of rates $\boldsymbol{r}$ and in interaction coefficients ($A \to A + \mu \mathbf{1}_n \mathbf{1}_n^T$), we can apply the same machinery that we have developed to provide approximations. This works because we know that the shift of $A \to A + \mu \mathbf{1}_n \mathbf{1}_n^T$ does not change properties like feasibility or invasibility (see section 4.5.6). What changes is that the rank of A goes up by one (see the observation below). Forgetting about this, we can use the same machinery as in the non-degenerate case: for feasibility this follows because only full rank subsets are considered, and the restriction of a singular Wishart to a block of $m \leq \ell$ subsets is a Wishart matrix. Further, the conditional distribution of blocks used for the derivation of the probability of non-invasibility holds in the non-degenerate case too [16].

Observation. The rank of $B = A + \mu \mathbf{1}_n \mathbf{1}_n^T$ is equal to the rank of A plus one. Indeed, let $\boldsymbol{w} \in \ker B$, then $\boldsymbol{w}^T B \boldsymbol{w} = \boldsymbol{w}^T A \boldsymbol{w} + \mu(\mathbf{1}_n^T \boldsymbol{w})^2 = 0$, hence $\boldsymbol{w} \in \ker A \cap \mathbf{1}_n^\perp$, and similarly any $\boldsymbol{w} \in \ker A \cap \mathbf{1}_n^\perp$ is in the kernel of B, hence $\ker B = \ker(A \cap 1^\perp)$. Unless $\ker A \subset \mathbf{1}_n^\perp$, $\dim(\ker B) = \dim(\ker A) - 1$, so the rank increases by one.

Consider then $A = CC^T$ for $C \in \mathbb{R}^{n \times \ell}$, and let $\{\boldsymbol{C}_i\}$ be the set of columns of matrix C. Then $\ker A$ is simply $U^\perp = \{\boldsymbol{C}_i\}^\perp$. As each column $\boldsymbol{C}_i$ is sampled independently from a continuous distribution then $W = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_\ell, \mathbf{1}_n\}$ is a linearly independent set almost surely, then $\dim W^\perp = n - \ell - 1$. Since $W^\perp = U^\perp \cap \mathbf{1}_n^\perp$, and $\dim U^\perp = n - \ell$ then $U^\perp$

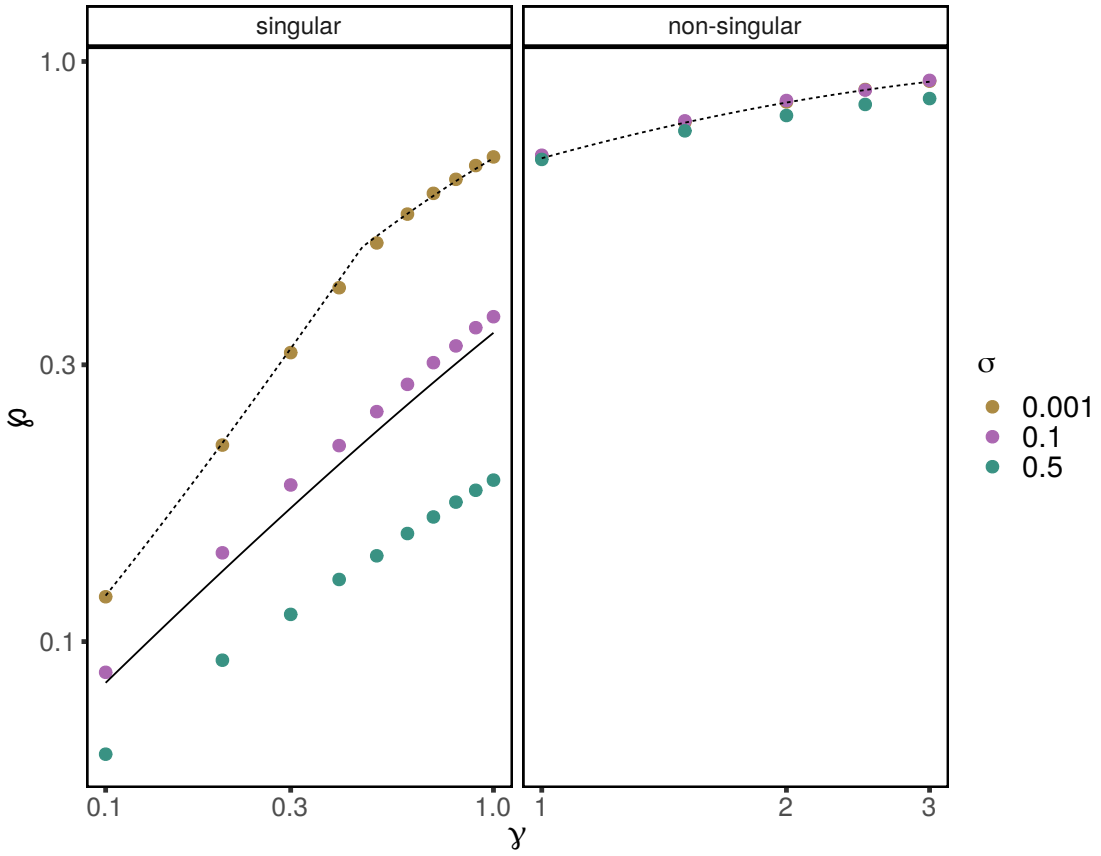

Figure 4.13: **Fraction of survivors under distinct levels of growth rate variability**. Dots mark the average values over simulations with $r \sim \mathcal{N}(1, \sigma^2)$ and $A \sim \mathcal{W}_\ell(I_n, n)$. In the singular case, the matrix A was perturbed by $A \to A + (b + 0.01)\mathbf{1}_n\mathbf{1}_n^T$ for $b = -\min(A)$. Dotted lines represent our analytical predictions assuming $\sigma = 0$. By Section 4.5.6 the shift in A does not affect $\wp$ when $\sigma = 0$. The initial decrease of $\wp$ in the singular case is due to this property not holding when $\sigma \neq 0$. The solid line is our analytical prediction for $\sigma = 0$, when $A \sim \mathcal{W}_\ell(\Sigma, n)$. Σ is a constant correlation matrix with $\rho = \frac{2\sigma_\ell + 0.01}{1 + 2\sigma_\ell + 0.01}$ and $\sigma_\ell = \sqrt{\mathbb{V}(A_{ij})}$ for $i \neq j$ which in this case is simply $\frac{1}{\sqrt{\ell}}$.

cannot be contained in $\mathbf{1}_n^\perp$.

Observe that the restriction on the size of the subsystems set $\gamma + 1/n$ as an upper bound for the mode q^*. In the singular case it may happen that q^* satisfying eq. (4.111) is bigger than $\gamma + 1/n$. Given that we expect the function to be unimodal and increasing with q, then our approximation for the mode in those cases is simply $\gamma + 1/n$.

CHAPTER 5

CONCLUSIONS

In this work, we have studied two types of assembly models: *top-down*, and *bottom-up* (or *sequential*) assembly. Under symmetric or strongly stable Lotka-Volterra dynamics for the local community, we were able to *analytically* compute many properties of the possible outcomes of the assembly process. In particular, we have shown that while the coexistence of an arbitrary subset of N species is highly unlikely, the assembly process can create highly diverse communities from arbitrary regional pools. This result hinges on the fact that the number of possible subsets of species for a given pool of N species is 2^N. As such, even if the probability that any particular subset of species coexist is extremely low, and tends to decline with size, the *combinatorial explosion* in the number of subsets as N grows allows the observation of diverse subcommunities.

5.1 Local Dynamics

Throughout the thesis, we have examined the case of Lotka-Volterra (LV) dynamics, because of the variety of tools available for their analysis [58]. Since the inception of the LV model—exactly a century ago—, countless studies have been devoted to the analysis of its properties [65]. For our purposes, the most useful known result is the explicit characterization of the final communities in terms of *feasibility*, *stability* and *non-invasibility* [58]. In order to analyze the statistical properties of these attractors, we combined their characterization with tools from probability and statistical physics. Extensions to cases where the dynamics do not necessarily converge to a fixed point have already been developed in the work of Bunin and colleagues [99, 100]. However, because these new modelling approaches are based on results from statistical physics, they rely heavily on an assumption of "statistical symmetry" between species. Interactions between species are usually taken to be *i.i.d.*

random variables—while each realization of the interaction matrix shows variability, there is no *true species identity*, given that the properties of the model are invariant to any arbitrary permutation of the species' indexing.

In Chapter 4, we showed that by introducing global correlations stemming from phylogenetic information the statistical symmetry no longer holds, and analytical treatment becomes highly non-trivial. Extending this type of model by explicitly considering the variability between species will allow the calculation of more biologically meaningful quantities such as the probability of survival of a subset of species as a function of their phylogenetic relatedness.

A more interesting approach to extend the LV framework is to introduce additional features that have been documented in natural systems. Two such candidates are the introduction of higher order interactions and phenotypic variability. Higher order interactions (HOI) have been found to stabilize cycling communities [53], and the extension from an LV model is straightforward with the equations being modified as follows:

$$\frac{dx_i}{dt} = x_i(r_i + \sum_j A_{ij}x_j + \sum_{ijk} Q_{ijk}x_j x_k + \ldots) \, . \tag{5.1}$$

While the pairwise interactions can be encoded in the matrix A, the three-way and higher-order interactions need to be encoded in tensors Q. Tensors are the natural generalization of a matrix A, in the same way that a matrix is an extension of a vector [68].

Importantly, in the presence of HOI equilibria are no longer easily computable: we are now searching for the zeros of a system of *non-homogeneous polynomial equations* of degree greater than or equal to two. Furthermore, while the LV model does not admit "true multistability" (in which the same set of species can coexist at different equilibrium points), including higher-order interactions allows the existence of distinct configuration of positive abundances for the same set of species. Recent work showed that the structure of the equilibra of subsystems following LV dynamics accurately reflects the data coming from controlled lab

experiments [85], and it would be interesting to see to what extent the analysis of higher-order interaction models refines these predictions. To this end, we expect that the development of new mathematical tools will be required to understand the structure of the possible equilibria in the model, and to study the effects on the stability and feasibility of the system.

Second, we can include in the model the fact that most species display intraspecific variation [18]. It has already been shown that the introduction of phenotypic variation in LV models positively influences coexistence [84]. A simple way to introduce phenotypic variation is to start with an LV model for N organisms (or phenotypes) and group them into m species, coupling their reproduction with a "redistribution matrix". In other words, we have the following equations (assuming only competitive and predatory interactions), for x_i^j the jth phenotype within species i(cf. [84]).

$$\frac{dx_i^j}{dt} = \sum_k Q_i^{kj} x_i^k (r_i^k + \sum_{A_{il}^{kh} \geq 0} A_{il}^{kh} x_l^h) - x_i^j (d_i^j + \sum_{A_{lk}^{jk} < 0} A_{lk}^{jk} x_l^k) \,, \tag{5.2}$$

where Q_i is the matrix for the ith species which encodes the transition probability Q_i^{kj} from phenotype $k \to j$ at reproduction. The interaction matrix is partitioned into species blocks such that A_{il}^{kh} denotes the interaction between phenotype k of species i and phenotype h of species l.

Importantly, due to the theory of quasi-polynomial systems, both of the models above can be embedded into a (possibly much larger) Lotka-Volterra model [110]. The difference is that the interaction matrix in this case is highly degenerate. To what extent this embedding can aid the analysis of such systems remains to be explored. What is clear, is that the degeneracy of the interaction matrix precludes the application of the classical results on Lotka-Volterra systems. Our only way forward is to devise wholly new ways of looking at the problem.

5.2 Species pool

In this work, we relied heavily on the existence of a constant, external species pool that supplies colonists to the local habitat, in close analogy with a mainland-island model [34]. In more general situations, we expect the existence of feedbacks between the composition of the regional pool and the local community, and two potential ways to correct for this are as follows. First, we can consider the pool to be the total set of species in a metacommunity M, with m sites connected by dispersal [59]: the connection between any two patches is encoded on a dispersion matrix, and each location has its own local dynamics, plus the effects from the immigration/emigration between patches. Importantly, we can then compare the mean size of the communities at the end of the dynamical pruning with our previous results, as was done in Roy et al. [100]. Contrary to a well-mixed system in which fixed points are usually the final states, by considering a metacommunity with enough heterogeneity and low dispersal, the system settles in fluctuating attractors which tend to achieve higher diversity than their fixed-point counterparts.

The second approach, which was suggested at the end of Chapter 4, is to model the development of the pool as the assembly proceeds (c.f. [83]). More precisely, one starts with a single species in the local habitat, and adds new species as "mutants" of the resident species. In this setting, local dynamics happen between the mutation events, and in this way the "pool" of potential colonizers—while potentially infinite—is shaped by the effect it has on the local community at each step of the assembly process.

5.3 Extensions

5.3.1 Low-Rank approximations

Staying within the Lotka-Volterra framework, a straightforward extension of our approach is given by considering models for which A is approximated by a low-rank matrix. Low-

rank approximations to an interaction matrix A simplify the analysis and help us identify the features of the matrix controlling the properties of the community. To do so, we can consider the following models:

$$\frac{dx_i}{dt} = x_i(1 - (Ax)_i) \ , \tag{5.3}$$

with

$$\begin{array}{ll} A = I + u1^t + 1u^t & \text{(additive)} \\ A = I + b11^t + avv^t & \text{(multiplicative)} \end{array} \ , \tag{5.4}$$

for u and v vectors in $\mathbb{R}^n$. To enforce competitive interactions in the additive model we assume $u \geq 0$ and similarly $b > 0$. The additive model has a hierarchical competitive structure. The multiplicative model creates a modular or bipartite structure, depending on the sign of a. As such, even though the model is simpler than what we had considered before, it is complex enough to create relevant interaction structures. It can be shown that, although we do not impose any condition on the stability of the system, the additive model is characterized by a unique, globally attractive fixed point, and the multiplicative model possesses at most two. The simple structure of these models allows for a characterization of the species present in the attractors in such a way that we can devise an efficient algorithm to find them.

5.3.2 Phylogenetic effects on local communities

The ideas contained in Chapter 4 suggest a way to test for the effect of phylogenies on local communities. An ongoing debate is if phylogenetic information actually helps explaining community properties [24]. Most of the current statistical approaches to test for this effect involve the use of summary statistics on the phylogenetic tree as covariates with respect to the total biomass of the community (PD metrics [22, 23]). On the other hand, assuming

a Lotka-Volterra framework, or more generally a linear regression framework as suggested by Maynard et al. [85], we can take into account the whole information encoded in the phylogenetic tree by the following: Let Σ be the variance-covariance matrix of an ultrametric tree T relating the species in the community. Then, consider a Lotka-Volterra model of the form:

$$\frac{dx_i}{dt} = x_i(1 - v_i \sum_j \Sigma_{ij} w_j x_j) \ , \tag{5.5}$$

where $w, v \in \mathbb{R}^n$ are vectors modeling the resource requirements and impacts of the species (c.f. Chapter 3). In this case for any sub-community comprising only species in the subset S, the equilibrium point, x^S, satisfies (in matrix notation):

$$D(v^S)\Sigma^S D(w^S)x^S = 1 \ . \tag{5.6}$$

where Σ^S is the sub-matrix containing the species in the subset S and similarly v^S and w^S are the sub-vectors containing only the entries for the species in S.

Thus, the parameters (Σ, v, w) determine a pattern of abundances $\hat{x}^S$ for any subset of species in the pool. Typical Biodiversity-Ecosystem functioning experiments [114] measure the pattern of abundances in distinct combinations of plant communities out of a pool of species. With these data in hand, one can perform the following: 1) Taking the tree topology as fixed, fit the branch lengths of the tree by minimizing the linear distance or log-distance between the observed x^S and $\hat{x}^S$ predicted by the tree; 2) To test for the effect of phylogeny, compare the fitting obtained for the actual phylogenetic tree with randomizations in which we swap the labeling of the species on the tree (thereby removing the effect of shared history on biomass). If phylogenies where to affect the distribution of abundances we would expect that the fitting using the actual tree topology would outperform most of the randomizations.

5.4 Final Thoughts: a niche for analytical approaches to assembly

"In fact, the dynamics seen in many species systems suggest little to no role for a purely analytical approach. At present, numerical simulations or approaches based on graphs and network theory would appear to be our best tools."

— Drake and Paul [39]

The early ideas of succession, developed by Henry Cowles [35], highlighted the dynamic nature of communities [87]. Building upon Cowles' ideas, researchers developed theories of community assembly. This research program aimed to find the regularities and mechanisms that control the assembly process [38, 69, 120]. Given that natural communities are the product of assembly, understanding the mechanics of this process sheds new light on the puzzle of coexistence [13].

In stark contrast with the vision of succession as an orderly process [92], the many studies on community assembly have highlighted the highly complex nature of the problem [39], with the result that a coherent theory of community assembly is still lacking. The advent of cheap computing power lead many to believe that simulations would be the best tool for solving this problem. Yet, without analytical tools and theory, it is difficult to determine what should be sought after in the rich output of complex simulations, and how simulations relate to experimental data. Without a guiding principle provided by good, solid theory, attempting to relate models and data could increase confusion, rather than clarifying the phenomenon of assembly.

The approach we took here was to study models that are simple enough so that analytical approaches are available, and yet complex enough to speak to the processes happening in nature. It is our hope that this study will show that making analytical progress in assembly, albeit difficult, is possible. We conclude by noting that, quite paradoxically, it seems that

our best tools to study the dynamic nature of assembly is to devise a framework in which the dynamics are "gone", i.e., dynamics are considered only indirectly as in the assembly graph.

REFERENCES

[1] P. B. Adler, J. HilleRisLambers, and J. M. Levine. A niche for neutrality. *Ecology Letters*, 10(2):95–104, 2007.

[2] S. Allesina and J. M. Levine. A competitive network theory of species diversity. *Proceedings of the National Academy of Sciences*, 108(14):5638–5642, 2011.

[3] S. Allesina and S. Tang. Stability criteria for complex ecosystems. *Nature*, 483:205–208, 2012.

[4] S. Allesina and S. Tang. The stability–complexity relationship at age 40: a random matrix perspective. *Population Ecology*, 57(1):63–75, 2015.

[5] S. Allesina, J. Grilli, G. Barabás, S. Tang, J. Aljadeff, and A. Maritan. Predicting the stability of large structured food webs. *Nature communications*, 6(1):1–6, 2015.

[6] D. R. Amor, C. Ratzke, and J. Gore. Transient invaders can induce shifts between alternative stable states of microbial communities. *bioRxiv*, 2019. doi: 10.1101/659052.

[7] G. Barabás, M. J. Michalska-Smith, and S. Allesina. The effect of intra-and interspecific competition on coexistence in multispecies communities. *The American Naturalist*, 188 (1):E1–E12, 2016.

[8] G. Barabás, M. J. Michalska-Smith, and S. Allesina. Self-regulation and the stability of large ecological networks. *Nature ecology & evolution*, 1(12):1870–1875, 2017.

[9] G. Barabás, R. D'Andrea, and S. M. Stump. Chesson's coexistence theory. *Ecological Monographs*, 88(3):277–303, 2018.

[10] M. Barbier, J.-F. Arnoldi, G. Bunin, and M. Loreau. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 115 (9):2156–2161, 2018.

[11] J. Bascompte, P. Jordano, C. J. Melián, and J. M. Olesen. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387, 2003.

[12] U. Bastolla, M. Lässig, S. C. Manrubia, and A. Valleriani. Biodiversity in model ecosystems, ii: species assembly and food web structure. *Journal of theoretical biology*, 235(4):531–539, 2005.

[13] L. R. Belyea and J. Lancaster. Assembly rules within a contingent ecology. *Oikos*, pages 402–416, 1999.

[14] G. Biroli, G. Bunin, and C. Cammarota. Marginally stable equilibria in critical ecosystems. *New Journal of Physics*, 20(8):083051, 2018.

[15] L. S. Bittleston, M. Gralka, G. E. Leventhal, I. Mizrahi, and O. X. Cordero. Context-dependent dynamics lead to the assembly of functionally distinct pitcher-plant microbiomes. *bioRxiv*, 2019. doi: 10.1101/727701.

[16] T. Bodnar and Y. Okhrin. Properties of the singular, inverse and generalized inverse partitioned wishart distributions. *Journal of Multivariate Analysis*, 99(10):2389–2405, 2008.

[17] T. Bodnar and Y. Okhrin. On the product of inverse wishart and normal distributions with applications to discriminant analysis and portfolio theory. *Scandinavian Journal of Statistics*, 38(2):311–331, 2011.

[18] D. I. Bolnick, P. Amarasekare, M. S. Araújo, R. Bürger, J. M. Levine, M. Novak, V. H. Rudolf, S. J. Schreiber, M. C. Urban, and D. A. Vasseur. Why intraspecific trait variation matters in community ecology. *Trends in ecology & evolution*, 26(4):183–192, 2011.

[19] H. C. Bravo, S. Wright, K. Eng, S. Keles, and G. Wahba. Estimating tree-structured covariance matrices via mixed-integer programming. In *Artificial Intelligence and Statistics*, pages 41–48, 2009.

[20] U. Brose, R. J. Williams, and N. D. Martinez. Allometric scaling enhances stability in complex food webs. *Ecology Letters*, 9(11):1228–1236, 2006.

[21] G. Bunin. Ecological communities with lotka-volterra dynamics. *Physical Review E*, 95(4):042414, 2017.

[22] M. W. Cadotte and T. J. Davies. *Phylogenies in ecology: a guide to concepts and methods*. Princeton University Press, 2016.

[23] M. W. Cadotte, T. Jonathan Davies, J. Regetz, S. W. Kembel, E. Cleland, and T. H. Oakley. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology letters*, 13(1):96–105, 2010.

[24] M. W. Cadotte, T. J. Davies, and P. R. Peres-Neto. Why phylogenies do not always predict ecological differences. *Ecological Monographs*, 87(4):535–551, 2017.

[25] J. A. Capitán, J. A. Cuesta, and J. Bascompte. Statistical mechanics of ecosystem assembly. *Physical Review Letters*, 103(16):168101, 2009.

[26] J. A. Capitn, S. Cuenda, and D. Alonso. How similar can co-occurring species be in the presence of competition and ecological drift? *Journal of The Royal Society Interface*, 12(110):20150604, 2015. doi: 10.1098/rsif.2015.0604.

[27] T. J. Case and R. G. Casten. Global stability and multiple domains of attraction in ecological systems. *The American Naturalist*, 113(5):705–714, 1979.

[28] B. Charlesworth and D. Charlesworth. *Elements of Evolutionary Genetics*. Roberts and Company Publishers, 2010. ISBN 9780981519425.

[29] P. Chesson. Mechanisms of maintenance of species diversity. *Annual review of Ecology and Systematics*, 31(1):343–366, 2000.

[30] P. Chesson. Quantifying and testing species coexistence mechanisms. *Unity in diversity: reflections on ecology after the legacy of Ramon Margalef*, pages 119–164, 2008.

[31] J. E. Cohen and D. W. Stephens. *Food webs and niche space*. Princeton University Press, 1978.

[32] J. E. Cohen, F. Briand, and C. M. Newman. *Community food webs: data and theory*, volume 20. Springer Science & Business Media, 1990.

[33] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. Computer science. MIT Press, 2009. ISBN 9780262533058.

[34] H. V. Cornell and S. P. Harrison. What are species pools and when are they important? *Annual Review of Ecology, Evolution, and Systematics*, 45:45–67, 2014.

[35] H. C. Cowles. The ecological relations of the vegetation on the sand dunes of lake michigan. part i.-geographical relations of the dune floras. *Botanical gazette*, 27(2): 95–117, 1899.

[36] O. Diekmann. A beginners guide to adaptive dynamics. *Summer School on Mathematical Biology*, pages 63–100, 2002.

[37] J. A. Drake. The mechanics of community assembly and succession. *Journal of Theoretical Biology*, 147(2):213–233, 1990.

[38] J. A. Drake. Community-assembly mechanics and the structure of an experimental species ensemble. *The American Naturalist*, 137(1):1–26, 1991.

[39] J. A. Drake and S. Paul. Assembly processes. In *Encyclopedia of Theoretical Ecology*, pages 60–63. University of California Press, 2012.

[40] J. A. Dunne, R. J. Williams, and N. D. Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

[41] J. A. Dunne, C. C. Labandeira, and R. J. Williams. Highly resolved early Eocene food webs show development of modern trophic structure after the end-Cretaceous extinction. *Proceedings of the Royal Society of London B: Biological Sciences*, 281 (1782):20133280, 2014.

[42] C. S. Elton. *The ecology of invasions by animals and plants*. Springer Nature, 2020.

[43] K. Eugenius and B. Amit. *Matrix Diagonal Stability in Systems and Computation*. Birkäuser Boston, 2000.

[44] M. A. Freilich and S. R. Connolly. Phylogenetic community structure when competition and environmental filtering determine abundances. *Global ecology and biogeography*, 24(12):1390–1400, 2015.

[45] T. Fukami. Historical contingency in community assembly: integrating niches, species pools, and priority effects. *Annual Review of Ecology, Evolution, and Systematics*, 46: 1–23, 2015.

[46] T. Fukami and M. Nakajima. Community assembly: alternative stable states or alternative transient states? *Ecology Letters*, 14(10):973–984, 2011.

[47] G. F. Gause. Experimental studies on the struggle for existence: I. mixed population of two species of yeast. *Journal of experimental biology*, 9(4):389–402, 1932.

[48] B. S. Goh. Global stability in many-species systems. *The American Naturalist*, 111 (977):135–143, 1977.

[49] B. S. Goh and L. S. Jennings. Feasibility and stability in randomly assembled Lotka-Volterra models. *Ecological Modelling*, 3(1):63–71, 1977.

[50] J. E. Goldford, N. Lu, D. Bajić, S. Estrela, M. Tikhonov, A. Sanchez-Gorostiaga, D. Segrè, P. Mehta, and A. Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, 2018.

[51] J. Grilli, T. Rogers, and S. Allesina. Modularity and stability in ecological communities. *Nature communications*, 7(1):1–10, 2016.

[52] J. Grilli, M. Adorisio, S. Suweis, G. Barabás, J. R. Banavar, S. Allesina, and A. Maritan. Feasibility and coexistence of large ecological communities. *Nature communications*, 8:14389, 2017.

[53] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[54] L. Hang-Kwang and S. L. Pimm. The assembly of ecological communities: a minimalist approach. *Journal of Animal Ecology*, pages 749–765, 1993.

[55] L. Harmon. *Phylogenetic comparative methods: learning from trees*. Self published under a CC-BY-4.0 license, 2018. URL `https://lukejharmon.github.io/pcm`.

[56] J. Hofbauer. On the occurrence of limit cycles in the Volterra-Lotka equation. *Nonlinear Analysis: Theory, Methods & Applications*, 5(9):1003–1007, 1981.

[57] J. Hofbauer. Saturated equilibria, permanences, and stability for ecological systems. In L. J. Gross, T. G. Hallam, and S. A. Levin, editors, *Mathematical Ecology - Proceedings Of The Autumn Course Research Seminars International Ctr For Theoretical Physics*. World Scientific Publishing Company, 1988.

[58] J. Hofbauer and K. Sigmund. *Evolutionary games and population dynamics*. Cambridge University Press, 1998.

[59] M. Holyoak, M. A. Leibold, and R. D. Holt. *Metacommunities: spatial dynamics and ecological communities*. University of Chicago Press, 2005.

[60] S. P. Hubbell. *The unified neutral theory of biodiversity and biogeography (MPB-32)*, volume 32. Princeton University Press, 2001.

[61] C. Hui, H. O. Minoarivelo, and P. Landi. Modelling coevolution in ecological networks with adaptive dynamics. *Mathematical Methods in the Applied Sciences*, 41(18):8407–8422, 2018. doi: 10.1002/mma.4612.

[62] A. James, J. W. Pitchford, and M. J. Plank. Disentangling nestedness from models of ecological complexity. *Nature*, 487(7406):227–230, 2012.

[63] C. R. Johnson. Positive definite matrices. *The American Mathematical Monthly*, 77 (3):259–264, 1970. ISSN 00029890, 19300972.

[64] S. Kéfi, E. L. Berlow, E. A. Wieters, L. N. Joppa, S. A. Wood, U. Brose, and S. A. Navarrete. Network structure beyond food webs: mapping non-trophic and trophic interactions on chilean rocky shores. *Ecology*, 96(1):291–303, 2015.

[65] S. Kingsland. Alfred J. Lotka and the origins of theoretical population ecology. *Proceedings of the National Academy of Sciences*, 112(31):9493–9495, 2015.

[66] I. Kotsiuba and S. Mazur. On the asymptotic and approximate distributions of the product of an inverse wishart matrix and a gaussian vector. *Theory of Probability and Mathematical Statistics*, 93:103–112, 2016.

[67] N. Kraft, W. Cornwell, C. Webb, and D. Ackerly. Trait evolution, community assembly,

and the phylogenetic structure of ecological communities. *The American Naturalist*, 170(2):271–283, 2007.

[68] S. Lane and G. Birkhoff. *Algebra*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1999. ISBN 9780821816462. URL `https://books.google.com/books?id=L6FENd8GHIUC`.

[69] R. Law and R. D. Morton. Alternative permanent states of ecological communities. *Ecology*, 74(5):1347–1361, 1993. ISSN 00129658, 19399170.

[70] R. Law and R. D. Morton. Permanence and the assembly of ecological communities. *Ecology*, 77(3):762–775, 1996. ISSN 00129658, 19399170.

[71] C. E. Lawson, W. R. Harcombe, R. Hatzenpichler, S. R. Lindemann, F. E. Löffler, M. A. OMalley, H. G. Martín, B. F. Pfleger, L. Raskin, O. S. Venturelli, et al. Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology*, pages 1–17, 2019.

[72] C. E. Lemke and J. T. Howson, Jr. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics*, 12(2):413–423, 1964.

[73] S. A. Levin. Community equilibria and stability, and an extension of the competitive exclusion principle. *The American Naturalist*, 104(939):413–423, 1970.

[74] J. M. Levine, J. Bascompte, P. B. Adler, and S. Allesina. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, 546(7656):56–64, 2017.

[75] A. J. Lotka. *Elements of physical biology*. Williams & Wilkins Company, Baltimore, MD, 1925.

[76] R. Mac Arthur. Species packing, and what competition minimizes. *Proceedings of the National Academy of Sciences*, 64(4):1369–1371, 1969.

[77] R. MacArthur. Fluctuations of animal populations and a measure of community stability. *ecology*, 36(3):533–536, 1955.

[78] R. MacArthur. Species packing and competitive equilibrium for many species. *Theoretical population biology*, 1(1):1–11, 1970.

[79] R. MacArthur and R. Levins. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921):377–385, 1967.

[80] R. Marsland, W. Cui, and P. Mehta. The minimum environmental perturbation principle: A new perspective on niche theory. *bioRxiv*, 2019. doi: 10.1101/531640.

[81] R. M. May. Will a large complex system be stable? *Nature*, 238(5364):413–414, 1972.

[82] M. M. Mayfield and J. M. Levine. Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology letters*, 13(9):1085–1093, 2010.

[83] D. S. Maynard, C. A. Serván, and S. Allesina. Network spandrels reflect ecological assembly. *Ecology letters*, 21(3):324–334, 2018.

[84] D. S. Maynard, C. A. Serván, J. A. Capitán, and S. Allesina. Phenotypic variability promotes diversity and stability in competitive communities. *Ecology letters*, 22(11): 1776–1786, 2019.

[85] D. S. Maynard, Z. R. Miller, and S. Allesina. Predicting coexistence in experimental ecological communities. *Nature ecology & evolution*, 4(1):91–100, 2020.

[86] K. S. McCann. The diversity–stability debate. *Nature*, 405(6783):228–233, 2000.

[87] R. P. McIntosh. The background and some current problems of theoretical ecology. *Synthese*, 43(2):195–255, 1980.

[88] P. Mehta, W. Cui, C.-H. Wang, and R. Marsland III. Constrained optimization as ecological dynamics with applications to random quadratic programming in high dimensions. *Physical Review E*, 99(5), 2019.

[89] K. E. Morrison. From bocce to positivity: some probabilistic linear algebra. *Mathematics Magazine*, 86(2):110–119, 2013.

[90] R. J. Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[91] R. Nichols. Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7):358–364, 2001.

[92] E. P. Odum. The strategy of ecosystem development. In *The Ecological Design and Planning Reader*, pages 203–216. Springer, 2014.

[93] J. M. Olesen, J. Bascompte, Y. L. Dupont, and P. Jordano. The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891–19896, 2007.

[94] S. B. Otto, B. C. Rall, and U. Brose. Allometric degree distributions facilitate food-web stability. *Nature*, 450(7173):1226–1229, 2007.

[95] M. Pascual, J. A. Dunne, et al. *Ecological networks: linking structure to dynamics in food webs*. Oxford University Press, 2006.

[96] S. Pimm. *The Balance of Nature?: Ecological Issues in the Conservation of Species and Communities*. University of Chicago Press, 1991. ISBN 9780226668307.

[97] A. Roberts. The stability of a feasible random ecosystem. *Nature*, 251(5476):607–608, 1974.

[98] R. P. Rohr, S. Saavedra, and J. Bascompte. On the structural stability of mutualistic systems. *Science*, 345(6195):1253497, 2014.

[99] F. Roy, G. Biroli, G. Bunin, and C. Cammarota. Numerical implementation of dynamical mean field theory for disordered systems: Application to the lotka–volterra model of ecosystems. *Journal of Physics A: Mathematical and Theoretical*, 52(48):484001, 2019.

[100] F. Roy, M. Barbier, G. Biroli, and G. Bunin. Complex interactions can create persistent fluctuations in high-diversity ecosystems. *PLoS computational biology*, 16(5):e1007827, 2020.

[101] E. L. Sander, J. T. Wootton, and S. Allesina. What can interaction webs tell us about species roles? *PLoS Computational Biology*, 11(7):e1004330, 2015.

[102] S. J. Schreiber and S. Rittenhouse. From simple rules to cycling in community assembly. *Oikos*, 105(2):349–358, 2004.

[103] C. A. Serván, J. A. Capitán, J. Grilli, K. E. Morrison, and S. Allesina. Coexistence of many species in random ecosystems. *Nature ecology & evolution*, 2(8):1237–1242, 2018.

[104] K. Sigmund. Darwin's "circles of complexity": Assembling ecological communities. *Complexity*, 1(1):40–44, 1995.

[105] C. Song, G. Barabás, and S. Saavedra. On the consequences of the interdependence of stabilizing and equalizing mechanisms. *The American Naturalist*, 194(5):627–639, 2019.

[106] P. P. Staniczenko, J. C. Kopp, and S. Allesina. The ghost of nestedness in ecological networks. *Nature Communications*, 4:1391, 2013.

[107] L. Stone. The Google matrix controls the stability of structured ecological and biological networks. *Nature Communications*, 7, 2016.

[108] D. B. Stouffer and J. Bascompte. Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences*, 108(9):3648–3652, 2011.

[109] G. Sugihara. Graph theory, homology and food webs. In *Proc. Symp. in Applied Mathematics*, pages 83–101. American Mathematical Society, 1984.

[110] G. Szederkenyi, A. Magyar, and K. M. Hangos. *Analysis and control of polynomial dynamic models with biological applications*. Academic Press, 2018.

[111] T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[112] V. Temperton, R. Hobbs, T. Nuttle, and S. Halle. *Assembly Rules and Restoration Ecology: Bridging the Gap Between Theory and Practice*. The Science and Practice of Ecological Restoration Series. Island Press, 2013. ISBN 9781597265904.

[113] D. Tilman. *Resource competition and community structure*. Princeton university press, 1982.

[114] D. Tilman, P. B. Reich, J. Knops, D. Wedin, T. Mielke, and C. Lehman. Diversity and productivity in a long-term grassland experiment. *Science*, 294(5543):843–845, 2001.

[115] Y. L. Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

[116] S. Valverde, J. Piñero, B. Corominas-Murtra, J. Montoya, L. Joppa, and R. Solé. The architecture of mutualistic networks as an evolutionary spandrel. *Nature ecology & evolution*, 2(1):94–99, 2018.

[117] C. Violle, D. R. Nemergut, Z. Pu, and L. Jiang. Phylogenetic limiting similarity and competitive exclusion. *Ecology letters*, 14(8):782–787, 2011.

[118] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118(2972):558–560, 1926.

[119] J. Wakely. *Coalescent Theory: An Introduction*. Macmillan Learning, 2016. ISBN 9780974707754.

[120] P. H. Warren, R. Law, and A. J. Weatherby. Mapping the assembly of protist communities in microcosms. *Ecology*, 84(4):1001–1011, 2003.

[121] C. O. Webb. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, 156(2):145–155, 2000.

[122] C. O. Webb, D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. Phylogenies and community ecology. *Annual review of ecology and systematics*, 33(1):475–505, 2002.

[123] M. G. Weber and A. A. Agrawal. Defense mutualisms enhance plant diversification. *Proceedings of the National Academy of Sciences*, 111(46):16442–16447, 2014. ISSN 0027-8424.

[124] R. J. Williams. Effects of network and dynamical model structure on species persistence in large model food webs. *Theoretical Ecology*, 1(3):141–151, 2008.

[125] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

[126] P. Yodzis. The stability of real ecosystems. *Nature*, 289(5799):674–676, 1981.

[127] P. Yodzis. *Introduction to Theoretical Ecology*. Cambridge: Harper & Row, 1989.

[128] N. Zhao, S. Saavedra, and Y.-Y. Liu. The impact of colonization history on the composition of ecological systems. *bioRxiv*, 2020. doi: 10.1101/2020.02.26.965715.