THE UNIVERSITY OF CHICAGO

COMPUTATIONAL APPROACHES TO CHARACTERIZING THE TATOOSH MIDDLE INTERTIDAL COMMUNITY

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
ELIZABETH SANDER

CHICAGO, ILLINOIS
JUNE 2017

In memory of Dr. Kurt Pickett, who introduced me to the joy of research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Writing a dissertation is a large undertaking, and I recieved help, advice, and support from many people in these past five years. First and foremost, I am grateful to my advisors, Stefano Allesina and J. Timothy Wootton. Tim, your deep ecological thought pushed me to look at my own research more deeply. Thank you for your ideas and for generous access to your data. Stefano, the Allesina Lab feels like home to me, and it's bittersweet to say goodbye. The creativity and and fun that you bring, your openness to all of the opportunities I wanted to pursue... I honestly can't express how thankful I am. I feel so lucky to have had you as an advisor.

I also want to thank my Thesis Committee, Greg Dwyer, Cathy Pfister, and Mei Wang, for support and guidance that improved the ecological and statistical aspects of my work. An additional thank you to Cathy for encouraging me to attend Pfister-Wootton lab meetings in my very first weeks at the university. Empirical and theoretical ecology can so easily stay separated in their own bubbles, and that push to escape my bubble helped ground my work many times throughout the years. Thanks also for the help and feedback from the Allesina, Wootton, and Pfister lab members past and present, and especially my collaborators Matt Michalska-Smith and Gyuri Barabás, for being so giving with your time and ideas.

I went on a lot of adventures as a graduate student, and I'm grateful to the support staff of E&E and the BSD for your help and patience: Bonnie Brown, Connie Homan, Alison Anastasio, Audrey Aronowsky, Ian Miller, Melissa Lindberg, and Diane Hall, and many others. Thank you to the Hinds Fund, the Makah Tribe, and to Tim and Cathy for the opportunity to visit Tatoosh Island. I could never have guessed how much I would learn about ecology from that week of exploration, observation, and ill-fated data collection. Thanks and stacks of octopus emoji to the amazing staff and alumni community of the Recurse Center. What an amazing and welcoming community. My summer at the Recurse Center deepened my knowledge and love of programming, and since then, the alumni community has always

been a place of support and pure intellectual joy for me. And of course, thank you to Mike Tessel from UChicagoGRAD and Sonali Sridhar from the Recurse Center for helping me find an internship and turn it into a job.

My friends who listened to my practice talks, who looked over my writing, and who supported me in numerous other ways: Darcy Ross, Hilary Katz, Peter Smits, and Gyuri Barabás, thank you all. Thanks also to my friends in the circus and gaming communities, with whom I could step into another world for a while.

To my husband, Brian Crucitti, I don't even know how to begin. You have been a steadfast, loving, and supportive partner through every moment of graduate school. You kept me grounded through all of the ups and downs. And my family: Don, Diane, and Dave Sander, you have always been there for me, and my time in graduate school was no different. I love you all so much.

CHAPTER 1

INTRODUCTION

We face a quickly changing world, where critical conservation decisions must be made based on limited ecological data. To make informed decisions, it is vital to understand the dynamics of ecological communities, and the underlying network of interactions that shapes those dynamics. As computing power continues to increase, we may benefit from a variety of sophisticated computational techniques, drawn from across the sciences, and use them to improve our ecological understanding.

In the following studies, I use computational approaches to characterize the structure and dynamics of ecological communities, with a focus on the Tatoosh Island middle intertidal. The Tatoosh Island intertidal is one of the longest-studied systems in ecology; first studied by Robert Paine in the early 1960s, the system has been used to study the influence of predators [191, 266], disturbance [194, 196], and indirect effects [266, 271] in ecological communities. This is an excellent system for the study of network structure and dynamics, both because of the diverse community of organisms, and because of the rich data available, including a network with trophic and nontrophic interactions [221] and a long-term dataset of community composition under control and experimental conditions [273]. I use data from this and other communities, in conjunction with machine learning and other computational methods, to make inferences about the structure and dynamics of ecological communities.

The remaining chapters are organized as follows. For consistency, I say that *I* did the work, but much of this work was performed in collaboration with others, who I note at the beginning of each chapter. In Chapter 2, I consider the problem of identifying which species occupy a similar niche; in particular, I ask if trophic and nontrophic interactions provide similar information about the roles species play in their communities. To address this question, I use the stochastic blockmodel, known in ecology as the group model, to identify groups of species, such that species in a group interact with other groups in a similar way.

I extend the group model, which uses an adjacency matrix to identify groupings, to group species based on a signed directed adjacency matrix. I find that trophic data generally provides most of the structural information needed to identify species roles, except when excluding interactions disconnects species from the network. I also find that the groups identified for the Tatoosh Island network correspond strikingly well to known niches in the community (*e.g.*, predatory snails, foraging birds), highlighting the strength of this approach for identifying relevant niches in the network.

In Chapter 3, I expand further on the group model, using the method to determine if parasites fill structurally distinct roles than free-living predators in ecological networks, by measuring the "imbalance" of the groupings. I then consider two possible reasons for the structural differences: concomitant predation and species degree, using the degree-corrected stochastic blockmodel. I find that parasites are structurally distinct from non-parasites (*i.e.*, they form highly imbalanced groups), and that including concomitant predation makes these groups even more distinct, but that degree information actually makes the groups worse. This highlights the importance of considering parasites when studying community structure, and identifies concomitant predation as a factor that differentiates parasites from free-living species.

In Chapter 4, I use food web data to answer a very different question: what factors predict species importance? To do so, I simulate dynamics and local extinctions for entire ecological communities, to directly measure the effect of species removal on the community. I then use a variety of ecological and network centrality measures and, using a hierarchical approach, identify which measures correlate with the effects of species removal. I find that high degree and trophic level correlate with high importance, but that high closeness centrality correlates with low importance, results which align well with previous work on keystone predators and network robustness.

In Chapter 5, I examine how well popular network inference methods identify the struc-

ture of known ecological networks. Network inference is common in microbial ecology, but since the true structures of these ecological networks is difficult to pinpoint and understudied, it is unclear how reliable the methods are. I use three common methods for inference of microbial and gene expression networks: dynamic Bayesian networks, Lasso regression, and a method based on Pearson correlation coefficients. I apply these methods to long-term presence-absence datasets from Tatoosh Island and France, and compare the inferred networks to the empirical structures. I find that these methods are good at predicting short-term dynamics, but have mixed success in identifying the underlying interaction structure. The results suggest that these methods can be effective, but should be used carefully, especially when applied to presence-absence data. This demonstrates the importance of understanding the limitations of different types of data, and the value in validating new methodologies using well-studied systems.

1.1 Conclusions

The work presented here is connected less by an overarching ecological question than by an approach to solving ecological problems: drawing computational methods from across the sciences, and applying them to find generalities between ecological communitites. I address several different topics related to ecological structure and dynamics, but there are many future avenues of research where these methods could provide insight. As I demonstrate in chapters 2 and 3, the group model is useful for coarse-graining the structure of a network. This property could be used to compare the core structures of different ecological networks. The group model makes it possible to condense the structures down to simpler structures with the same number of groups, allowing us to answer a variety of questions. Do communities of all kinds have the same essential group structure? Are the same structural niches present in terrestrial and aquatic systems, freshwater and marine, tropical and temperate? What about microbial communities? The group model is a promising way to zoom out from the

details of individual species, and identify the structural niches that are found in ecological communities.

In chapter 4, I rely on large-scale community simulations. These dynamical simulations are still computationally challenging, slow, and require many parameterization and modelling decisions. Still, they are uniquely able to test community-wide effects given the model assumptions. In particular, they are well-suited to model ecological effects which shift the community from its original equilibrium point, since these effects are difficult to study analytically. For example, community simulations could be used to model different types and frequencies of disturbance, and how they affect community structure and robustness. Simulations with many species are complex and often fragile, and so should be used sparingly. Although two-species models can sometimes be used, the conclusions do not always extend to larger communities (*e.g.*, [173]). Used with caution, these models are a useful tool for community ecologists.

Finally, in chapter 5, I focus on network inference, using methods which have existed for some time in the computer science and statistical literature [247], but have only recently become popular to apply to biological problems [159]. A robust network inference method would be extremely powerful, and could be used to infer the structure and predict the dynamics of many ecological systems where long-term monitoring data have been gathered. However, we have a long way to go to determine the power and limits of these methods in different situations: different amounts or types of data, resiliency to measurement and process error, and so forth. Further efforts to test these methods against real and simulated community data will help build a fuller picture of how best to apply these approaches in ecological systems.

Overall, the work I have presented speaks to the power of connecting computational methods with carefully collected ecological data. In the technology sector, much has been made of the advent of "big data" [158, 167, 32, 166], and the power of machine learning

algorithms as applied to massive datasets. Big data may be powerful in some situations, but it relies on the fact that the data are easy and convenient to collect. Many systems that are useful and interesting to community ecologists are highly inconvenient; remote mountains [144, 254], wave-battered shores [171, 196], and thick rainforests [58, 53] among them. But I do not think that these powerful algorithms and methods need leave the smaller-scale systems behind. Instead, I would like to advocate for the value of *medium data*: far less attention-grabbing, but used carefully with computational approaches and supported with natural history knowledge, a powerful resource for ecological understanding.

CHAPTER 2

WHAT CAN INTERACTION WEBS TELL US ABOUT SPECIES ROLES?[1]

Abstract

The group model is a useful tool to understand broad-scale patterns of interaction in a network, but it has previously been limited in use to food webs, which contain only predator-prey interactions. Natural populations interact with each other in a variety of ways and, although most published ecological networks only include information about a single interaction type (*e.g.*, feeding, pollination), ecologists are beginning to consider networks which combine multiple interaction types. Here we extend the group model to signed directed networks such as ecological interaction webs. As a specific application of this method, we examine the effects of including or excluding specific interaction types on our understanding of species roles in ecological networks. We consider all three currently available interaction webs, two of which are extended plant-mutualist networks with herbivores and parasitoids added, and one of which is an extended intertidal food web with interactions of all possible sign structures (+/+, -/0, etc.). Species in the extended food web grouped similarly with all interactions, only trophic links, and only nontrophic links. However, removing mutualism or herbivory had a much larger effect in the extended plant-pollinator webs. Species removal even affected groups that were not directly connected to those that were removed, as we found by excluding a small number of parasitoids. These results suggest that including additional species in the network provides far more information than additional interactions for this aspect of network structure. Our methods provide a useful framework for simplifying networks to their essential structure, allowing us to identify generalities in network structure and better

1. This manuscript was originally published under the CCBY license and is reprinted with permission: E.L. Sander, J.T. Wootton, and S. Allesina. 2015. PLoS Computational Biology 11(7):1-22.

understand the roles species play in their communities.

Author Summary

Ecological interactions are highly diverse even when considering a single species: the species might feed on a first, disperse the seeds of a second, and pollinate a third. Here we extend the group model, a method for identifying broad patterns of interaction across a food web, to networks which contain multiple types of interactions. Using this new method, we ask whether the traditional approach of building a network for each type of interaction (food webs for consumption, pollination webs, seed-dispersal webs, host-parasite webs) can be improved by merging all interaction types in a single network. In particular, we test whether combining different interaction types leads to a better definition of the roles species play in ecological communities. We find that, although having more information necessarily leads to better results, the improvement is only incremental if the linked species remain unchanged. However, including a new interaction type that attaches new species to the network substantially improves performance. This method provides insight into possible implications of merging different types of interactions and allows for the study of coarse-grained structure in any signed network, including ecological interaction webs, gene regulation networks, and social networks.

2.1 Introduction

Networks are a useful tool to understand patterns of interactions in an ecological community. As ecologists have collected more and more network data, the size of published networks has grown dramatically, with many networks now containing hundreds of species. To make sense of these increasingly complex data, we need tools to simplify the network down to its essential structure, allowing us to identify general patterns of interaction in the community.

The group model (equivalent to the stochastic block model from the social science literature, [230]) is a useful way to simplify and understand ecological networks. It has previously been common to characterize species in terms of their ecological niches, that is, by the resources or predators of a given species. Species with identical niches were considered "trophic species", and ecological networks were often simplified by combining them [162]. However, this approach is highly sensitive to small changes or errors in the network structure, since a single missing or false interaction can change which species may be combined. The group model [11] models the concept of *ecological equivalence* [155] (distinct from the term as used in neutral theory). Species are considered to be ecologically equivalent if their predators and prey are equivalent, who are equivalent if their predators and prey are equivalent, and so on. In other words, species are grouped together if they eat and are eaten by the same other groups. This recursive definition implies that species which are far from each other in the network may still impact each other's grouping. This reflects the ecological reality of the complex ways in which species in a network influence each others' dynamics, for example, via trophic cascades or apparent competition [225, 252]. Since ecologically equivalent species prey on and are preyed on by the same other groups, species within a group can be thought of as filling the same role in the community, and may be expected to operate in the community in similar ways. The group structure is also able to capture both modular (compartmental) and anti-modular (*i.e.*, trophic levels) aspects of the network, both of which are found in ecological networks. Thus, the group model is a useful way to gain a coarse-grained view of ecological dynamics and the niches that are filled in the community.

A limitation of the group model is the fact that it can only group species based on a single interaction type (usually predator-prey interactions, although it could in principle be applied to any one interaction type). Of course, species in ecological communities interact in diverse ways, and different interaction types operate simultaneously to influence community dynamics [237, 269]. Although ecologists have traditionally built separate networks for each

interaction type, such as food webs (containing only feeding interactions), or plant-pollinator and plant-seed-disperser networks (containing only mutualistic interactions) [69, 235, 31, 219, 188, 189], there is a growing recognition that different interaction types may work in concert to influence communities. Both empirical and simulation studies have demonstrated the complex ways in which mutualisms and antagonisms may interact [237, 33, 197]. For example, recent work has begun to explore the possible effects of including cheaters in mutualistic networks [33, 84, 222], modelling communities with multiple interaction types [97, 19], and combining mutualistic networks and food webs [175, 83, 133].

Here, we extend the group model from unsigned (single interaction type) to signed directed adjacency matrices, allowing ecologists to study the general structure of merged interaction networks. Using this extension of the group model, species in a group tend to interact with other groups *in the same way*. We demonstrate one possible use of this method by considering how including or excluding different interaction types changes our understanding of group structure in three interaction webs (the only three such networks currently available). Despite the growing body of work on potential impacts of merging networks with multiple interaction types, it is unknown whether these merged networks provide new, meaningful information about species roles at the network level. While it is intuitive that more types of interaction data would provide more (or more accurate) information about the roles species play in their communities, it is valuable to study this question directly. Clearly, species groupings will be contingent on the species and interactions that are included in the network. Adding interactions may reinforce, refine, or contradict the previous understanding of species roles (Fig. 2.1). We study how our understanding of species roles changes based on different types using three networks of two types. The Tatoosh mussel bed network is an intertidal food web with additional interaction types included. This network contains feeding (+/-), competitive (-/-), mutualist (+/+), commensal (+/0), and amensal (-/0) interactions. For this network, we compare how species group based on all interaction types, only

trophic interactions, and only nontrophic interactions. The other interaction networks, from Doñana Biological Reserve [169] and Norwood farm [210], are terrestrial networks which include plants, plant mutualists, plant herbivores, and in the Norwood web, parasitoids which parasitize herbivores. These networks are structurally different from the Tatoosh web in that they are almost entirely multipartite; that is, they are composed of "layers" of species which only interact with the layers above and below (*i.e.*, mutualists interact with plants, plants interact with mutualists and herbivores, and herbivores interact with plants and parasitoids). In these networks, only plants are involved in both feeding and mutualistic interactions, so we consider how the grouping of plants is affected by including mutualists, herbivores, or both. For the Norwood web, we also consider the effect of including or excluding parasitoids on plant groupings. Using this framework, we study if and how omitting specific interaction types changes our understanding of network structure and species roles.

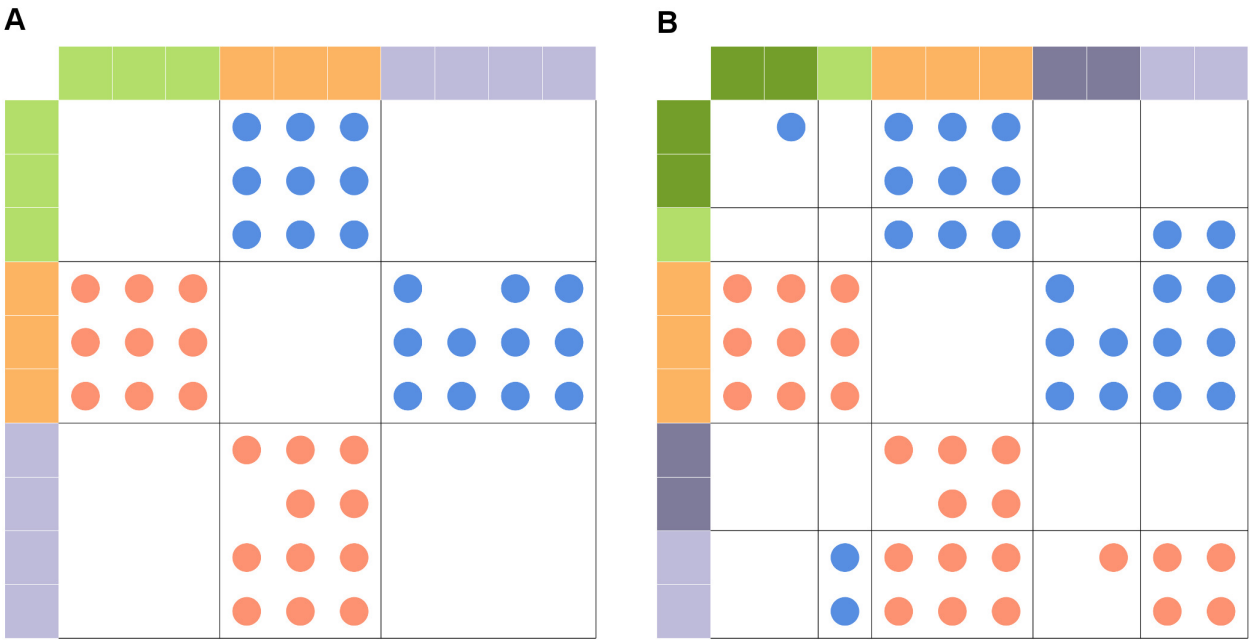

Figure 2.1: **Example 10-species network partitioned using the group model.** Each row and column represents a species, and each dot in the heatmap represents an interaction between two species (red for negative impact of column on row, blue for positive impact of column on row, white for no interaction). Colors on the outer edge correspond to group membership. In (A), only trophic links are included, and the network is partitioned into 3 groups. In (B), both trophic and nontrophic interactions are included. The mutualism between the light purple and light green groups has caused the green and purple groups from part (A) to split into two subgroups. In this example, nontrophic interactions serve to refine trophic groups into subgroups, but additional interactions could potentially reinforce or directly conflict with groupings based on a single interaction type.

2.2 Methods

2.2.1 Ethics Statement

The Makah Tribal Council has granted permission to the Wootton lab for access to Tatoosh Island.

2.2.2 Signed Interaction Networks

A food web composed of S species may be represented by an adjacency matrix A, where A_{ij} is 1 if j consumes i, and 0 otherwise. Similarly, interaction networks may be represented by

a signed adjacency matrix where A_{ij} is 1 if the growth rate of j positively depends on the presence of i, -1 if its growth rate negatively depends on i, and 0 otherwise. Such a matrix may be thought of as containing the signs of the community matrix (the Jacobian evaluated at equilibrium), as opposed to a matrix of zero-sum energy or nutrient flow throughout the system (*sensu* [214]). Some interactions, such as competition for carbon or another nutrient, may be considered an indirect interaction which is the product of two consumer-resource interactions (two direct consumer-carbon interactions in this case). In this example, carbon would be incorporated into the differential equations underlying the community matrix.

Since we are interested in how species group within an interconnected network, we require that the complete interaction networks are a single weakly connected component (that is, isolated subgraphs were removed).

2.2.3 Network Data

Interaction data for Tatoosh Island were collected from the intertidal middle zone based on observed interactions and natural history information. This middle zone on Tatoosh is dominated by the mussel *Mytilus californianus*. This mussel-dominated band is defined from below by the presence of *Pisaster ochraceus*, which consumes *M. californianus* [193], and from above by physiological constraints, such as time spent submerged [196]. The signed interaction network contains 110 taxa and 1898 interaction pairs (869 +/-, 5 +/+, 208 +/0, 492 -/0, and 324 -/-). This dataset is available on Dryad (`DOI:10.5061/dryad.39jv1`)

The largest weakly connected component was taken from Doñana Biological Reserve and Norwood Farm (data made available in [169] and [210], respectively). The Doñana network contains 391 species total, with 170 plants, 207 mutualists (576 mutualistic interactions), and 14 herbivores (221 feeding interactions). The Norwood network contains 445 species, with 91 plants, 251 mutualists (569 mutualistic interactions), 62 herbivores (570 herbivory interactions), and 43 parasitoids (367 parasitic interactions). Two species were classified in

two categories: one which interacted both as a mutualist and as an herbivore, and another as both a mutualist and a parasitoid.

Because taxonomically similar species are generally expected to fill similar roles in a community [263] (but see [87]), taxonomic data provide a potential natural grouping. Tatoosh taxa were classified to kingdom and phylum, and plants in the Doñana and Norwood webs were classified to the order level. Taxonomic levels were chosen to have a number of groups that was similar to the number of groups found by the group model for the complete networks. The high phylogenetic diversity of the Tatoosh system meant that taxonomic groupings beyond the phylum level included too many groups to provide useful information about the system. Taxonomic data for all three networks were gathered from the Integrated Taxonomic Information System (ITIS) database and Encyclopedia of Life (see SI for details).

2.2.4 Group Model for Signed Directed Graphs

Consider an interaction web with S species and L links, K of which are positive and $L-K$ negative. The data can be represented using a signed directed adjacency matrix N. What is the probability of obtaining N by chance? A simple model of random signed network structure is similar to an Erdős-Rényi random graph with S species and a fixed probability c of connecting any two nodes, with an additional probability π that a link is positive. Then the probability of obtaining exactly N using this model is:

$$P\left(N(S,L,K)\middle|c,\pi\right) = c^L \pi^K (1-c)^Z (1-\pi)^{L-K} \tag{2.1}$$

where $Z = S^2 - L$ is the number of zeros in the matrix. This likelihood is maximized when $\hat{c} = \frac{L}{L+Z}$ and $\hat{\pi} = \frac{K}{L}$.

Now to see this in the context of the group model, consider N when divided into two groups, X and Y. If N is a mutualistic web, these groups might correspond to plants and pollinators. Now the random network process involves eight probabilities: c_{xx}, the

probability of a species in group X connecting to another species in group X, π_{xx}, the probability of a link between two species in X being positive, c_{xy}, the probability of a species in X connecting to a species in Y, and so on for c_{yx}, c_{yy}, π_{yy}, π_{xy}, and π_{yx}, which are defined similarly. Note that c_{xy} and c_{yx} are not necessarily equal (nor are π_{xy} and π_{yx}), since N need not be symmetric. Then the probability of obtaining N given the two groups is:

$$P\left(N(S,L,K)\middle|c_{ij},\pi_{ij},i,j\in x,y\right)=\prod_{i\in(X,Y)}\prod_{j\in(X,Y)}c_{ij}^{L_{ij}}\pi_{ij}^{K_{ij}}(1-c_{ij})^{Z_{ij}}(1-\pi_{ij})^{L_{ij}-K_{ij}} \tag{2.2}$$

Analagous to equation 1, this likelihood is maximized when $\hat{c}_{ij}=\frac{L_{ij}}{L_{ij}+Z_{ij}}$ and $\hat{\pi}_{ij}=\frac{K_{ij}}{L_{ij}}$ for all combinations of groups. This can be generalized to g groups as follows:

$$P\left(N(S,L,K)\middle|c_{ij},\pi_{ij},i,j\in 1:g\right)=\prod_{i=1}^{g}\prod_{j=1}^{g}c_{ij}^{L_{ij}}\pi_{ij}^{K_{ij}}(1-c_{ij})^{Z_{ij}}(1-\pi_{ij})^{L_{ij}-K_{ij}} \tag{2.3}$$

When $g=1$, this is equivalent to equation 1. When $g=S$, each species is in its own group, and the likelihood is 1. Such a grouping is not very informative, so we need to perform model selection. Using a uniform prior (such that the probability of each model is $\frac{1}{2}$), it is possible to analytically calculate a Bayes factor to compare two groupings. For groupings G_1 and G_2, the Bayes factor is given by:

$$B=\frac{P\left(N|G_1\right)}{P\left(N|G_2\right)} \tag{2.4}$$

where $P\left(N|G_i\right)$ is the marginal likelihood

$$\int_0^1 \cdots \int_0^1 P\left(c_{ij}, \pi_{ij}, i, j \in 1:S \middle| G_i\right) P\left(N \middle| c_{ij}, \pi_{ij}, i, j \in 1:g, G_i\right) \mathrm{d}c_{11} \ldots \mathrm{d}c_{gg} \mathrm{d}\pi_{11} \ldots \mathrm{d}\pi_{gg} \tag{2.5}$$

which can be analytically integrated to give:

$$\prod_{i=1}^{g} \prod_{j=1}^{g} \frac{K_{ij}! Z_{ij}! (L_{ij} - K_{ij})!}{(1 + L_{ij})(1 + L_{ij} + Z_{ij})!} \tag{2.6}$$

Because there are many possible groupings to choose from, we compared the marginal likelihoods of the groupings when searching for the best grouping, rather than explicitly calculating B for each pair.

We searched for the optimal grouping using Metropolis-Coupled Markov Chain Monte Carlo (MC^3) with a Gibbs sampler (see SI). It is not feasible to exhaustively search the space of all possible groupings, so the best groupings found are not guaranteed to be the optimal ones, but for simplicity, we refer to them as "best groupings" throughout.

Box 1. Calculation of MI for ecological partitions.

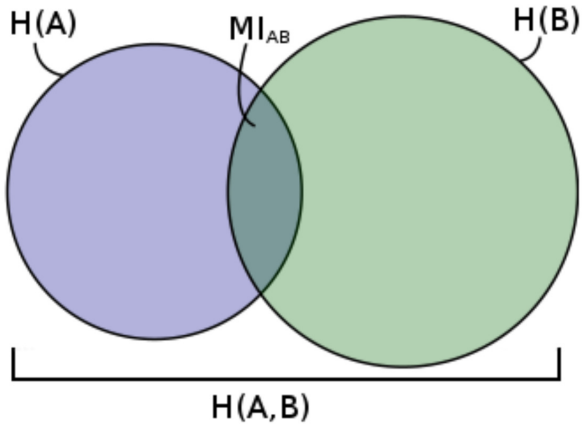

Figure 2.2: **Mutual Information Venn Diagram for 5-species partitions A and B.** Left circle represents $H(A)$, right circle represents $H(B)$, and the intersection represents MI_{AB}. All areas are proportional to the values they represent.

Consider the following two partitions for a five-species grouping:

$$\text{Partition } A\text{: } 1 \quad 2 \quad 1 \quad 2 \quad 1$$

$$\text{Partition } B\text{: } \alpha \quad \beta \quad \gamma \quad \beta \quad \beta$$

where each column is a species, and numbers and Greek letters correspond to group identity within partitions A and B, respectively. Using these groupings, we can create a joint count matrix:

	1	2	$n_{i\cdot}$
α	1	0	1
β	1	2	3
γ	1	0	1
$n_{\cdot j}$	3	2	5

where each table entry n_{ij} is the number of species which are in group i in partition A and in group j in partition B. The row totals $n_{i\cdot}$ and column totals $n_{\cdot j}$ are the marginal counts, *i.e.*, the total number of species in group i in partition A or the total number of species in group j in partition B, respectively. These counts can easily be converted into probabilities by dividing by the total number of species N (in this case, 5). Then $p(a) = \frac{n_{a\cdot}}{N}$, $p(b) = \frac{n_{\cdot b}}{N}$, and $p(a,b) = \frac{n_{ab}}{N}$. This gives us

$$MI_{AB} = \sum_{i=1}^{g_A} \sum_{j=1}^{g_B} \frac{n_{ij}}{N} \ln \left(\frac{n_{ij}}{N} \frac{1}{\frac{n_{i\cdot}}{N}} \frac{1}{\frac{n_{\cdot j}}{N}} \right) \tag{2.7}$$

$$= \sum_{i=1}^{g_A} \sum_{j=1}^{g_B} \frac{n_{ij}}{N} \ln \left(\frac{n_{ij} N}{n_{i\cdot} n_{\cdot j}} \right) \tag{2.8}$$

for our example:

$$MI_{AB} = \frac{1}{5} \ln \left(\frac{1 \cdot 5}{1 \cdot 3} \right) + \cdots + \frac{0}{5} \ln \left(\frac{0 \cdot 5}{1 \cdot 2} \right) \approx .102 \tag{2.9}$$

Because the MI is the shared entropy between two partitions, it can be represented as a Venn Diagram, with circle areas proportional to $H(A)$ and $H(B)$, and the area of overlap between the circles proportional to the mutual information. The corresponding diagram for our example is given in Figure 2.2, with $H(A) = .673$, $H(B) = .950$, and $MI_{AB} = .102$.

2.2.5 Partition Similarity

The entropy of a partition A is an information theoretic measure of the information content or uncertainty of that partition, measured in nats [51]. A partition where all species are in the same group would have low entropy, because we can be quite certain of which group any given species belongs to. In contrast, a partition with many groups would have higher entropy, since it is difficult to make an *a priori* guess about the group identity of a given species. Entropy is calculated as:

$$H(A) = -\sum_{a \in A} p(a) \ln \big(p(a)\big) \tag{2.10}$$

This metric is known as Shannon entropy, commonly used in ecology to measure the diversity of a community [125]. The joint entropy of two partitions A and B is similarly defined:

$$H(A, B) = -\sum_{a \in A} \sum_{b \in B} p(a, b) \ln \big(p(a, b)\big) \tag{2.11}$$

This can be thought of as the union between $H(A)$ and $H(B)$, since it sums over all joint probabilities of the two entropies. Note that for all entropies, $0 \ln(0)$ is given to be 0, so that including values with probability zero does not change the entropy [51].

To measure the similarity between two partitions, we then wish to know how much entropy the partitions share. This is known as the mutual information (MI), which quantifies

Figure 2.3: The Allesina Lab Field Station (also known as Miriam's Cafe at the Smart Museum). All research performed in this text relied heavily on caffeine samples collected daily at this site.

the reduction in entropy of partition B when partition A is known. It is calculated as

$$MI_{AB} = H(A) + H(B) - H(A, B) \tag{2.12}$$

This can be thought of as the intersection between $H(A)$ and $H(B)$. Converting this measure into probabilities gives us

$$MI_{AB} = -\sum_{a \in A} p(a) \ln \big(p(a)\big) - \sum_{b \in B} p(b) \ln \big(p(b)\big) + \sum_{a \in A} \sum_{b \in B} p(a, b) \ln \big(p(a, b)\big) \tag{2.13}$$

$$= -\sum_{a \in A} \sum_{b \in B} p(a, b) \ln \big(p(a)\big) - \sum_{a \in A} \sum_{b \in B} p(a, b) \ln \big(p(b)\big) + \sum_{a \in A} \sum_{b \in B} p(a, b) \ln \big(p(a, b)\big) \tag{2.14}$$

$$= \sum_{a \in A} \sum_{b \in B} p(a, b) \ln \left(\frac{p(a, b)}{p(a)p(b)} \right) \tag{2.15}$$

To see how this is calculated for a partition generated by the group model, see Box 1.

Significance of MI values was estimated based on a randomization test. To estimate how likely it was to get an equal or higher MI by chance, each of the two partitions were shuffled, such that the randomized partitions conserved the number of species in each group (and therefore the upper bound on the MI, see SI for details), but not their identities. The MI was then calculated for the randomized partitions. This process was repeated one million times, and the p-value was estimated as the probability of getting an MI greater than or equal to the observed MI for the two partitions. Since the probability of getting a given MI is based both on the entropies and the groupings, it is possible to get a low p-value for a relatively low MI, or a high p-value for a high MI. Code for calculating partition similarity, obtaining taxonomic data, and running the search algorithm are available on GitHub at

https://github.com/esander91/SignedGroupModel.

2.3 Results

2.3.1 Tatoosh Island

Both the partitions for the network with all interactions and the network with trophic interactions grouped species in a similar way (Fig. 2.4). Though the complete web grouping divided taxa into more groups than the trophic grouping did (19 and 13 groups, respectively), many of these additional groups were simply nested within groups from the trophic one (Fig. 2.5A). Many groups corresponded strikingly well to known ecologically relevant groups in this community, including predatory snails ($n = 4$), kelps ($n = 5$), limpets ($n = 4$), and foraging birds ($n = 3$; Fig. 2.6).

The complete grouping was also quite similar to the nontrophic grouping. In contrast to the trophic partition, which captured the general structure of the complete grouping across the entire web (Fig. 6.1), the nontrophic partition captured portions of the complete one very precisely, but grouped many species into one of two broad groups. Although the nontrophic network contained more interactions than the trophic one overall, these interactions were unevenly spread across species; in particular, sessile species tended to competitively interact with other species, while mobile species often only interacted with a few other species in a nontrophic fashion. As a result, many sessile species (particularly algae and barnacles; see Fig. 6.2 and Table 6.1) were organized into similar groups as in the complete grouping, while most other species were placed into one of two large groups which were sparsely connected to the rest of the network. The trophic and nontrophic groupings were less similar to each other than to the complete grouping (Fig. 2.5C), but were much more similar to each other than expected by chance. Jackknife resampling of the complete network showed that group structure was robust to removal of individual species, as measured by ratio between the

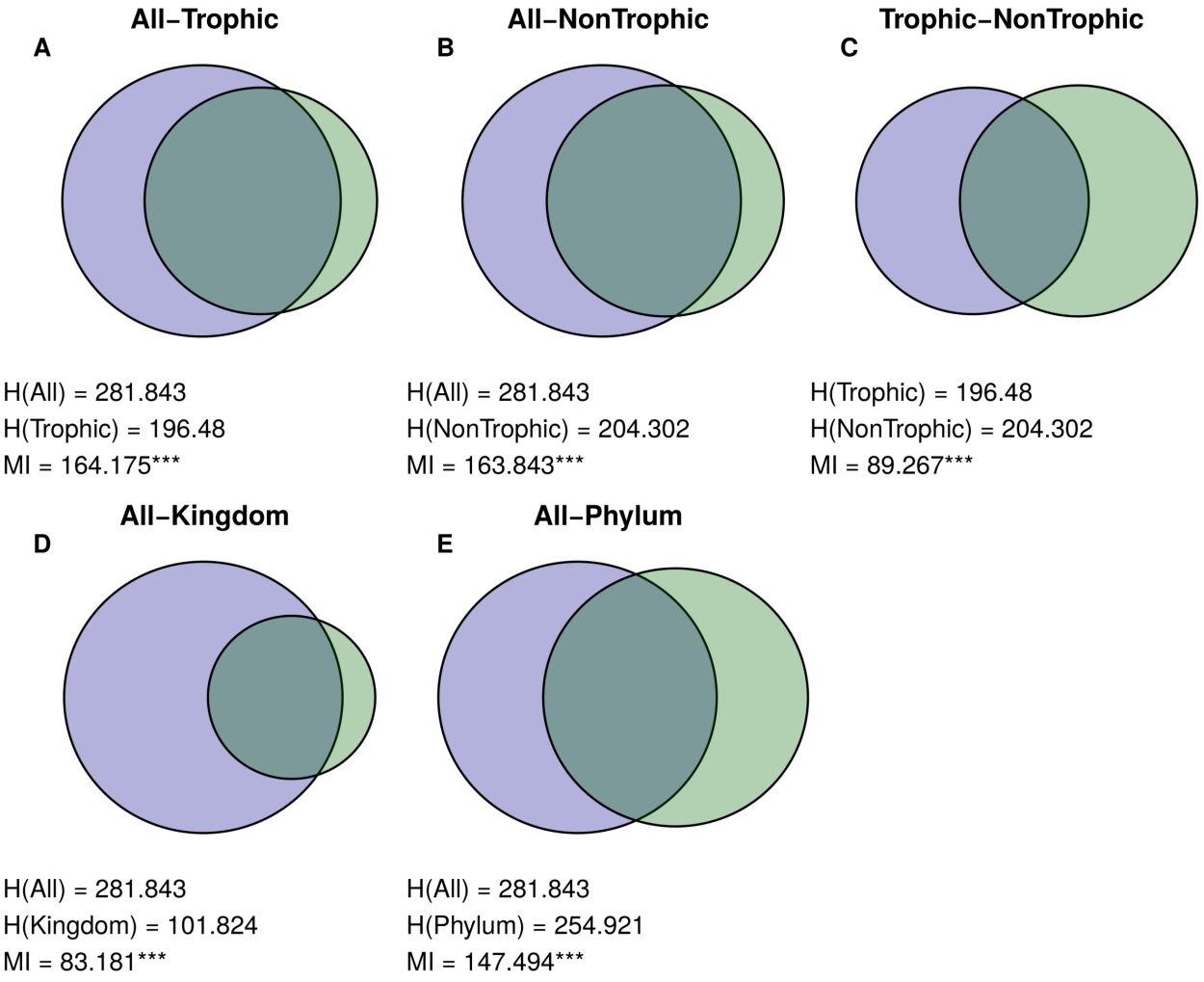

Figure 2.4: **Similarities between Tatoosh Mussel Bed partitions.** Venn Diagrams showing the similarity between pairs of partitions in the Tatoosh Mussel Bed: (A) the complete and trophic networks, (B) complete and nontrophic networks, (C) trophic and nontrophic, and complete and taxonomic groupings (D and E). Venn Diagrams are structured as in Fig. 2.2, where the size of the left circle is proportional in area to the entropy of the first partition listed ($H(A)$), the right circle's area represents the entropy of the second partition listed ($H(B)$), and the overlap between the circles is proportional to the Mutual Information values (MI). Stars next to MI values denote significance level ($^* < .05$, $^{**} < .01$, $^{***} < .001$). Note that this figure includes only the partition comparisons that are discussed in the main text. For all partition comparisons, see Fig. 6.3.

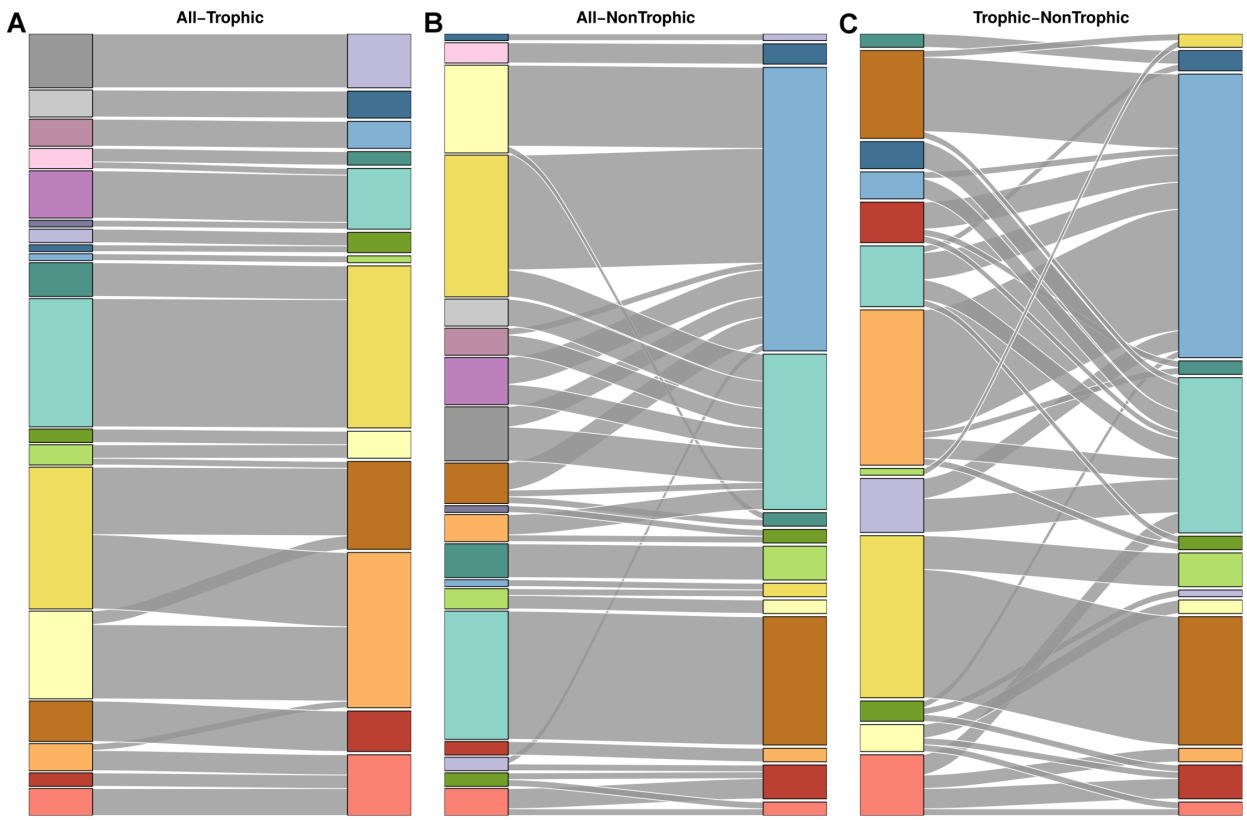

Figure 2.5: **Similarity between Tatoosh network groupings.** Alluvial diagrams comparing the species groupings for (A) complete and trophic webs, (B) complete and nontrophic webs, and (C) trophic and nontrophic webs. Complete network coloring matches colors in Fig. 2.6. Note that the light red group in the complete grouping does not necessarily correspond to the light red group in the trophic group, and so on. Flows between groupings show species in common between two groups; line thickness is proportional to number of species in common. The complete Tatoosh network is organized into groups that are almost perfectly nested in the trophic grouping. The complete grouping also matches closely with the nontrophic groupings, but the trophic and nontrophic groupings are comparatively dissimilar.

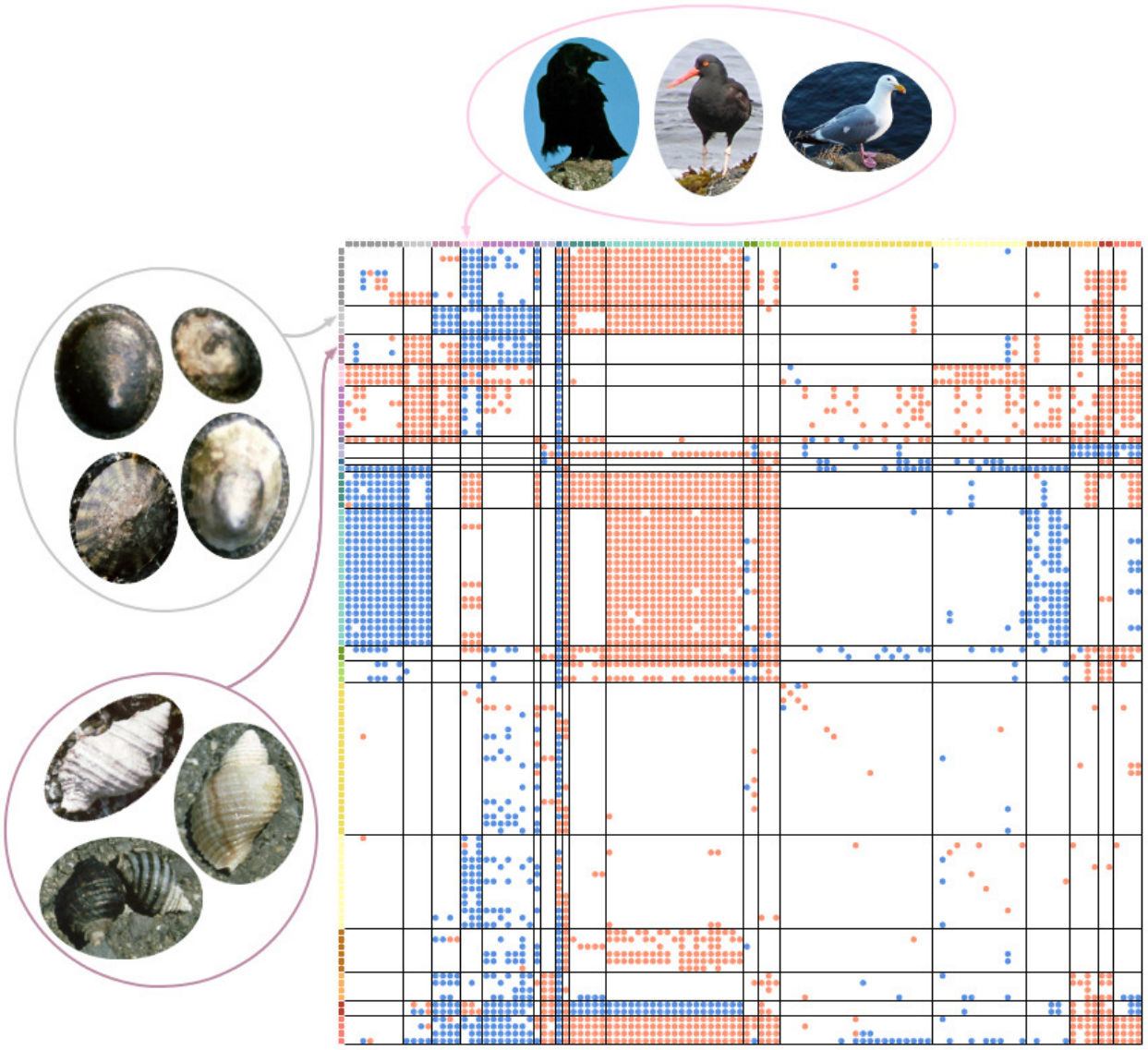

Figure 2.6: **Matrix structure of complete Tatoosh network, organized by groups.** The best complete Tatoosh network grouping, displayed in matrix form. Dot colors in the top row and leftmost column represent group identity (19 groups total). Red and blue dots in the matrix are defined as in Fig. 2.1. Many of the groups in the partition correspond closely to *a priori* ecological knowledge about the system, for example in the foraging birds (dusty purple), limpets (light blue), and predatory snails (dark aqua). This highlights the success of this method in identifying relevant groups, even in the absence of specific ecological information. Full list of species and their group identities given in Table 6.1.

MI for the Jackknifed and original groupings and the maximum MI possible given their entropies (mean $\frac{MI}{MI_{MAX}} = .99, \sigma = .014$; see SI for methodological details).

2.3.2 Doñana Biological Reserve

Plants in the complete Doñana network grouped in a similar way to both the herbivore-removal and mutualist-removal networks. The herbivore-removal and mutualist-removal partitions were much less similar to each other than to the complete partition, although still more similar than expected by chance (Fig. 2.7, Fig. 2.8). The herbivore-removal grouping contained much more information about the complete grouping than the mutualist-removal one did, possibly because mutualists greatly outnumbered herbivores in this network, both in number of species (207 and 14 species, respectively) and interactions with plants (576 and 221 interactions).

2.3.3 Norwood Farm

When parasitoids were excluded from the network, results for the Norwood community were qualitatively similar to Doñana. Mutualist-removal and herbivore-removal groupings were similar to the grouping with both mutualists and herbivores (but not parasitoids), but were less similar to each other (Fig. 2.9, Fig. 2.10). Interestingly, removing herbivores in this network changed group structure more than removing mutualists, even though there were many more mutualists than herbivores (251 and 62 species), and mutualists and herbivores had almost exactly the same number of interactions with plants (569 and 570 interactions).

Including parasitoids in the network markedly changed the resulting group structure. The complete grouping remained similar to the herbivore-removal grouping (which also removes parasitoids, since they only interact with herbivores). However, the mutualist-removal partition was no more similar to the complete one than expected by chance. Surprisingly, the partition for the mutualist-parasitoid-removal was more similar to the complete partition

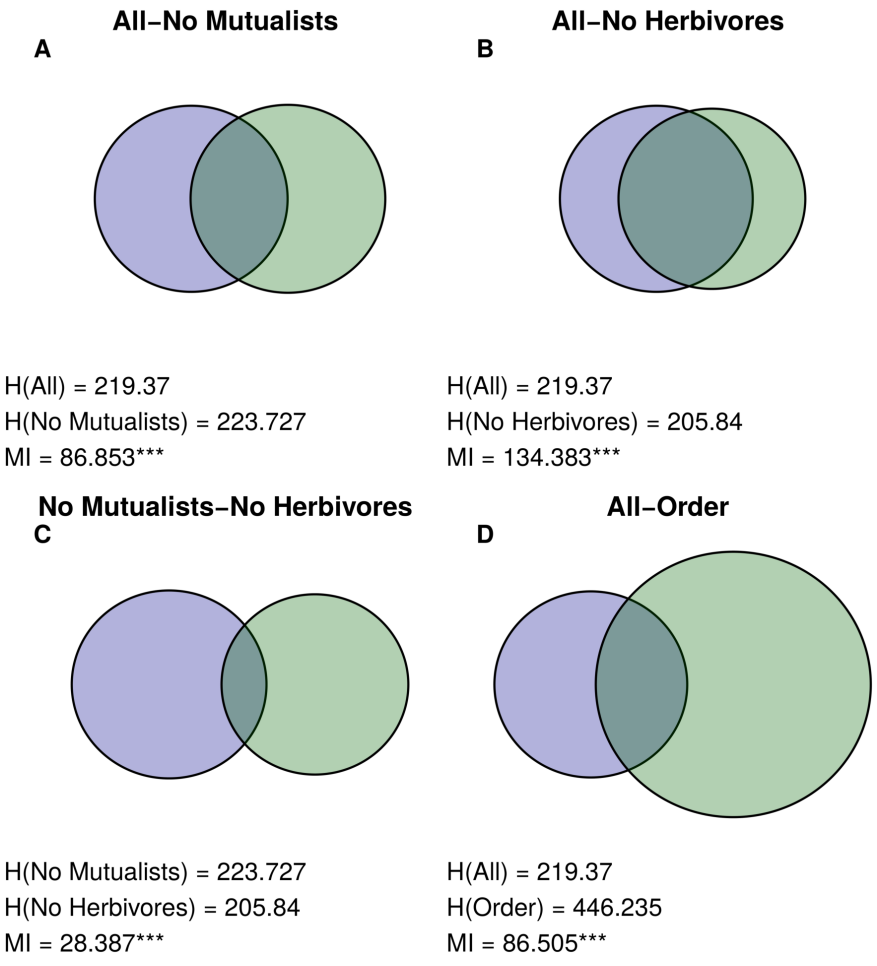

Figure 2.7: **Similarities between Doñana Biological Reserve plant partitions.** Venn Diagrams for similarity between pairs of plant partitions for the Doñana web: (A) complete and mutualist-removal webs, (B) complete and herbivore-removal webs, (C) mutualist-removal and herbivore-removal webs, and (D) complete network and taxonomic order. Figure structured as in Fig. 2.4. This Figure includes only comparisons relevant to the main text; for all comparisons, see Fig. 6.4.

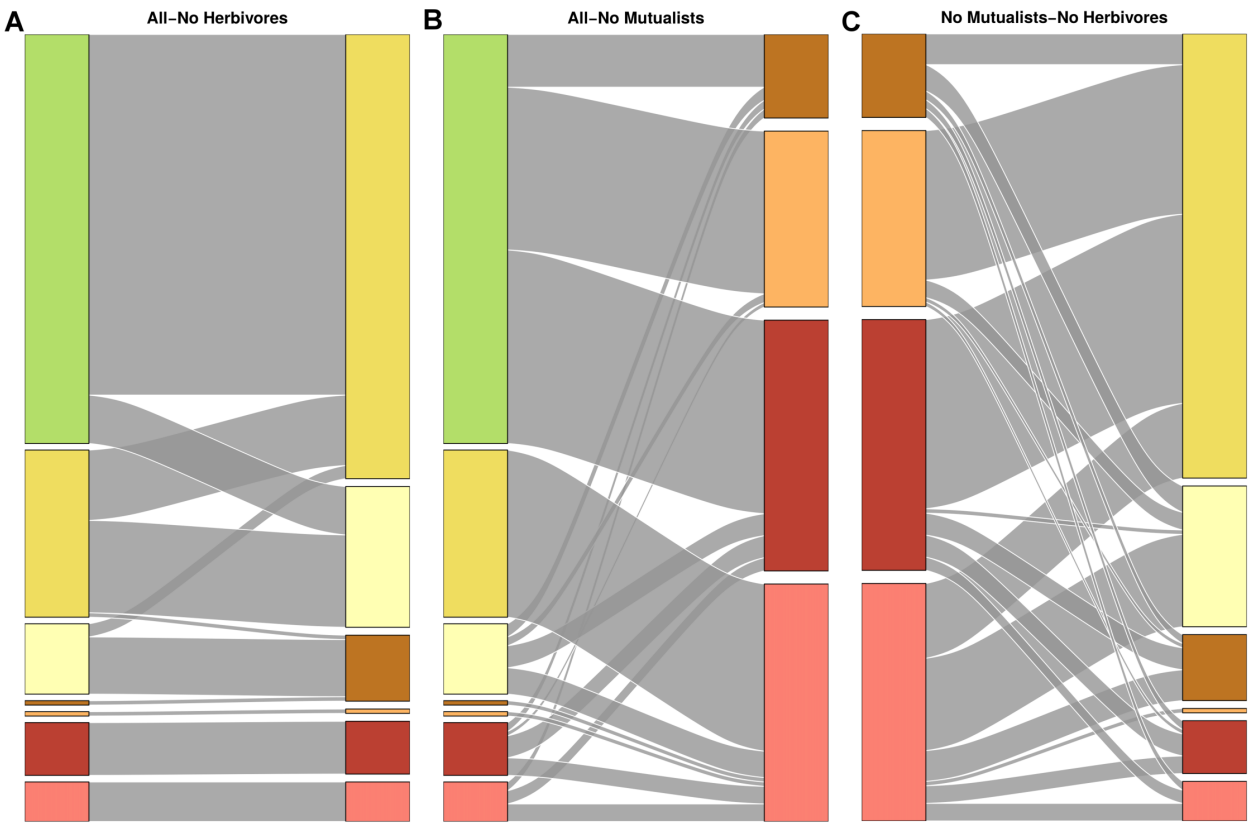

Figure 2.8: **Similarity between Doñana plant groupings.** Alluvial diagrams comparing the plant groupings for (A) complete and herbivore-removal webs, (B) complete and mutualist-removal webs, and (C) herbivore-removal and mutualist-removal webs. All three comparisons show major areas of similarity, but the groupings in (C) have many more conflicts than (A) and (B).

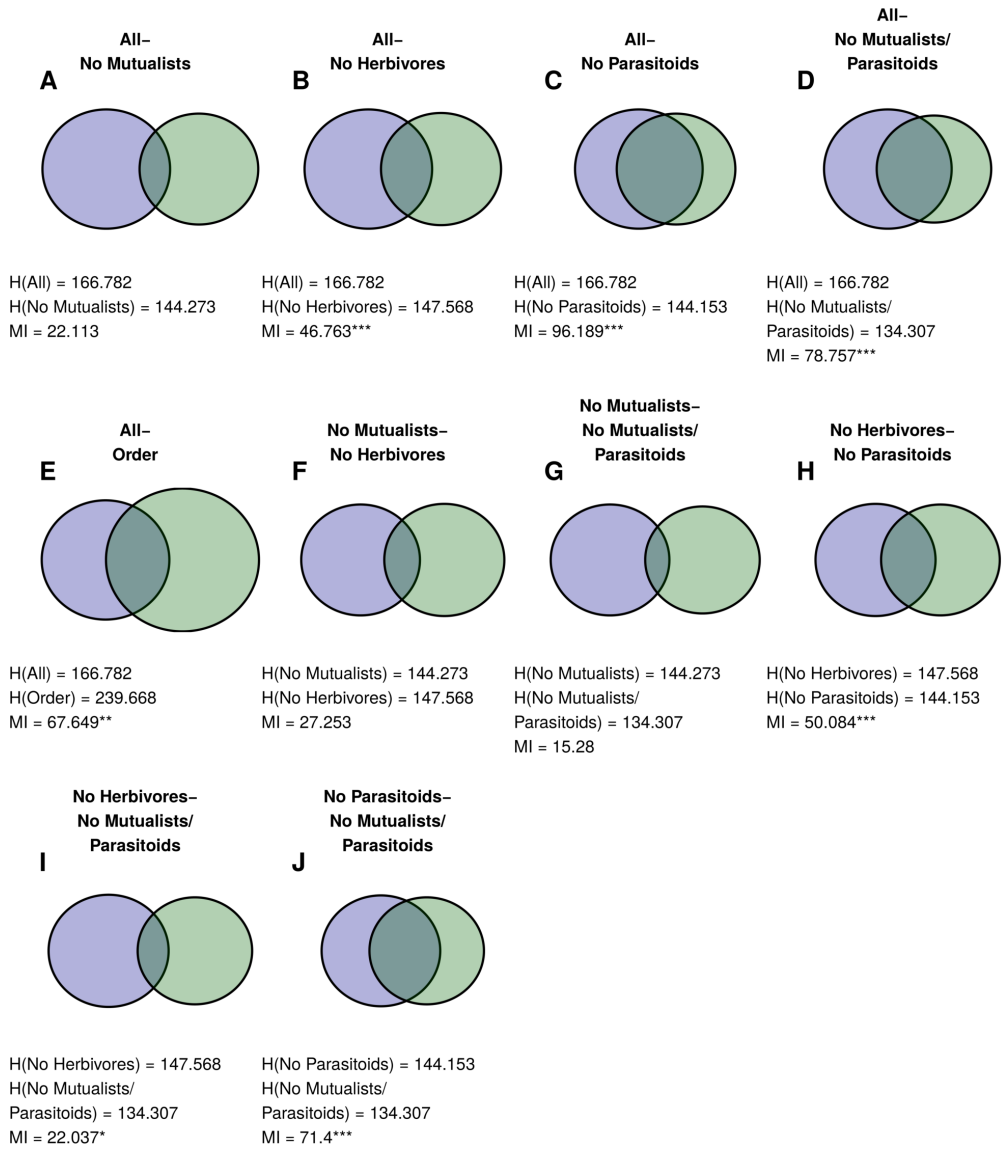

Figure 2.9: **Similarities between Norwood Farm plant partitions.** Venn Diagrams for similarity between pairs of plan partitions for the Norwood Farm webs: (A) complete mutualist-removal webs, (B) complete and herbivore-removal webs, (C) complete and parasitoid-removal webs, (D) complete and mutualist-and-parasitoid-removal webs, (E) complete web and taxonomic order, (F) mutualist-removal and herbivore-removal webs, (G) mutualist-removal and mutualist-and-parasitoid-removal webs, (H) herbivore-removal and parasitoid-removal webs, (I) herbivore-removal and mutualist-and-parasitoid-removal webs, and (J) parasitoid-removal and mutualist-and-parasitoid-removal webs. Figure structured as in Fig. 2.4. Note that comparisons H-J are equivalent to the comparisons in Doñana, in that they show the effect of removing mutualists and herbivores in the absence of parasitoids. As in Figs. 2.4 and 2.7, only partition comparisons relevant to the main text are included; for all comparisons, see Fig. 6.5.

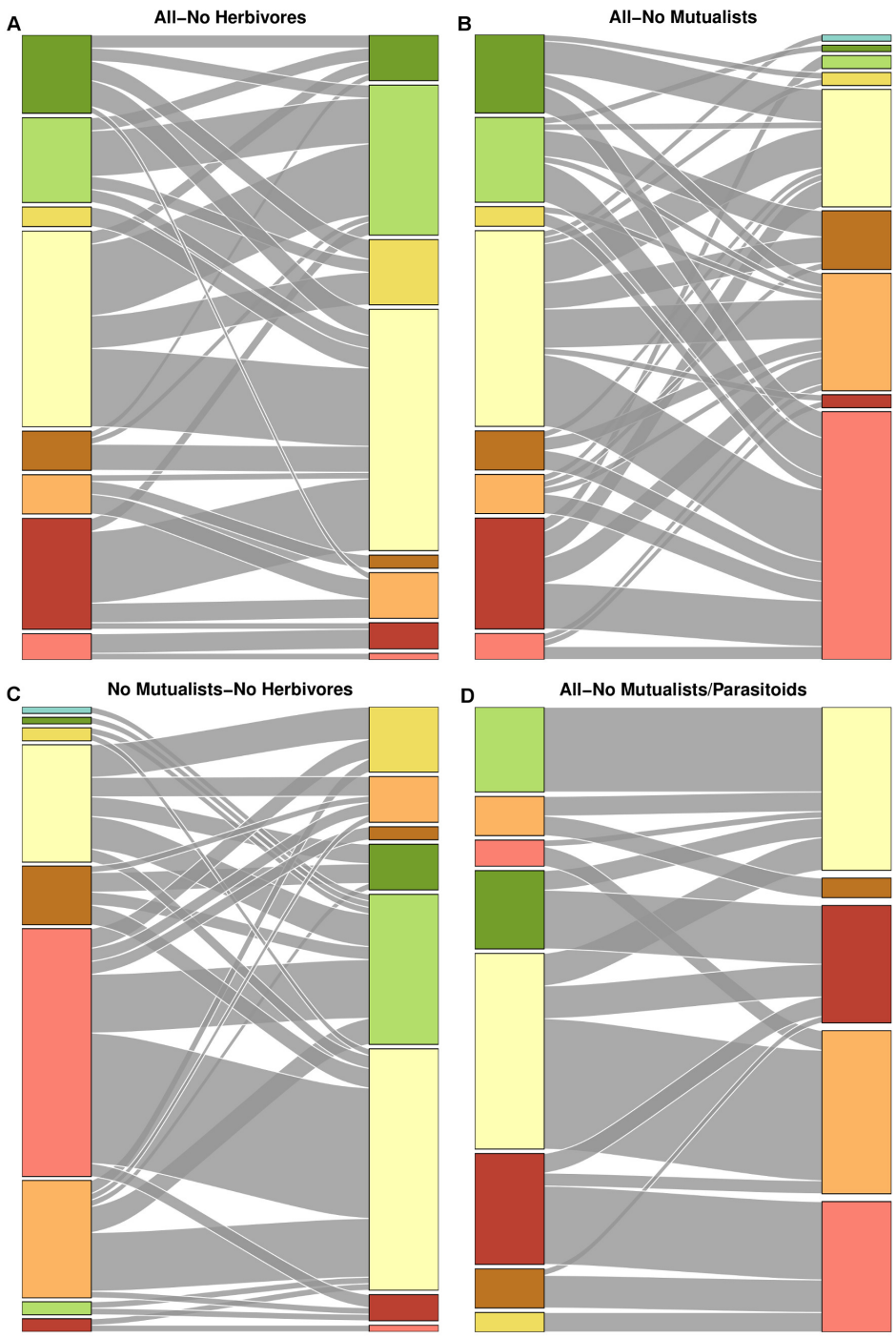

Figure 2.10: **Similarity between Norwood plant groupings.** Alluvial diagrams comparing the plant groupings for (A) complete and herbivore-removal webs, (B) complete and mutualist-removal webs, (C) herbivore-removal and mutualist-removal webs, and (D) complete and mutualist-and-parasitoid-removal webs. In general, these grouping are more dissimilar than seen in the Tatoosh and Doñana systems, and only (A) and (D) show more similarity than expected by chance.

than either the herbivore or mutualist removal groupings.

2.3.4 Taxonomic Groupings

Taxonomic grouping provided some information about complete groupings for all three networks. The Tatoosh complete grouping is almost perfectly nested within the species classification by kingdom (Fig. 2.4, Fig. 2.11). However, because this classification is so broad, it provides less information than phylum, even though the phylum grouping and complete grouping are dissimilar in many areas. In the Doñana and Norwood webs, taxonomic order was significantly similar to the complete groupings (Figures 2.7 and 2.9, respectively), but this similarity was not even across orders: some orders strongly grouped together in the complete groupings, while many others were scattered between several groups (Fig. 2.11).

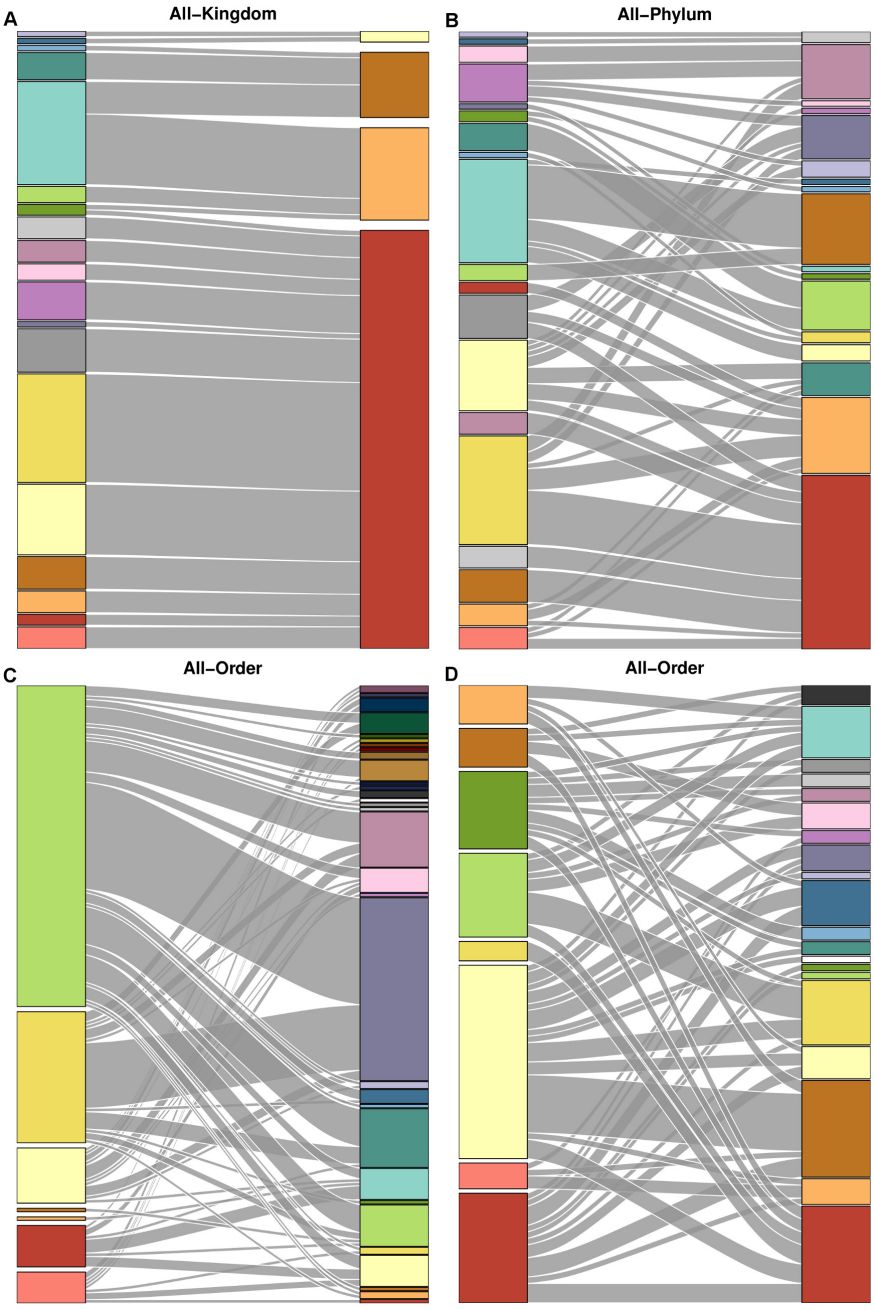

Figure 2.11: **Comparison between complete and taxonomic groupings.** Alluvial diagrams comparing complete web groupings with taxonomic groupings for (A) Tatoosh and kingdom, (B) Tatoosh and phylum, (C) Doñana and plant order, and (D) Norwood and plant order. All groupings are more similar than expected by chance. Kingdom matches very closely with the complete Tatoosh grouping, but has so few categories that it still provides very limited information. The other taxonomic groupings have more categories but still provide relatively little information.

2.4 Discussion

The extended group model is able to take large networks of great complexity, with many types of interactions, and condense them down to their essential structure. This results in a significant decrease in network complexity. It is able to reduce the Tatoosh intertidal network from 110 species down to 19 groups of ecologically equivalent taxa. Using a subset of these interaction types reduces the number of groups simply because the model has less information to work with, and indeed we see that the number of groupings in Tatoosh is greater with all interactions than with trophic interactions only (19 and 13 groups, respectively). Thus, using this extension of the group model in conjunction with interaction web information gives us a slightly more refined view of the network structure. It is notable that the Tatoosh groupings corresponded closely to many ecologically natural sets of species. The model does not use any ecological information outside of the network structure itself, but these patterns of interaction alone are enough to make highly specific distinctions, such as between limpets and other types of grazers.

As one possible use of the extended group model, we consider the effects of including or excluding interaction types from a network. In the Tatoosh network, removing interactions did not exclude species from the network, and even removing large numbers of interactions—nontrophic interactions constitute 54% of interactions in this system—had relatively little effect. This means that in these networks, species which have similar patterns of predation also have similar patterns of competition and mutualism, and so forth. In Doñana and Norwood, however, removing interaction types mean that entire classes of species were also included, and these removals had a comparatively large effect on the group structure. This suggests that plants which are similar to mutualists are not necessarily also similar to herbivores.

The grouping differences between these two network types could arise for many reasons. Sampling effects could play a role, since only three networks were available for study. In-

trinsic differences between terrestrial and intertidal systems might also have an effect, since marine systems exhibit strong trophic control [225, 226]. Because terrestrial mutualists and herbivores are not as tightly linked by these top-down forces, plant groupings based on these different groups might not be tightly linked either. Another possibility relates to the biological traits which underly species interactions. In the intertidal, traits which are relevant to predators, such as mobility and presence of a shell, are likely also relevant for other types of interactions. For example, sessile species will tend to compete for space, and shelled species may benefit other species by providing shelter. In the Tatoosh community, mobile and sessile species rarely group together, and this is also true for shelled and shell-less species (Fig. 2.6, Table 6.1). In terrestrial plants, traits and structures that are relevant to mutualists (flowers, fruits) are relatively distinct from those that are relevant to herbivores (foliage, defense compounds). This specificity of traits relevant to particular interactions could decrease the group similarity when considering different parts of the network.

Taxonomic classification provides an obvious natural grouping for species. However, although taxonomic grouping provided some information about the complete group structure (as has been found for food webs in [75]), they were never the best way to estimate it. Taxonomic groupings were either too broad to provide much information, or grouped species differently than the complete network. This coincides with recent findings that phylogenetic relatedness poorly predicts interaction patterns and species roles in green algae [180, 87, 5].

The recursive definition of the group can lead to interesting outcomes. For example, parasites have a dramatic effect on Norwood group structure in the absence of mutualists. This is likely the result of a domino effect where parasitoids influence the grouping of herbivores, and herbivores influence the grouping of plants. Thus, when mutualists are removed, parasitoids have a major effect on the broad structure of the system. But in the presence of mutualists, plants are being influenced by both mutualists and herbivores, and the signal is lost. This result adds to the abundant evidence for the importance of including parasites in networks

[114, 246, 142, 41] (but see [68]), but more generally, it demonstrates that species need not be directly connected to influence each other. This situation reflects ecological reality, in that species may place evolutionary pressures on each other via a common species, which has been documented specifically between plant pollinators and plant herbivores [3, 237].

The extended group model may help us study and understand interaction networks in a variety of ways. One possible approach is simply to examine the grouping and look for surprises. For example, only crustose and coralline algae form a group separate from other algae based on trophic information in the Tatoosh network, but when nontrophic information is also incorporated, several kelp species form an additional distinct group. This suggests that these two groups interact differently in the network, in a way that specifically relates to their nontrophic interactions. On closer examination of the network structure, this difference is likely related to the fact that these kelps have a negative effect on the growth of the other algal group, but the other algae do not negatively affect the kelps.

Similarly, because the group model identifies ecologically equivalent species, it can be used to identify species which are performing unique roles in the community. In the Tatoosh network, there are three species which are not grouped with any others: detritus, diatoms, and *Anthopleura elegantissima*, a sea anemone. Detritus and diatoms are both relatively unique food sources that are present in the water column, rather than attaching to the rock. It is, perhaps, less obvious why anemones are so unique as to be placed in their own group. However, they are unlike all other species in the system in that they are predatory but sessile, unlike other predators which move to find and consume their prey. *Anthopleura* also has endosymbiotic algae which are implicitly included in the network through the anemone's interactions. Although the existence of a group does not guarantee that it is essential for ecosystem functioning, looking at groups with one or few species may be a useful way to identify species which play unique roles and whose removal might have a larger effect on the system, since no other species are able to take their place.

Another possible application of the group model is to have a simpler version of the network to work with. These simplified networks are easier to take in and comprehend by eye. They may also be useful for finding generalities across networks. This is currently difficult to do, since there are few interaction networks currently available. In the future, it would be interesting to see if communities tend to form similar numbers of groups, if specific species always perform unique roles, if similar groups tend to form at specific trophic levels, and so forth. Since species within a group perform similar roles in the community, we speculate that these species might exhibit similar population dynamics. It is possible that simplifying networks down to their group structure could be a useful way to simplify multi-species dynamical models.

2.4.1 Conclusions

The extended group model is a general method for identifying functionally equivalent nodes in signed directed networks. We have discussed the method as applied to ecological interaction webs, but the methodology could also be used to study the structure of networks of gene regulation [99, 215], sensors [30], and even social networks which incorporate both positive and negative social interactions [148]. The generality of the method does not detract from its usefulness in ecology; in fact, the model is able to identify highly specific ecological roles. This model is a new and useful exploratory tool to understand and compare the coarse-grained structure of ecological communities.

2.5 Acknowledgments

Thanks to E. Baskerville, G. Barabás, M. Michalska-Smith, C. Pfister, M. Wang, and G. Dwyer for their ideas and suggestions. We are grateful to the Makah Tribal Council for providing access to Tatoosh.

CHAPTER 3

UNDERSTANDING THE ROLE OF PARASITES IN FOOD WEBS USING THE GROUP MODEL

3.1 Summary

1. Parasites are ubiquitous and have been shown to influence macroscopic measures of ecological network structure, such as connectance and robustness, as well as local structure, such as subgraph frequencies. Nevertheless, they are often underrepresented in ecological studies due to their small size and often complex life cycles.

2. We consider whether or not parasites play structurally unique roles in ecological networks; that is, can we distinguish parasites from other species using network structure alone?

3. We partition the species in a community statistically using the group model, and we test whether or not parasites tend to cluster in their own groups, using a measure of "imbalance."

4. We find that parasites form highly imbalanced groups, and that concomitant predation, in which a predator consumes a prey and its parasites, but not the number of interactions, improves the group model's ability to distinguish parasites from non-parasites.

5. This work demonstrates that parasites and non-parasites interact in networks in statistically distinct ways, and that these differences are partly, but not entirely, due to the existence of concomitant predation.

3.2 Introduction

Parasites are ecologically significant players in many communities, and several authors have urged the incorporation of these species into ecological networks [161, 160, 141]. While many modern networks are well resolved with respect to most free living species, parasites are often excluded entirely. In networks which do incorporate parasites, these species affect several general aspects of food web structure; for example, increasing the number of species (richness), proportion of possible links that are observed (connectance), and number of consumptive links between the highest and lowest trophic levels (trophic chain length) [114, 246, 142, 17, 68]. Conversely, many parasite species may decrease network robustness—usually quantified as the proportion of species lost following a given number of primary extinctions [70]—because highly complex and specialized life cycles may make them prone to secondary extinction in response to host removal [141].

Parasites also affect local network structure. In a comprehensive analysis on the topic, [68] show that parasites change the relative frequency of certain network subgraphs, and tend to have niches which are broader, but contain more gaps, than predators in aquatic food-webs. The intimate connection between parasites and their hosts results in another major effect. Concomitant predation, wherein a predator consumes both the prey and its parasites, is sometimes a necessary part of parasitic life cycles. These interactions increase the connectance of the network and affect the degree, *i.e.* the number of consumptive interactions involving a given species, of both parasites and their incidental predators.

There is some disagreement in the literature about how exactly parasite degree differs from that of free-living species. Parasites are often highly specialized (*e.g.* [16, 73]), suggesting that parasites might have a lower in-degree (number of prey/hosts) than free-living predators. However, parasites have been found to increase overall connectance, depending on how the calculation is done [142]. Concomitant predation will also affect degree, increasing the in-degree of free-living predators and the out-degree (number of predators)

of parasites. This suggests that degree is a structurally distinguishing feature of parasitic species, especially when in-degree and out-degree are considered separately.

Ecological networks often contain hundreds of species and thousands of consumer-resource interactions. To study these complex networks, it is useful to understand the general roles species play in the community. Species roles are sometimes classified based on phylogenetic (*e.g.* a terrestrial ungulate can be assumed to be herbivorous) or *a priori* trophic strategies (e.g. an herbivore consumes exclusively primary producers by definition), but they may also be identified statistically. *Ecologically equivalent* species (also known as trophic species) have the same set of predators and prey, and therefore play identical roles in the network structure [155, 69]. This concept can be relaxed and generalized using the group model [11], which organizes species into groups, such that species in a group tend to eat and be eaten by the same other groups of species (Fig. 3.1). Equivalent to the stochastic blockmodel from the social science literature [230, 132], the group model uses network structure exclusively to form groups that often have straightforward ecological interpretations [221]. Indeed the species roles defined by the group model are essentially functional groups, in that species within a group tend to interact with the same sets of species in the same way.

Using the group model, we consider the structural distinction between parasites and free-living predators. Whether or not parasites alter general network metrics, if the patterns of their interactions are structurally unique within the network [43], then the groups identified by the model should reflect this distinction. For our study, we consider a set of large food webs that include information on both parasites and free-living species ([113, 106, 176, 243, 276]; Table 7.1), and we quantify how well the network's group structure matches broad trophic strategies. In general, it is difficult to identify the specific ecological drivers that contribute to the group structure. Here, we are able to isolate the effects of two ecological factors, degree and concomitant predation, which may influence group structure and how well it corresponds to the trophic strategies we expect. To examine the effect of degree, we compare

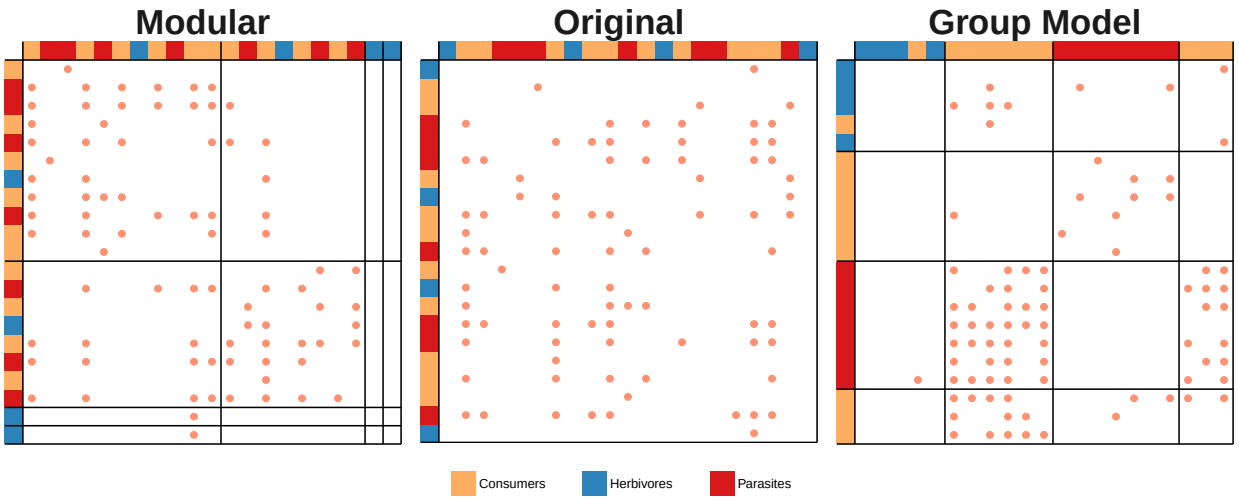

Figure 3.1: **Example of a group-model-produced grouping of an empirical adjacency matrix.** The center matrix represents a subset of one of the empirical webs used in this analysis. To the right and left are the same subset, but with the rows/columns reordered to maximize the modularity [183] (left) or the group model (right). Both modularity and the group model attempt to condense the links into groups of species that are strongly connected, producing a pattern in which the matrix is divided into areas with either very high or very low connectance. Note that, though the links appear randomly distributed before sorting, applying a walk-trap algorithm to find modules partitions the matrix into four groups, with links concentrated within modules (in blocks along the diagonal). Applying the group model (unrestricted for number of groups) also finds four groups, but partitions the matrix differently, creating more strongly connected blocks that are often (though not necessarily) off the diagonal. In all cases, links are indicated by orange dots and groupings by black lines. The trophic species of each node is indicated with a colored box along the margins.

the groupings found using a standard group model, and a variant of the model that removes the effect of degree on the group structure. To study concomitant predation, we compare groupings found when concomitant links are included and excluded.

We find that parasites perform unique roles in ecological communities, whether or not concomitant links are included and whether or not the model is corrected for degree. However, although concomitant links help distinguish parasites from free-living species, degree does not.

3.3 Materials and methods

3.3.1 Data

We analyzed the seven well-resolved marine and estuarine food webs described in [106], [276], [176], [243], and [113] [67]. We analyzed two versions of each network: one which includes concomitant links, and one which excludes them. Concomitant interactions are inferred links based on the assumption that predators eat all parasite species of their prey [68]. For all webs, parasites with complex life-cycles had their various life-stages aggregated into a single node.

Species were sorted into four trophic strategies: primary producer, herbivore, predator, or parasite. Primary producers were identified as any species with no prey. Herbivores were identified as species which consumed only primary producers. Parasites were identified based on [68]. All other species were labelled as consumers; therefore, this group contains both carnivores and omnivores.

3.3.2 Group Model

Metrics of categorizing network structure are common in analyses of ecological networks. One of the most popular of such metrics is modularity [183], which evaluates the presence of compartments within a community. These compartments contain individuals/species which interact more strongly with fellow members of their compartment than they do with members of other compartments (*e.g.* benthic verses pelagic species or flowers which bloom in the early versus late summer). This results in a structure of dense blocks along the diagonal of a matrix when properly ordered (Fig. 3.1). The group model can be thought of as a generalization of modularity, in which compartments are not defined exclusively by strong within-compartment connections, but rather by patterns of strong connections between compartments (*e.g.* between herbivores and primary producers). Note that the group

model does not exclude the possibility of strong connections with one's own compartment, such that modularity is a subset of the possible groupings identified by the group model.

The group model provides a likelihood-based framework to calculate how well a specific grouping fits the observed network structure. High-likelihood groupings will tend to have groups which act as functional groups, that is, species within a group tend to eat and be eaten by the same other groups. Consider a food web with S species and L links, represented by directed adjacency matrix A. Modelling the network as an Erdős Rényi random graph with connectance (the proportion of possible links that are realized) c, the likelihood of obtaining A can be given by:

$$Pr\left(A(S,L)|c\right) = c^{L}(1-c)^{S^2-L} \tag{3.1}$$

The likelihood is maximized when $\hat{c} = \frac{L}{S^2}$, the observed connectance. Using a partition (grouping) G containing g groups, we can split the network into a series of blocks, where each block represents all of the interactions from group r to group s, and where the groups contain S_r and S_s species, respectively. A block has L_{rs} links and connectance c_{rs} (note that because the network is directed, block rs is distinct from block sr). Then the full likelihood can be calculated as the product of the likelihoods of each individual block, as follows:

$$Pr(A(S,L)|c_{rs}, r, s \in 1:g) = \prod_{r=1}^{g}\prod_{s=1}^{g} c_{rs}^{L_{rs}}(1-c_{rs})^{S_rS_s-L_{rs}} \tag{3.2}$$

which is maximized when $\hat{c}_{rs} = \frac{L_{rs}}{S_rS_s}$ for every r, s.

Model selection can be performed by calculating the Bayes factor, or, equivalently, by choosing the partition with the highest marginal likelihood, which can be calculated as:

$$Pr(A|G) = \prod_{r=1}^{g}\prod_{s=1}^{g} \frac{L_{rs}!(S_rS_s - L_{rs})!}{(1+L_{rs})(1+S_rS_s)!} \tag{3.3}$$

For a full derivation of the group model and the Bayes factor, see [11] and [221].

The group model may be extended to correct for degree. For this version of the model, the marginal likelihood may be calculated as:

$$Pr(A|G) = \left[\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{g^2} \prod_{r=1}^{g}\prod_{s=1}^{g} (1+\beta)^{-(\alpha-L_{rs})}\,\Gamma(\alpha+L_{rs})\right] \times \left[\prod_{r=1}^{g} \frac{\prod_{i=1}^{S_r}\left(k_i^{\text{in}}\right)!\left(k_i^{\text{out}}!\right)}{\Gamma(S_r)^2\,\Gamma\left(S_r+\sum_{i=1}^{S_r}k_i^{\text{in}}\right)\Gamma\left(S_r+\sum_{i=1}^{S_r}k_i^{\text{out}}\right)}\right] \quad (3.4)$$

where α and β are parameters on a Gamma prior ($\alpha = 1$ and $\beta = 1$ used in our analyses), and k_i^{in} and k_i^{out} are the in-degree (number of prey/hosts) and out-degree (number of predators/parasites) for species i, respectively. For a derivation of the likelihoods and Bayes factor for the degree-corrected model, see the Supplemental Information and [132].

We searched for partitions that best fit the group model, one for each combination of the following variables: including/excluding concomitant predation, standard group model/degree corrected model, and maximum number of groups (2, 3, 5, 10, or 100). When up to 100 groups were allowed, the groupings collapsed down to a number that was more statistically parsimonious. Allowing for 100 groups gives the model the flexibility to find a truly optimal grouping, but constraining the number of groups makes the structure easier to visualize, understand, and interpret. In addition, it allows for a clear comparison between partitions with the same number of groups. Therefore, although we present results for all groupings, we focus on the 10 group case, which gives the model some flexibility, but is feasible to visualize and compare.

We used Metropolis-Coupled Markov Chain Monte Carlo (MC^3) with a Gibbs sampler to search for the partition of species into groups that maximizes the marginal likelihood (for details, see Supplemental Information and [221]). Since exhaustively searching all possible groupings is computationally infeasible, we performed 200 independent MC^3 runs for each

grouping reported, with 10 chains and 200,000 steps. Differences from the true optimum are likely to be small, so finding the true optimum is unlikely to have a large effect on the results. For convenience, we refer to the best partitions found as "best groupings", although they are not guaranteed to be optimal.

We studied the effect of degree correction on the groupings by calculating the mutual information between degree-corrected and degree-uncorrected partitions. For details, see the Supplementary Information.

3.3.3 Imbalance

Once the group structure was inferred, we evaluated how well these statistically defined groups correspond to ecologically relevant *a priori* partitions, such as those specifying general trophic strategies (*e.g.*, herbivores or parasites). We did this by characterizing the "imbalance" of the distribution of species employing a given strategy across the various groups specified by the group model. For instance, considering the distribution of parasites across the group structure, we can measure the imbalance by counting the number of parasites and the number of non-parasites in any given group:

$$\psi_i^{parasites} = \frac{max(\pi_i, \phi_i)}{\pi_i + \phi_i} \tag{3.5}$$

where π_i is the number of parasites in group i and ϕ_i is the number of free-living, *i.e.* non-parasitic, species. This index can range from $\frac{1}{2}$ in the case where both trophic strategies are present in equal numbers, to 1 when all species in the group employ the same strategy. For a given network and partition, we can calculate the full imbalance for a given trophic strategy by taking the product across all groups in the partition:

$$\Psi^{parasites} = \prod_{i=1}^{g} \psi_i = \prod_{i=1}^{g} \frac{max(\pi_i, \phi_i)}{\pi_i + \phi_i} \tag{3.6}$$

We calculated imbalance in this way for all trophic strategies we considered: primary producers, herbivores, free-living consumers, and parasites. We used a generalization of this measure to calculate the imbalance value for the full network (incorporating all strategies and groups simultaneously)[1]:

$$\Psi^{All} = \prod_{i=1}^{g} \prod_{k=1}^{c} \psi_i^k = \prod_{i=1}^{g} \prod_{k=1}^{c} \frac{max(\chi_i^1, \chi_i^2, \ldots, \chi_i^c)}{\sum_{k=1}^{c} \chi_i^k} \tag{3.7}$$

where g is the number of groups in the partition, c is the number of unique trophic strategies, and χ_i^k is the number of species with strategy k in group i. Note that $\left(\frac{1}{c}\right)^g \leq \Psi^{All} \leq 1$. To determine whether a given value for Ψ is higher than expected by chance, and therefore whether parasites significantly aggregate with other parasites in the partition, we wanted to associate Ψ with a p-value, measuring the probability of obtaining an equal or greater imbalance at random.

To get at this, we consider the following example. Suppose we partition S species into two groups ($g = 2$). The first group contains ϕ_1 free-living species and π_1 parasites, while the second group has ϕ_2 free-living species and π_2 parasites. Thus, the total number of parasites in the network is $P = \pi_1 + \pi_2 = \sum_{i=1}^{g} \pi_i$ and the total number of free-living species is $F = \phi_1 + \phi_2 = \sum_{i=1}^{g} \phi_i$. Clearly, $S = P + F$. The probability of obtaining exactly ϕ_1, ϕ_2, π_1, π_2 at random can be computed using the hypergeometric distribution:

$$\begin{aligned} Pr(\phi_1, \phi_2, \pi_1, \pi_2 | \vec{\mathcal{P}}, P, F) = \\ = \frac{\binom{P}{\pi_1}\binom{F}{\phi_1}}{\binom{P+F}{\pi_1+\phi_1}} \frac{\binom{P-\pi_1}{\pi_2}\binom{F-\phi_1}{\phi_2}}{\binom{P+F-\pi_1-\phi_1}{\pi_2+\phi_2}} = \frac{\binom{P}{\pi_1}\binom{F}{\phi_1}}{\binom{P+F}{\pi_1+\phi_1}} \cdot 1 = \frac{\binom{P}{\pi_1}\binom{F}{\phi_1}}{\binom{P+F}{\pi_1+\phi_1}} \\ = \frac{\binom{P}{\pi_2}\binom{F}{\phi_2}}{\binom{P+F}{\pi_2+\phi_2}} \frac{\binom{P-\pi_2}{\pi_1}\binom{F-\phi_2}{\phi_1}}{\binom{P+F-\pi_2-\phi_2}{\pi_1+\phi_1}} = \frac{\binom{P}{\pi_2}\binom{F}{\phi_2}}{\binom{P+F}{\pi_2+\phi_2}} \cdot 1 = \frac{\binom{P}{\pi_2}\binom{F}{\phi_2}}{\binom{P+F}{\pi_2+\phi_2}} \end{aligned} \tag{3.8}$$

1. Note that in the case of just two categories (*e.g.* parasites and non-parasites, this equation collapses into Eqn. 3.6 and is the same for both categories.

where $\vec{\mathcal{P}}$ is the partition structure (in this case there are only two groups, *i.e.*, $|\vec{\mathcal{P}}| = 2$) provided by the group model. Therefore, we can associate a probability of obtaining this result at random to each possible partition encompassing a given number of parasites and free-living species. The formula above can be generalized to an arbitrary number of groups $|\vec{\mathcal{P}}| = g$:

$$Pr(\vec{\phi}, \vec{\pi}|\vec{\mathcal{P}}, P, F) = \prod_{i=1}^{g} \frac{\binom{P-\sum_{j=0}^{j<i} \pi_j}{\pi_i} \binom{F-\sum_{j=0}^{j<i} \phi_j}{\phi_i}}{\binom{P+F-\sum_{j=0}^{j<i} (\pi_j+\phi_i)}{\pi_i+\phi_i}} \tag{3.9}$$

and trophic strategies c:

$$Pr(\vec{\chi^1}, \vec{\chi^2}, \ldots, \vec{\chi^c}|\vec{\mathcal{P}}, X^1, X^2, \ldots, X^c) = \prod_{i=1}^{g} \frac{\prod_{k=1}^{c} \binom{X^k-\sum_{j=0}^{j<i} \chi_j^k}{\chi_i^k}}{\binom{S-\sum_{j=0}^{j<i} \sum_{k=1}^{c} \chi_j^k}{\sum_{k=1}^{c} \chi_i^k}} \tag{3.10}$$

where by definition $\phi_0 = \pi_0 = \chi_0 = 0$, $X^k = \sum_{i=1}^{g} \chi_i^k$ is the total number of species with strategy k, and, as above, $S = \sum_{k=1}^{c} X^k$ is the total number of species in the network. For a description of how to calculate p-values based on these probabilities, see Box 1.

All data and code needed to run the search algorithm and perform all analyses may be found at `https://git.io/vXciH`.

Box 1. Calculation of partition imbalance p-values.

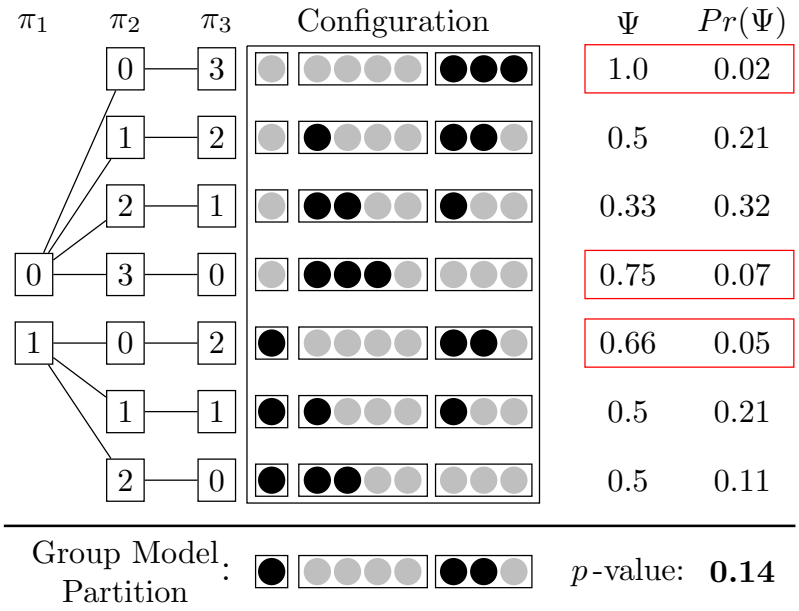

Figure 3.2: **Process for calculating the p-value for a hypothetical group model partition.** Under "Configuration", we list all seven unique configurations for the three parasites (black circles) and five free-living species (grey circles) into three groups (boxes) whose sizes have been determined by the partition produced by the group model. π_1, π_2, and π_3 show the combinatorial tree to obtain these configurations. Using equation 3.6, we associate each configuration with an imbalance value ψ. Next, using equation 3.9, we compute the probability of obtaining each configuration at random. Finally we sum the probabilities for all configurations with equal or greater imbalance than the empirical partititon (those in red boxes) to compute a p-value.

We are interested in the probability of observing an equal or larger value of imbalance at random. For networks with few groups and/or trophic strategies, we can compute this probability analytically by enumerating all possible cases and adding the probabilities of observing each imbalance value greater or equal to that observed. For example, take a network that is composed of 8 species ($S = 8$), of which 3 are parasites ($P = 3$) and 5 are free-living ($F = 5$). Suppose that when we use the group model to find the optimal partitioning into 3 groups we find that $\vec{\pi} = [1, 0, 2]$ and $\vec{\phi} = [0, 4, 1]$. The imbalance is $\Psi = \frac{2}{3}$. We can compute all the possible cases in which we arrange the P parasites and

F free-living species into 3 boxes of sizes $\vec{\mathcal{P}} = [1, 4, 3]$, with each configuration having an associated imbalance value and probability of obtaining this configuration at random. We can then compute a p-value for the empirical distribution of trophic strategies across the partition produced by the group model by summing the probabilities associated with the configurations yielding imbalance equal or higher than that found in the partition produced by the group model (Fig. 3.2).

Though this brute-force method becomes infeasible for networks with many groups or trophic strategies, we can still calculate the p-value numerically by comparing the observed imbalance to a large number of randomized species strategy distributions across the provided group structure, with fairly rapid convergence (Fig. S16).

3.4 Results

Results were similar across networks. We report statistical results across all networks, but for simplicity we display figures and imbalance scores only for the largest network (PuntaBanda) in the main text. Figures for the other six networks may be found in the Supplemental Information.

Partitions were significantly imbalanced in almost all cases, across different numbers of groups and all networks (Tables 3.1, S2-S7). Hence, the groupings maximizing the marginal likelihoods separated parasites (and other trophic strategies) from other strategies more than expected by chance alone. Results were significant whether or not the model corrected for degree, and whether or not concomitant predation was considered. The only exception was that producers and herbivores were often highly imbalanced, but not significantly. Since these networks generally had few producers and herbivores, this could be due to the relatively low statistical power. It is fairly easy for trophic strategies with few species to appear in the same groups simply by chance; as a result, trophic strategies with few species tend to have higher

imbalance, but lower significance. This is a common problem for permutation tests and other procedures involving discrete outcomes, such as in Fisher's exact test. Since raw imbalance scores depend both on the number of groups and the number of species in each trophic strategy, the scores alone can be misleading; for this reason, we focus our interpretation on the significance rather than the scores themselves. Producers did tend to group together, but these groups often contained non-producers as well. This could be because they were being consumed by a similar group of predators, *e.g.* by a group of omnivores. Parasites and predators in particular tended to form groups which were distinct from all other strategies (Fig. 3.3).

Average in- and out-degree varied across trophic strategies, whether or not concomitant predation was included (one-way ANOVA, $p < .0001$ for all four tests: in-degree without concomitant, $F = 31.01$; in-degree with conomitant, 126.2; out-degree without concomitant, $F = 84.93$; out-degree with concomitant, $F = 174.5$) (Fig. 3.4). Mean in-degree for consumers was significantly higher than other trophic strategies (mean in-degree $\mu_{in} = 15.64$ without concomitant predation, $\mu_{in} = 30.54$ with concomitant predation), followed by parasites ($\mu_{in} = 12.80$ without concomitant, $\mu_{in} = 13.16$ with concomitant), followed by producers and herbivores, which were not statistically distinct ($\mu_{in} = 0$ and 1.65 for producers and herbivores, respectively, both with and without concomitant predation). Mean out-degree was highest for producers, herbivores, and consumers without concomitant predation ($\mu_{in} = 14.71$, 13.73, and 13.25, respectively), with lower out degree for parasites ($\mu_{in} = 6.79$). When concomitant predation was added, the pattern flipped: out-degree for parasites was highest ($\mu_{in} = 31.00$), with the other trophic strategies significantly lower ($\mu_{in} = 14.71$, 13.88, and 13.25 for producers, herbivores, and consumers, respectively). In all four cases, parasites were significantly different from free-living consumers.

Despite the differences in degree between trophic strategies, the degree-corrected group model produced significantly more imbalanced groups overall (across webs, trophic strate-

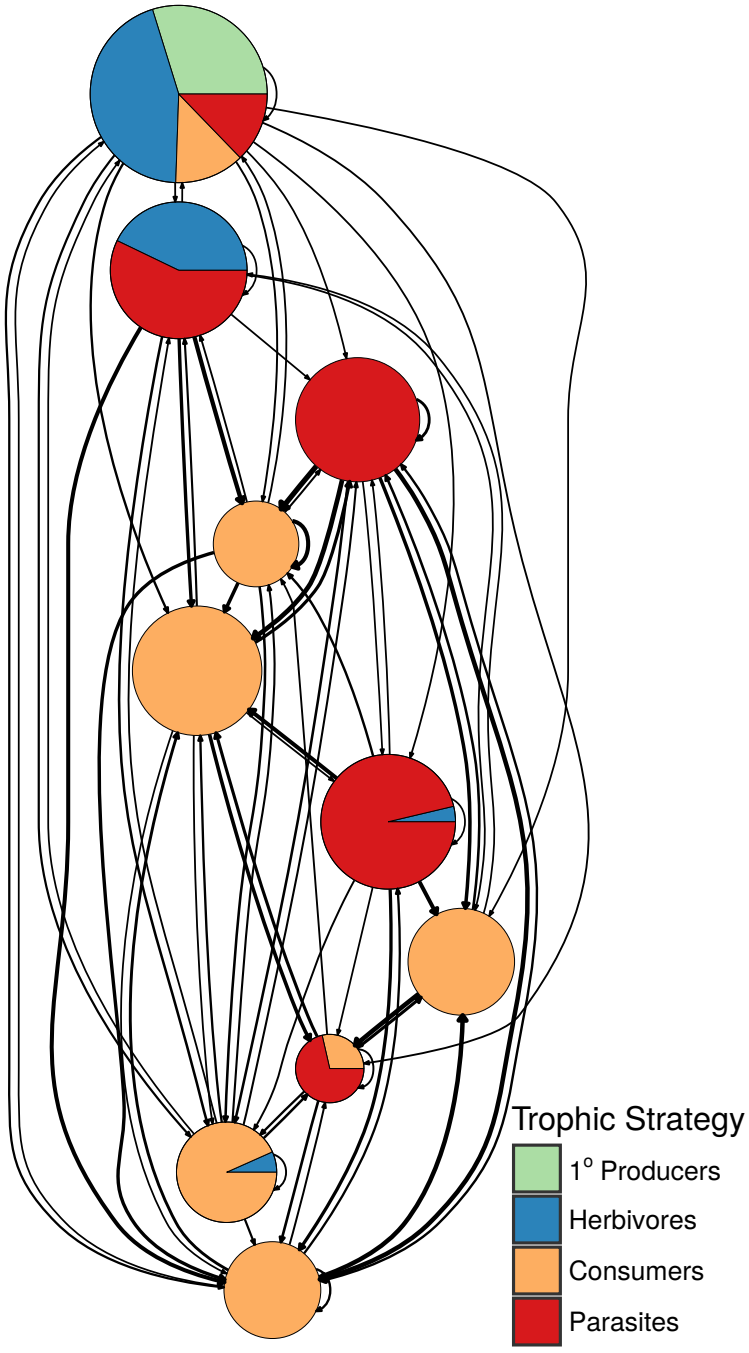

Figure 3.3: **Condensed graph representation of the Punta Banda network using the best partitioning found by the group model.** Each group is depicted by a pie-chart in which the fraction of nodes of each trophic strategy are indicated by the colored slices and the overall size is proportional to the number of nodes in the group. Number of links between (or within) groups is given by the thickness of each arrow.

gies, and number of groups), both when concomitant predation was included (paired t-test, estimated difference: .12, $p < .001$) and excluded (esimated difference: .067, $p < .001$). Including concomitant links also produced more imbalanced groups, under both the degree-corrected (paired t-test, estimated difference: .034, $p = .0049$) and uncorrected (estimated difference: .085, $p < .001$) models.

In general, degree-corrected and uncorrected partitions contained similar information when allowed to form more than 3 groups (Table S14). The mutual information between corrected and uncorrected partitions increases as the number of groups increases. As expected, degree-corrected partitions tended to have fairly evenly sized groups [132], whereas group size was significantly less even for uncorrected partitions, both with concomitant predation (paired t test, estimated difference in Pielou's evenness [208]: $-.030$, $p = .031$) and without (estimated difference: $-.025$, $p = .001$).

Finally, repeating the imbalance analysis for groupings of species according to subgraph-role participation (Supplemental Information, [236]) yield less consistently significant imbalance values (Tables S8-S13).

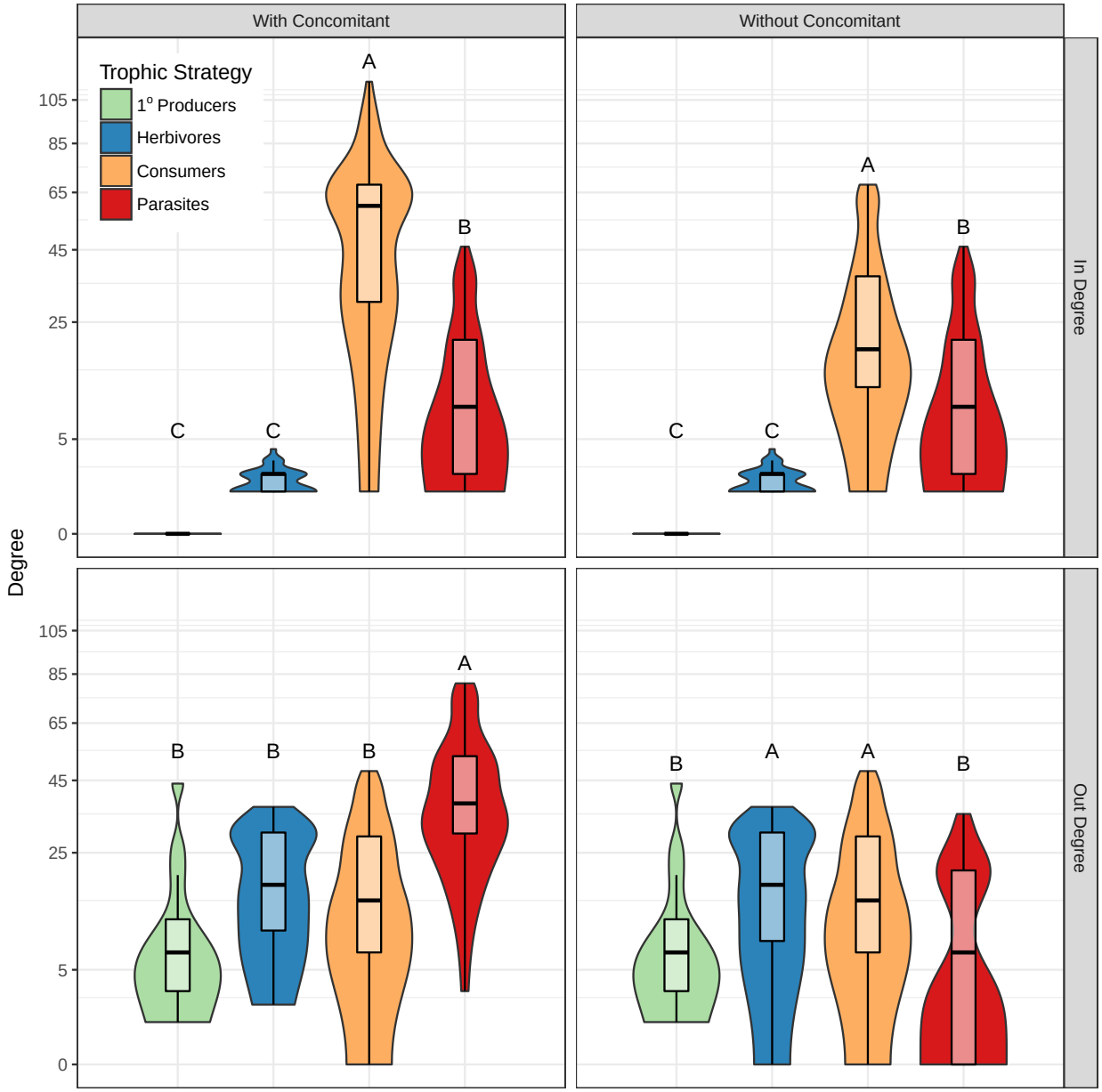

Figure 3.4: **Violin and boxplots of in-degree (number of prey) and out-degree (number of predators) for different trophic strategies in the Punta Banda network.** Degree is plotted on a square root scale. Boxes indicate the traditional 25^{th}, 50^{th} and 75^{th} quartiles, with wiskers extending to 1.5 times the inter-quartile range. Above each violin are grouping letters as indicated by a Tukey's HSD (honest significant difference) test.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.895***	0.737***	0.491***	0.421	0.246***
		3/3	0.826**	0.589	0.331***	0.260	0.120***
		5/5	0.757***	0.551***	0.628***	0.430***	0.307***
		10/10	0.585*	0.448***	0.439***	0.445***	0.214***
		18/100	0.481**	0.441***	0.494***	0.292***	0.103***
No	Yes	2/2	0.874	0.692	0.660***	0.449***	0.397***
		3/3	0.815	0.565	0.381***	0.521***	0.250***
		5/5	0.689	0.350	0.289***	0.559***	0.109***
		10/10	0.487	0.145	0.244***	0.138***	0.027***
		43/100	0.162***	0.036***	0.003***	0.005***	0.000***
Yes	No	2/2	0.898**	0.745***	0.912***	0.555***	0.555***
		3/3	0.839**	0.598	0.835***	0.603***	0.370***
		5/5	0.731	0.440*	0.702***	0.710***	0.327***
		10/10	0.702***	0.284***	0.582***	0.343***	0.164***
		23/100	0.411***	0.253***	0.590***	0.504***	0.169***
Yes	Yes	2/2	0.874	0.709***	0.812***	0.609***	0.537***
		3/3	0.833**	0.609***	0.564***	0.515***	0.282***
		5/5	0.716	0.360	0.365***	0.506***	0.173***
		10/10	0.451	0.198***	0.364***	0.445***	0.114***
		54/100	0.148***	0.006***	0.000***	0.024***	0.000***

Table 3.1: **Imbalance values for the PuntaBanda network (with or without concomitant predation) found via uniform sampling of 10^6 possible distributions of trophic strategies across the partitionings found by the (degree-corrected or otherwise) group model with g groups.** Significance is indicated by the trailing asterisks, with $p < 0.05$, $p < 0.01$, and $p < 0.001$, corresponding to $*$, $**$, and $***$, respectively. Note that the lower bound for the imbalance varies with the number of groups, so it is only appropriate to compare raw imbalance scores when the number of groups is the same.

3.5 Discussion

The group model can be used to find the coarse-grained ecological roles, similar to functional groups, that are present in a community. Here, we use the group model to identify general patterns in groupings across networks, to determine if parasites are structurally unique. Since groupings are based on the entire network structure, and the quality of a group depends on the quality of all other groups in the network, it is generally difficult to study how an ecologically relevant trait affects the group structure. In this study, we are able to consider the effect of two ecologically distinguishing features of parasites: concomitant predation and degree, by including or excluding concomitant links and by using a degree-corrected variant of the group model, respectively. We find that parasites are, in general, structurally distinct from free-living species, regardless of number of groups in the model, the inclusion or exclusion of concomitant predation, and whether or not the model corrects for degree.

Concomitant predation tends to increase the distinction between parasites and free living species. This is not very surprising; concomitant interactions can create loops in the network, causing it to look less "cascade-like", that is, less like a network where species only consume species which are below them along some niche axis (Figs. S2-S8). In the absence of parasites, food webs tend to follow a largely cascade-like structure [47], so the group model can easily use these interactions to distinguish parasitic species from free-living ones. Without concomitant links, parasites which have similar prey to free-living predators might end up in mixed groups of free-living and parasitic species; however, by including these links, parasitic species have additional predators that distinguish them from otherwise similar free-living species.

More surprising is the relationship we find with degree. Since parasites have different average in- and out-degree than free-living species, we might expect that degree would help the group model cluster parasites together. We found that the opposite was true: accounting for degree information in the group model still produced more imbalanced groups. This

suggests that degree is not the strongest structural signal that separates parasites from non-parasites. Indeed, although the mean degree is significantly different, the degree distribution of parasitic species overlaps considerably with the degree distributions of free-living categories. When uncorrected for degree, the group model tends to form a few groups with a small number of high-degree species. If the number of groups is constrained, as we have done here, this results in a few small groups and several larger groups. Thus, while the small groups may be highly imbalanced, the larger groups are often less imbalanced. The degree-corrected model counteracts this effect, producing groupings are significantly more even in size and even more strongly imbalanced. Relaxing the constraint on the number of groups improves the fit of the uncorrected model, and indeed, we see that the two models form more similar groupings as the number of groups increases (Table S14). This pattern suggests that the corrected and uncorrected models are identifying similar underlying structures, but that the uncorrected model "prioritizes" grouping off high-degree species over grouping species which are structurally similar. Put another way, the uncorrected model can be affected by high-degree outliers, especially when the number of groups is heavily constrained.

Finding little consolation in degree, one might think that the key to parasite structural uniqueness could be found in a slightly higher-order form of network structure, such as the local patterns of connectance termed "motifs" or subgraphs [68, 236]. Unfortunately, though some subgraph-roles have been associated with particular trophic strategies [43], we found that these trends were not as consistent as the group model at distinguishing strategies, and yielded less significant groupings on average.

These results provide evidence that parasites are structurally distinguished, not by how many predators and hosts they have, but by who those predators and hosts are at a global scale. What are the ecological drivers of this difference? The groups identified by the group model do not lend themselves to easy interpretation in terms of one, or even a combination of several, node-specific properties of the network (*e.g.* degree, subgraph-role, taxonomy,

body-size, trophic-level, centrality, *etc.*). Instead, the group model coalesces nodes that share similar roles within the network, *i.e.* species which interact with similar sets of other species in similar ways. Put another way, the group model acts upon the links between species, grouping links that go from and to similar species. Previous results have suggested that the group model is able to find more informative groupings than any one property alone [234], but perhaps some combination of properties not investigated here could lend a more ecological explanation for the similarity of species within groups.

Another area for further research is an investigation of the relationship between stages of complex parasitic life-cycles and the group model results. The food webs used in this analysis aggregated parasite life-stages into a single node. It would be interesting to repeat this analysis on food webs in which the life-stages were preserved as separate nodes, as such stages often have very different patterns of interaction with other species. Would parasites of all life-stages form imbalanced groups together, or are some life-stages more similar to predators or herbivores, resulting in less imbalanced groups overall? If they do group together initially, could this be a potential explanation for the later subdivision of parasites into multiple groups?

Parasites constitute a very broad set of organisms. They can vary in many ways: size (from microscopic viruses to parasitic worms reaching a meter or longer [217]); life cycle complexity; level of specialization; presence of free living stages; and whether they live in or on their hosts. They are also extraordinarily phylogenetically diverse. Given these major differences, it is encouraging to see that a human-chosen categorization as parasite is indeed structurally relevant in food webs.

3.6 Conclusion

Network structure has been found to influence many important features of ecological systems, including robustness [70], stability [10], and resilience [134]. General patterns of network

structure are also used to develop structural [47, 264, 206] and dynamic [28, 100] models. However, many of these models were developed from data that excluded parasites, and parasites violate many of the patterns that they are based on. For example, concomitant predation creates loops that violate the cascade model, and allometric patterns which hold for free-living species (*e.g.*, [35]), such as predator:prey body mass ratios, are inverted for parasitic interactions [220]. Models such as Allometric Diet Breadth Model [206] and the Allometric Trophic Network [28], which are based on body size data, are unlikely to capture parasites successfully.

Our finding that parasites have unique structural roles – in essence, form unique functional groups – suggests that existing food web models should be reevaluated to better fit these distinct structural patterns. This stands in contrast to previous work suggesting that parasites' effect on network structure is mainly due to changes in connectance and diversity [68]. Using the same set of networks, we instead find that parasites perform statistically distinct roles in networks, even when correcting for degree, and even when concomitant links are excluded. These results add to the growing evidence that parasites must be considered as we continue to study and model ecological networks.

3.7 Acknowledgements

Thanks to G Barabás, A Dobson, J Dunne, J Grilli, K Lafferty, and N Martinez for helpful discussions. ELS is supported by the NSF GRFP. MJM is supported by the U.S. Department of Education grant P200A150101. SA is supported by NSF DEB-1148867. This work was inspired by the Parasites and Food Webs Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (DEB-0553768), the University of California, Santa Barbara and the State of California.

3.8 Data accessiblity

All food web data used in this project can be found in the Dryad Digital Repository [67].

CHAPTER 4

ECOLOGICAL AND STRUCTURAL FACTORS PREDICT IMPORTANCE IN FOOD WEBS

Abstract

Local extinction of a species can have surprising and extreme impacts on a community. These effects are determined in part by the food web structure, that is, who eats whom in the ecological community. Here, I simulate equilibrium and extinction dynamics for large empirical and simulated food webs. I find that several network and population metrics can predict species importance, measured as the effect of species removal. In particular, high degree and trophic level correlates with higher importance, but high closeness centrality is assoicated with lower importance. These findings support previous empirical work on keystone predators and network robustness, and provide potential ways to identify species of conservation importance.

4.1 Introduction

Not all species in a community are created equal. Some may appear and vanish from the community to little effect, while the losing others can cause the system to collapse. Identifying important species is challenging, in part because of the complex network of interactions that occur in an ecological community. A species may affect others through both direct and indirect interactions. This means that, even if a species has low biomass or few interactions, its removal may have disproportionate effects on the system [192, 70].

A natural definition of species importance is how strongly its removal affects the rest of the community. As species across the planet continue to go locally and globally extinct, understanding and predicting the effects of extinction becomes increasingly important. Re-

moving species from a system can also provide ecological insights. Species removal is a common and useful tool in field ecology, and has been used to study theories such as keystone predation [191, 66], trophic cascades [223, 78], and the relationship between functional diversity and productivity [248].

What makes a species important? In this study, I consider predictors based on food web structure and ecological factors. As structural predictors, I consider three measures of centrality: degree, closeness centrality, and eigenvector centrality [12]. We would expect highly central species to be more important, since their removal would more likely have cascading effects on the rest of the community. As ecological predictors, I consider two measures: trophic level and population variability. Trophic level could be positively or negatively related to species importance, depending on whether top-down or bottom-up forces dominate the dynamics. Population variability may also underlie species importance. Important species might be expected to have low variability in population abundance, since species which are both important and variable could destabilize the system with their large fluctations.

To study if these factors predict species importance, I used large-scale simulations of entire ecological communities. Such an approach has only recently become feasible as computing power has increased, with only a couple of studies using either mass-balance [187] or stochastic [152] models. Here, I simulate the dynamics of networks using a discretized Lotka-Volterra model with growth rate perturbations to simulate environmental variability. I consider open and closed ecological communities, using both empirical network structures, and structures based on popular food web models. I find that many of the structural and ecological factors are predictive, and in particular, that high degree and trophic level are associated with high species importance. These results align with empirical and theoretical work on the importance of keystone predators and hub species for network robustness.

4.2 Materials and Methods

4.2.1 Food Web Simulation

I simulated dynamics for nine well-resolved food webs from a variety of ecosystems: the Tatoosh intertidal [221], Ythan Estuary, Scotland [101], the Sylt tidal basin of Germany and Denmark [244], the Caribbean Reef shelf of the Virgin Islands [190], the St. Marks National Wildlife Refuge seagrass bed in Florida, USA [42], the mudflat of Otago Harbor, New Zealand [177], the Flensburg Fjord of Germany and Denmark [277], and the Cariçaie Marsh at Lake Neuchâtel, Switzerland [34]. I chose these webs because they are large and well-resolved, and because the dynamics of these network structures proved numerically robust for simulation.

I also used networks generated from food web models, which allowed me to study how the results varied across networks generated in the same way. I used three common food web models that have structural similarities to empirical ecological networks: the Cascade Model [47], the Niche Model [264], and the Minimum Potential Niche, or MPN, Model [8]. For the MPN model, I generated networks with niche gap probabilities of 25, 35, and 45%. I generated thirty networks for each food web model, including thirty networks for each MPN gap probability. The size of each network was fixed at fifty species.

For each network, I simulated time series data using two simple generalized Type I Lotka-Volterra models: one closed model, and one allowing immigration. The standard continuous-time model with immigration is given by

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = N_i \left(b_i + \sum_{j=1}^{S} \alpha_{ij} N_j \right) + I_i \tag{4.1}$$

where N_i is the population size of species i, b_i is the intrinsic growth rate of species i, S is the number of species in the network, α_{ij} is the change in the per capita growth rate of species i to a unit change in species j's abundance, and I_i is the immigration rate of species i. In the closed model, I_i values were set to 0. A discretized model was used for simulation

purposes, since it improved the speed and robustness of the simulations. The discretized model is described by

$$N_{i,t+\Delta t} = N_{i,t} exp\left(\Delta t\left(b_i\left(1+\epsilon_{i,t}\right)+\sum_{j=1}^{S}\alpha_{ij}N_{j,t}+\frac{I_i}{N_{i,t}}\right)\right) \tag{4.2}$$

where $N_{i,t}$ is the population size of species i at time t, and $\epsilon_{i,t}$ is a function describing a small, time-varying perturbation of species i's growth rate. These perturbations were added to simulate background environmental variation.

I parameterized each network with α_{ij}, b_i, and equilibrium population size (N_i^*) values thirty times. Interaction strengths α_{ij} ($i \neq j$) were drawn from a lognormal distribution $\mathcal{LN}(0,1)$, truncated 5 standard deviations above the mean, to prevent outlier values from influencing the dynamics too strongly. All values past the truncation point were discarded and redrawn. Positive interaction strengths, denoting the positive effect of prey on predator, were multiplied by conversion rates that were drawn from a normal distribution ($\mathcal{N}(0.2, 0.01)$).

All simulations were initialized at a locally stable, feasible ($N_i^* > 0$ for all species) equilibrium. A system is locally stable when the real part of the leading eigenvalue λ_1 of the community matrix M is negative. The community matrix is defined as the partial of the growth rates with respect to the population sizes evaluated at equilibrium, given by

$$M_{ij} = \left.\frac{\partial r_i}{\partial N_j}\right|_* \tag{4.3}$$

$$= \alpha_{ii} - \frac{I_i}{(N_i^*)^2} \qquad \text{when } i = j, \tag{4.4}$$

$$= \alpha_{ij} \qquad \text{otherwise} \tag{4.5}$$

where r_i is the total per capita growth rate of species i, which reads

$$r_i = b_i + \sum_{j=1}^{S} \alpha_{ij} N_j + \frac{I_i}{N_i} \tag{4.6}$$

For the closed model, I added self-regulation for all species to ensure that the system was stable. First, I calculated the leading eigenvalue λ_1 for the community matrix M, with diagonal terms α_{ii} set to 0. If the system was not already stable, I enforced stability by setting self-regulation terms α_{ii} to $-1.1Re(\lambda_1)$. I then drew equilibrium population sizes from $\mathcal{LN}(0,1)$ (again truncated 5 standard deviations above the mean). Choosing a distribution with a positive support guaranteed that the system was feasible.

For the immigration model, I enforced stability through immigration rather than self-regulation. I set all α_{ii} terms to 0, meaning that the immigration terms needed to be large enough to stabilize the system. As in the closed model, I calculated the leading eigenvalue λ_1 for the community matrix M, with α_{ii}s and I_is set to 0. Also following the closed model, I wanted the diagonal terms of the community matrix to be smaller than $-\lambda_1$. To ensure this, a lower bound can be determined for the immigration terms using Eqn. 4.5:

$$\alpha_{ii} - \frac{I_i}{(N_i^*)^2} > \lambda_1 \tag{4.7}$$

$$I_i > \lambda_1 (N_i^*)^2 \tag{4.8}$$

A natural upper bound for the immigration terms is N_i^*, so that the immigration rate does not drown out the internal dynamics of the system. This implies that N_i^*s can be no larger than $\frac{1}{\lambda_1}$; otherwise, the lower bound for I_i would be greater than the upper bound. Following these constraints, N_i^*s were drawn from $\mathcal{LN}(0,1)$, truncated either 5 standard deviations above the mean or $\frac{1}{\lambda_1}$, whichever was smaller. Then the immigration rates were drawn from the uniform distribution $\mathcal{U}(\lambda_1 (N_i^*)^2, N_i^*)$.

Intrinsic growth rates b_i were then back-calculated as $b = -AN^*$, where b and N^* are

column vectors containing the intrinsic growth rates and equilibrium population sizes of each species, respectively, and A is a matrix containing the per-capita interaction strengths. For randomly constructed webs, I examined 30 parameterizations within each of the 30 different network structures within each type.

Small perturbations $\epsilon_{i,t}$ were added to the intrinsic growth rate to incorporate environmental stochasticity in the system. These perturbations were drawn from $\mathcal{N}(0, 0.01)$ for each time step t. The system was initialized at equilibrium and simulated for 1000 time steps with $\Delta t = 0.01$. All simulations and analyses were performed in R using the `igraph` and `lme4` packages [242, 52, 24]. Note that, since simulations were started at the (unperturbed) equilibrium state, there was no need to discard any transients from the beginning of the time series. Time series simulation models and code were developed in collaboration with György Barabás, and are available on GitHub at `https://github.com/elsander/PredictImportance` as an installable R package.

4.2.2 Measuring Variability

Based on the simulated population dynamics, I calculated mean abundance (μ) and temporal variance (σ^2) for each species in each parameterization of each network. According to Taylor's Law, log mean abundance is expected to scale linearly with the log standard deviation [241]. This was found to be the case for the vast majority of simulations of these networks.

I calculated the best-fit regression line for $ln(\sigma) \sim ln(\mu)$ for each simulation, and used the residuals as a scale-free metric of population variability. This residual variability metric represents the population variability not explained by abundance alone. Any simulations which did not result in a significant $ln(\sigma) \sim ln(\mu)$ relationship were dropped from the analysis.

4.2.3 Network Metrics

Closeness centrality was calculated as the inverse of the sum of the lengths of the shortest directed paths from node i to all of the other nodes in the network [85]:

$$\frac{1}{\sum_{i \neq j} d_{ij}} \tag{4.9}$$

where d_{ij} is the shortest directed distance from node (species) i to node j.

Eigenvector centrality was calculated following [12]. This measure is mathematically related to Google's PageRank$^{\text{TM}}$, which ranks webpages as more important if important pages link to them. Intuitively, this measure can be thought of as an extension of degree, in that nodes are more central not just because they are highly connected, but because they are connected to highly connected nodes, which are connected to highly connected nodes, and so on. This recursive problem may be solved using a column-averaged transformation $\bar{C}$ of the adjacency matrix C, such that

$$\bar{C}_{ij} = \frac{C_{ij}}{\sum_j C_{ij}} \tag{4.10}$$

For webpages, C_{ij} represents a link from page i to page j, and $\bar{C}_{ij}$ represents the fraction of i's importance that is given by its connection to page j. Then, the eigenvector centrality may be calculated using the eigenvector associated with the dominant eigenvalue of $\bar{C}$. To ensure that this eigenvector exists, I follow [12] and add a detritus node, to which every other node contributes and which contributes to all primary producers.

Trophic Level (T) was calculated as

$$T = (I - \bar{C})^{-1} u \tag{4.11}$$

where I is the identity matrix, and u is a vector with each entry equal to one [265]. The matrix inverse above is known as the Leontief inverse; note that this formulation of T assumes

that there is no consumer preference for any particular resource. The formula works as long as the inverse exists, which was always the case with our data.

4.2.4 Species Importance

Species importance was calculated as the change in community composition due to the removal of a species. I calculated this by setting the abundance of the focal species to zero, then simulating the deterministic community dynamics for 1000 time steps, long enough for the community to reach a new equilibrium. If any remaining species reached an abundance below a precision cutoff of 10^{-6} at the end of the simulation, it was considered extinct and its abundance was set to 0. I then calculated importance as the Jaccard distance between the community abundance distributions before and after simulation. The focal species was held at abundance 0 in the "before" community for purposes of this calculation, to avoid biasing the measure toward more abundant species. Jaccard distance(J) is a measure of community dissimilarity that can be used to compare two abundance distributions and is calculated as:

$$J = \frac{2B}{1+B} \tag{4.12}$$

$$B = \frac{\sum_{i=1}^{S} x_i - y_i}{\sum_{i=1}^{S} x_i + y_i} \tag{4.13}$$

where B is the Bray-Curtis dissimilarity, S is the number of species, x_i is the abundance of species i in the intact community, and y_i is the abundance of species$i$ in the community with the focal species removed.

4.2.5 Analysis of Simulated Data

I used multilevel regression models to investigate the relationship between importance and residual variability and the network metrics. Jaccard distance was log odds transformed,

which improved the normality of the distribution and matched its support with that of the other variables. All predictors except for trophic level were log transformed to improve the normality of the distributions. Each food web model type and each empirical network was analyzed in a separate multilevel regression. Due to a high correlation between slope and intercept for all predictors, slope but not intercept was allowed to vary by parameterization. I standardized all independent variables (subtracted the mean and divided by the standard deviation) before modelling. For the empirical network analysis, the slope for residual variability was allowed to vary by parameterization, for a regression with the following structure:

$$\begin{aligned} Importance_{p,s} \sim \beta_0 + \beta_{1,p} ln(CV_{p,s}) + \beta_2 ln(Degree_s) \\ + \beta_3 ln(Closeness_s) + \beta_4 ln(Eigenvector_s) + \beta_5 (TrophicLevel_s) \end{aligned} \quad (4.14)$$

where p is a specific parameterization of the network and s is a species in the network. The food web model simulations contained an additional layer of hierarchy, resulting in a regression with the structure:

$$\begin{aligned} Importance_{n,p,s} \sim \beta_0 + \beta_{1,n,p} ln(CV_{n,p,s}) + \beta_{2,n} ln(Degree_{n,s}) \\ + \beta_{n,3} ln(Closeness_{n,s}) + \beta_{n,4} ln(Eigenvector_{n,s}) + \beta_{5,n} (TrophicLevel_{n,s}) \end{aligned} \quad (4.15)$$

This is identical to the regressions for the empirical networks, except that multiple network structures are being incorporated into the same multilevel model. This means that β_1 through β_5 now vary by n, the network structure.

The significance of each predictor in each model was tested by fitting the model without the predictor, and performing a likelihood ratio test between the full model and the model without the predictor of interest, using the `lme4` package in the language `R` [242, 24].

Significance levels were Bonferroni corrected in four sets: immigration model with simulated networks, immigration model with empirical networks, closed model with simulated networks, and closed model with empirical networs. These multiple comparison correction groups were chosen because conclusions were drawn based on general trends at this level of the data.

4.3 Results

In the closed model (no immigration), all variables significantly predicted species importance in almost all webs, both simulated and empirical (Figs. 4.1a, 4.1b). Degree positively predicted importance, and was the strongest predictor overall, with a median standardized effect size of .56 in empirical webs and .53 in simulated webs. Trophic level (median effect size .24 empirical, .28 simulated) and eigenvector centrality (.09 empirical, .04 simulated) also positively predicted importance. Closeness centrality (-.22 empirical, -.18 simulated) and residual variability (-.12 empirical, -.14 simulated) negatively predicted importance.

Degree, closeness centrality, and trophic level were significant predictors in all webs. Variability and eigenvector centrality were each significant in 7 out of 9 empirical webs (Table 4.1). All predictors were significant for all simulated webs (Table 4.2).

Trends were very similar in the immigration model (Figs. 4.1c, 4.1d). All variables significantly predicted species importance in most empirical webs. Degree (median effect size .63 empirical, .64 simulated), trophic level (effect size .32 empirical, .42 simulated), and eigenvector centrality (.11 empirical, .05 simulated) positively predicted importance, and closeness centrality (-.16 empirical, -.34 simulated) negatively predicted importance. Although residual variability negatively predicted importance in empirical webs (median effect size -.12), it was non significant for all simulated webs but the cascade model (median effect size 0).

As in the closed model, degree and trophic level were significant in all webs (Tables 4.3,

4.4). Variability was significant in 7 out of 9 empirical webs and 1 out of 5 simulated webs. Closeness centrality was significant in 8 out of 9 empirical webs, and eigenvector centrality was significant in 6 out of 9 empirical webs. All predictors except variability were significant in all simulated webs.

Web	Variability	Degree	Closeness	Eigenvector	Trophic Level
Caricaie	**17.61**	**603.40**	3.40	0.65	**121.60**
Flensburg	**20.20**	**1742.70**	**128.58**	**58.01**	**130.83**
Otago	**11.03**	**886.85**	**46.33**	**10.94**	**135.22**
Reef	**195.55**	**579.78**	**12.13**	**81.41**	**346.61**
Serengeti	**22.90**	**1194.64**	**419.37**	**36.53**	**53.95**
Stmarks	**21.41**	**503.81**	**59.79**	**37.57**	**102.27**
Sylt	**27.15**	**698.55**	**21.09**	0.43	**333.92**
Tatoosh	0.36	**672.31**	**66.78**	9.25	**54.96**
Ythan	0.00	**1241.60**	**107.87**	**12.43**	**28.22**

Table 4.1: **Significance of predictors for the closed model with empirical network structures.** Rows are specific network structures, and columns are predictors. Values in the table are test statistics based on a likelihood ratio test comparing the full model to the model with the specified predictor excluded. Significant results ($\alpha = .05$, Bonferroni corrected to $\alpha = .0011$) are indicated in bold.

Web	Variability	Degree	Closeness	Eigenvector	Trophic Level
Cascade	**394.71**	**2117.52**	**297.10**	**55.82**	**345.55**
MPN25	**110.18**	**6674.53**	**588.90**	**74.33**	**1050.35**
MPN35	**188.56**	**5471.80**	**411.88**	**70.17**	**1412.35**
MPN45	**246.50**	**6202.32**	**575.41**	**67.83**	**1170.57**
Niche	**350.82**	**4904.14**	**284.42**	**68.22**	**1464.58**

Table 4.2: **Significance of predictors for the closed model with simulated network structures.** Structured as in 4.1, with α levels Bonferroni corrected to $\alpha = .002$.

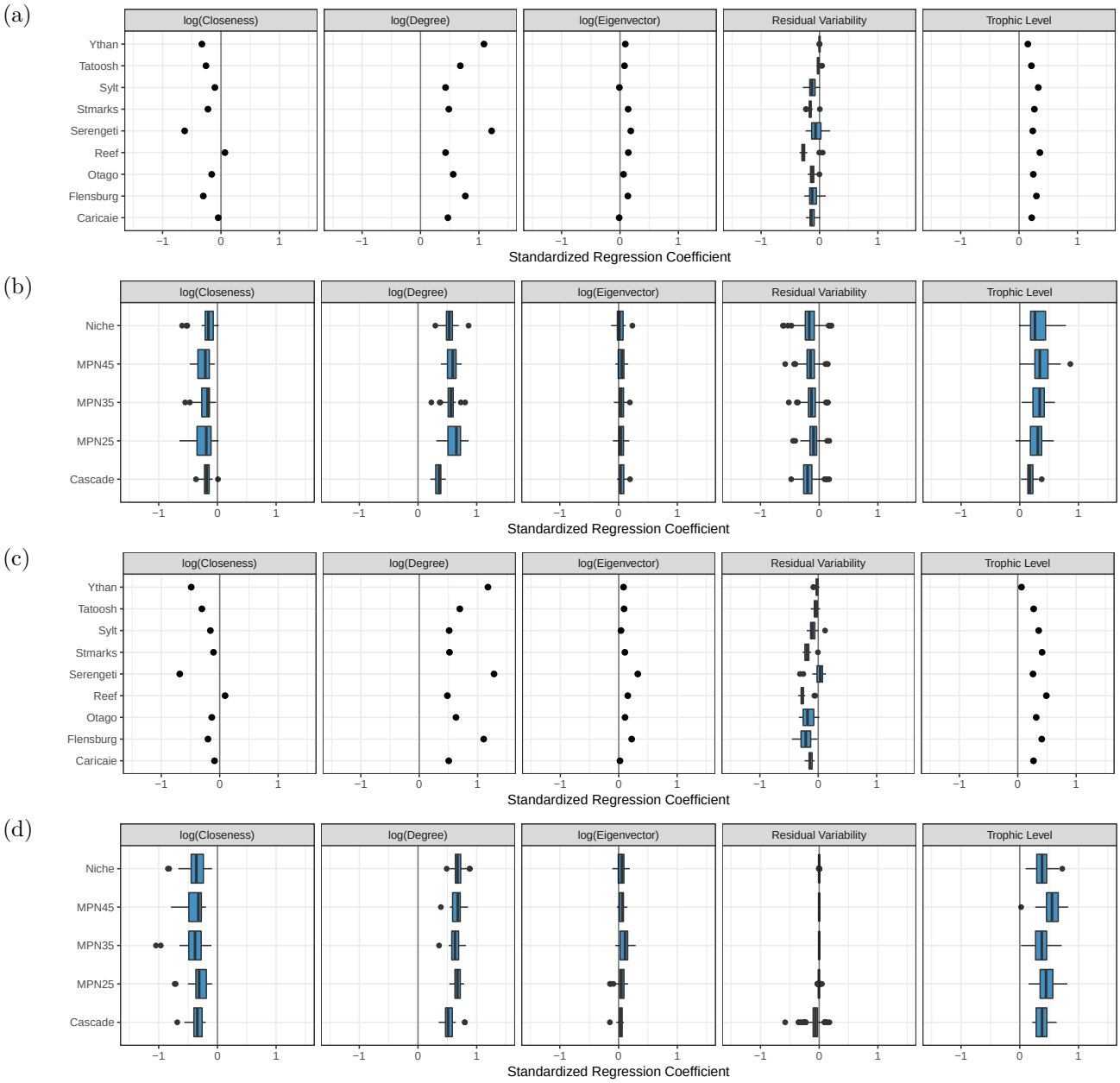

Figure 4.1: **Distributions of predictor effect sizes from hierarchical models.** Plots shown for closed model with empirical (a) and simulated (b) networks, and immigration model with empirical (c) and simulated (d) networks. Each row represents a network structure (empirical networks) or food web model which was used to generate network structures (simulated networks). Fixed effects in (a) and (c) are plotted with points at the mean and lines for the standard errors (too small to be visible). Random effects are represented as boxplots.

Web	Variability	Degree	Closeness	Eigenvector	Trophic Level
Caricaie	**32.44**	**1160.53**	**17.77**	2.23	**323.07**
Flensburg	**34.92**	**1913.50**	**30.97**	**89.89**	**136.48**
Otago	**40.05**	**1253.12**	**41.04**	**41.21**	**271.73**
Reef	**287.29**	**1009.61**	**29.20**	**126.37**	**872.22**
Serengeti	**11.10**	**716.01**	**246.31**	**63.93**	**39.97**
Stmarks	**42.56**	**771.20**	**18.05**	**28.79**	**333.46**
Sylt	**15.99**	**816.38**	**37.39**	4.68	**303.35**
Tatoosh	2.77	**693.56**	**96.51**	**12.97**	**91.03**
Ythan	0.30	**1317.12**	**172.05**	8.76	4.45

Table 4.3: **Significance of predictors for the open model with empirical network structures.** Structured as in 4.1, with α levels Bonferroni corrected to $\alpha = .0011$.

Web	Variability	Degree	Closeness	Eigenvector	Trophic Level
Cascade	**69.14**	**3717.57**	**870.94**	**41.86**	**1341.94**
MPN25	0.55	**5790.12**	**796.56**	**66.77**	**1967.19**
MPN35	0.00	**5188.99**	**1300.75**	**157.28**	**1599.26**
MPN45	0.00	**5861.96**	**1186.55**	**56.02**	**1632.47**
Niche	0.00	**6296.94**	**1208.84**	**97.08**	**1527.72**

Table 4.4: **Significance of predictors for the open model with simulated network structures.** Structured as in 4.1, with α levels Bonferroni corrected to $\alpha = .002$.

4.4 Discussion

Across multiple types of models and network structures, I consistently found that degree and trophic level positively predicted importance, while closeness centrality negatively predicted importance. Eigenvector centrality was also a fairly consistent positive predictor, although its effect was very weak. It is not possible to test all variations of all models, but the fact that these predictors were consistent across different parameterizations, network structures, and models (open or closed) suggests that they are somewhat robust.

These results also connect well to previous findings in the food web and community ecology literature. Many studies have found that removing high-degree species results in more secondary extinctions than removing low-degree species or randomly removing species [70, 71, 231], but see [209]; this is also true in pollination networks [170]. Since degree, a local centrality measure, is such a strong positive predictor, one might expect that closeness

centrality would be a positive predictor as well. Surprisingly, I found the opposite pattern. Closeness centrality comes from the study of social networks and does not have an obvious ecological interpretation; therefore, the reason for this unexpected inverse relationship is not obvious. However, this result does underline the fact that traditional network centrality measures are not necessarily simple proxies for ecological importance. Using them as measures of "positional importance", as some studies have suggested [253, 80, 124, 123], may produce misleading results for many ecological questions. As these authors suggest, it is vital to carefully consider the question, and if the chosen index will truly answer it.

Eigenvector centrality, unlike closeness centrality, was a fairly consistent positive predictor. Still, its low predictive power was surprising, given that Allesina and Pascual found it to predict the severity of secondary extinctions [12]. Allesina and Pascual were addressing a slightly different question than is being considered here; the Jaccard Distance acts as a measure of community change, not community collapse. Unlike most previous studies in this area, I simulated the abundances of the species, and was therefore able to capture not just secondary extinctions, but changes in abundance as well. Some species might be important not because they maintain community diversity, but because they suppress populations of their prey and competitors. It is possible that eigenvector centrality captures secondary extinctions, but not these other changes in community composition.

Variability was generally a negative predictor, but had a low effect size and was not very consistent, especially in simulated network structures with immigration. This could be due in part to the source of variability in the models used. Like all models, the models used here are a simplification, and assume that fluctuations in the system occur as a result of random fluctuations in intrinsic growth rates. Although it is obvious that weather and resource availability can impact birth and death rates, environmental factors may affect the system in other ways as well. Disturbances may act on the population size directly, as from forest fires [4] or wave action [196]. External factors may also affect interactions between species,

as has been seen in the decoupling of plants and their pollinators due to climate change [109]. The current study weakly suggests that more variable species are less important, but further empirical and computational work may reveal the strength and consistency of this effect, and how it is influenced by the type of environmental perturbation.

The effect of trophic level was much stronger and more consistent. Trophic level's positive relationship with species importance suggests that the simulated dynamics were strongly affected by top-down forces. The higher importance of species with higher trophic level could reflect trophic cascades [39, 195]. This pattern appeared despite a lack of clearly delineated trophic levels in either the empirical or the simulated network structures which I considered. Although the classic examples of trophic cascades occur in systems with a clear trophic structure (*e.g.*, [66, 78]), cascade-like patterns have been observed empirically even in systems without distinct trophic levels [213, 172].

Conclusion

Simulations are a powerful tool for ecologists, and can be used to support and generalize from empirical work. Using extensive simulation of large ecological communities, I find support for the importance of top-down forces and highly connected species in maintaining the composition of ecological communities. Identifying important species is vital both for understanding ecological processes and prioritizing conservation effeorts. Targeting these conservation efforts is especially relevant, given that top predators are particularly vulnerable to extinction due to environmental and anthropogenic factors [205, 55, 207]. By targeting our efforts appropriately, we may hope to mitigate the damage to complex ecological communities.

CHAPTER 5

ECOLOGICAL NETWORK INFERENCE FROM LONG-TERM PRESENCE-ABSENCE DATA

5.1 Abstract

Ecological communities are characterized by complex networks of trophic and nontrophic interactions, which shape the dynamics of the community. Machine learning and correlational methods are increasingly popular for inferring networks from co-occurrence and time series data, particularly in microbial systems. In this study, we test the suitability of these methods for inferring ecological interactions by constructing networks using Dynamic Bayesian Networks, Lasso regression, and Pearson's correlation coefficient, then comparing the model networks to empirical trophic and nontrophic webs in two ecological systems. We find that although each model significantly replicates the structure of at least one empirical network, no model significantly predicts network structure in both systems, and no model is clearly superior to the others. We also find that networks inferred for the Tatoosh intertidal match the nontrophic network much more closely than the trophic one, possibly due to the challenges of identifying trophic interactions from presence-absence data. Our findings suggest that although these methods hold some promise for ecological network inference, presence-absence data does not provide enough signal for models to consistently identify interactions, and networks inferred from these data should be interpreted with caution.

5.2 Introduction

Species interactions are a major driver of population dynamics. Interactions may produce predator-prey [239] or host-pathogen [18] cycling, competitive exclusion [20], or complex dynamics resulting from multiple interactions [154, 72]. In general, it is easier to collect data

about dynamics than observing interactions directly; as a result, a major goal in ecological modeling is the inference and study of ecological interactions based on dynamical data. Mechanistic models are a popular approach, and can provide a way to reveal underlying biological processes [136]. However, these models often require detailed information about the system, including the values of underlying parameters and the functional forms describing the ecological interaction itself. To study the interactions between more than a few species, it is necessary to find a more scalable approach.

At the community level, ecological interactions produce a complex network. Food web and interaction web structures affect dynamical properties such as community stability [165, 13], reactivity [240], and robustness to extinction [70, 9]. As with individual two-species interactions, we expect the structure of interaction networks to be reflected in the community dynamics. Whole network inference from community dynamics is intractable using standard mechanistic approaches. State-space reconstruction has been suggested as a model-free method for inferring pairwise ecological interactions [238, 44], but has been critiqued for its sensitivity to common features of natural systems, such as process error [46].

Machine learning and correlational methods have increasingly been used to infer microbial [81, 138] and gene expression [275, 150, 159] networks from time-series and co-occurrence data. These methods are convenient in that underlying parameters need not be estimated or even specified; instead, the model is fit based on observational data alone, and, when possible, tested using cross-validation or out-of-sample data (data that is used not to fit the model, but to test its predictive accuracy). This process produces "association networks" that reflect co-occurrence or correlational patterns. The hope is that these association networks reflect ecological dynamics, based on the much-debated theory that nonrandom co-occurrence distributions result from ecological interactions [59, 256, 96].

Experimentally or observationally verified microbial interaction networks are not commonly available, so it is difficult to determine if these methods effectively capture ecological

interactions [81]. Macroscopic communities are useful to study here, because a combination of observational, experimental, and network data are available for a few well-studied systems. Testing the validity of inferred networks for these systems can benefit both microbial ecology and traditional community ecology. If the models produce high-quality networks, we can have more trust in these methods in systems at either scale; if the networks do not correspond to empirical knowledge, we need to be cautious in interpreting the association networks currently being produced for microbial systems. Only one study to our knowledge has applied network inference methods to macro-ecological systems [174]. This study found that Bayesian networks successfully identified many known interactions among species and habitat variables. We expand on this work by considering presence-absence data for multiple large communities, using time series data for dynamic rather than static methods.

It is important to consider what the inferred co-occurrence networks mean ecologically. An inferred link represents a predictive or correlational link between one species and another species in the future, links which do not necessarily require interaction or causality. However, it seems reasonable to expect that when two species interact, that interaction will be reflected in the dynamics, and therefore, that we could identify many interactions based on observed dynamics. Presence-absence data dampens this dynamical signal, making it more challenging to correctly infer interactions. With this weaker signal, it may be easier to infer some types of interactions better than others. For example, predator-prey interactions tend not to cause total extinction of either species, and therefore might be difficult to identify from presence-absence alone. In contrast, competition can result in local extinction from competitive exclusion, and facilitation may allow for the successful invasion of a locally absent species. For this reason, we expect nontrophic interactions such as competition and facilitation to be more easily identified than trophic ones.

We use three network inference methods, which represent major classes of algorithms used to construct gene and microbial networks. The first approach uses Dynamic Bayesian

Networks (DBNs), which uses patterns of conditional independence between variables to infer a graphical structure. The second is Lasso regression, which constrains the inclusion of regression coefficients in the model. The final method uses Pearson's correlation coefficients with a significance threshold. All methods are presented in detail in the methods. Correlational approaches are especially common in microbial ecology [40, 23, 29, 201], and are distinct in that they are not predictive models (*i.e.*, the Pearson's correlation coeffient method can be used to construct an association network, but not to predict future dynamics, as DBNs and Lasso can). We train the models to predict which species are present based on which were present in the previous sample time. This allows us to capture the relationships between species through time, rather than assuming that species respond to each other instantaneously.

To evaluate how successfully these three models infer networks from ecological data, we apply them to two long-term presence-absence datasets: a riverine fish community in France, and a diverse intertidal community on Tatoosh Island in Washington state. These datasets differ in several ways: the France dataset spans a larger spatial scale, but contains fewer species and less phylogenetic diversity than the Tatoosh system. Since it includes only fish, it might be considered a guild, rather than a complete community. This allows us to evaluate the robustness of these inference methods under different data scenarios, and for different types of communities. We consider how well the models infer both trophic networks in both systems, and the nontrophic network from the Tatoosh intertidal. We find that each class of model inference significantly predicts at least one network structure, but that no class of model successfully predicts network structures for both France and Tatoosh, and no method outperforms the others in general. We also find that the models are completely unable to capture the Tatoosh trophic network, but that multiple models infer networks which are similar to the Tatoosh nontrophic network, suggesting either that nontrophic interactions drive dynamics in this system, or that the dynamical signal of trophic interactions gets lost

when using presence-absence data. Overall, although these methods have some predictive ability even with the limitation of presence-absence data, they should be used and interpreted with care.

5.3 Methods

We used multi-site, presence-absence, time series datasets from two systems: stream fish in France, and the middle intertidal zone on Tatoosh Island. We predicted network structures using three approaches: Dynamic Bayesian Networks, Lasso regression, and a method based on Pearson's correlation coefficients. We then compared the model-inferred network structures to networks based on empirical data. For the two predictive models (DBNs and Lasso), we also used the models to make short-term (next time step) predictions using out-of-sample data; that is, data which were not used to fit the model, and were set aside to test prediction accuracy. Code used to analyze these data is available on GitHub at `https://git.io/vDEFu`.

5.3.1 Data

The French stream fish dataset is a long-term monitoring dataset collected by ONEMA, the French National Agency for Water and Aquatic Environments. We used the data available on Dryad [49], which includes presence-absence information for 32 common species of fish, collected at 794 sites from 1992 to 2011, and split into eight time periods, with each time period spanning either two or three years. Not all sites contained data from all time periods. Because we used models that relied on the time series nature of the data, only consecutive sets of time points were used. For more information on how these data were collected, see [48, 212].

We constructed a food web for these species based on gut content analyses from 88 articles. This count excludes studies that only identified cannibalistic interactions, only identified interactions with species not present in the network, or did not identify prey

to the species level. Feeding interactions were incorporated into the network if they were observed in at least one study. Prey items were excluded when identity was unclear or not resolved to the species level, and when interactions were based on a single or spurious observation. Consumption of eggs, larvae, and fry were included as interactions if they were identified to the species level. Cannibalistic interactions were excluded from the food web, since self-interactions in the inferred networks mostly represent autocorrelation rather than cannibalism. Since this dataset contains only fish, all interactions in this food web are piscivorous.

While it is possible that this literature-based food web is missing some feeding interactions, all species in the dataset are quite common and many articles containing dietary information were available. Another possible concern is that the "sampling effort" is different for different species, since species may differ in interest for research or fishery purposes. Fortunately, many gut content analyses were available for most fish, especially for highly piscivorous species. Of all species in the web identified as partially or entirely piscivorous on FishBase [88], we only had difficulty finding dietary information for the barbel *Barbus barbus*, which we consequently removed from both the food web and time-series data. A listing of the feeding interactions and the citations supporting them can be found in Table S3.

The Tatoosh Island presence-absence dataset is a long-term observational and experimental dataset from the middle intertidal zone, non-destructively sampled at 30 plots annually. Ten control and ten experimental removal plots were sampled from 1994 to 2012, with an additional five control and five experimental plots sampled from 1998 to 2012. In the experimental plots, the competitive dominant *Mytilus californianus* was chronically and selectively removed. Censuses were performed in each plot by recording percent cover (sessile species) or numbers seen (mobile species) with the aid of a quadrat subdivided with microfilament lines into 121 sub-squares. For further details on the census methods, see [272, 274]. Census

data were converted to presence-absence after collection.

The middle intertidal zone is dominated in biomass by the mussel *Mytilus californianus*, which grows in dense mats. These mats are occasionally torn from the rock by strong wave action, revealing patches of bare rock, which then follows a predictable successional pattern of colonizing species, before being taken over by *M. californianus* again [196, 270]. The mussel beds provide physical structure for a diverse set of sessile and mobile species. The non-destructive nature of the sampling procedure makes it possible to study the undisturbed dynamics of the community, but as a result may miss some species that live deeper in the mussel bed. 107 species were identified across 530 site-years, but only the 60 species which were observed 10 or more times were used in the analysis.

The composition of plots with and without mussels is strikingly different, with control plots generally dominated by *M. californianus*. *Mytilus californianus* is so dominant in these plots that it may mask dynamics between other species. There is no single dominant species across the experimental plots, and including these data may reveal species interactions that would be hard for a model to identify from the control data alone. We attempted to fit separate models for the control and experimental plots, but the data were too limited. Because *M. californianus* removal has been found to dampen environmental stochasticity, but not to affect the temporal order of the dynamics [272] (that is, it changes the size of fluctuations in the system, but not the timescale over which they occur), we believe that it is reasonable to model the system using both experimental and control plots.

Interaction data for Tatoosh Island were collected based on observational and natural history information obtained in nearly a half century of study at the site [221]. We used two versions of the network: a trophic web, with only feeding interactions, and a nontrophic web with only non-feeding interactions, including competition, mutualism, facilitation, commensalism, and amensalism.

5.3.2 Model Training and Cross-Validation

Networks were constructed using three methods: Dynamic Bayesian Networks, Lasso regression, and Pearson's correlation coefficient. Models were fit using training data, and tested for predictive ability on a separate set of test data. Since the method based on Pearson's correlation coefficients does not produce a predictive model, it was used only to fit a network structure. With the exception of DBN structure learning, all model fitting and prediction was performed in R [242] using the package glmnet for Lasso regression [86] and helper functions from several other packages [262, 260, 36, 257, 258, 261, 259, 38, 54].

The France dataset was split into training and test sets (60 and 40 percent of sites, respectively). The Tatoosh dataset was small enough (30 sites with annual samples for 15 or 19 years each) that setting aside a test set was not feasible. Instead, one data point was randomly excluded from each site. The model was fit to the remaining data, and the excluded points were used as a small test set. This procedure was performed 14 times, such that all data points but the first year served as an out-of-sample point at most once. The first year could not be used for out-of sample prediction since the network models use data from the previous time point to make predictions. Following this procedure, 14 separate models were fit for each method, so that a distribution of model performance could be obtained.

5.3.3 Static and Dynamic Bayesian Networks

A Bayesian Network is a probabilistic model that uses a graphical structure to represent relationships between the random variables. The structure $\mathcal{G}$ contains vertices $\mathcal{V}$ and directed edges $\mathcal{E}$. The vertices (nodes) represent the random variables of interest, in this study, presence/absence of species. Edges (links) represent conditional relationships from one node (the "parent") to another (the "child"); put more precisely, the absence of an edge between two nodes represents a *conditional independency* between the two nodes. The probability distribution of a child node C given parents $\boldsymbol{P}$ is conditionally independent of all other

vertices in $\mathcal{G}$; that is, $P(C|\mathcal{V}) = P(C|\boldsymbol{P})$.

These conditional independencies allow the joint probability distribution of $\mathcal{G}$ to be given in terms of simple conditional probabilities. The joint probability distribution of $\mathcal{G}$ for random variables $X_1, X_2, \ldots, X_n$, with parents $\boldsymbol{P}_i$, is given by

$$P\left(X_1, X_2, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i|\boldsymbol{P}_i\right) \tag{5.1}$$

Bayesian networks of this kind, which do not have a time component, are known as *Static Bayesian Networks* (SBNs). SBNs are limited by a statistical constraint: the network must be a directed acyclic graph (DAG); that is, it may not contain directed cycles. This means that if species A depends on species B, B may not depend on A. This does not conform to ecological reality at all. For example, since predators gain biomass by consuming prey, and prey populations are reduced by predation, we expect predators and prey to mutually influence each other, as is reflected in dynamical ecological models such as Lotka-Volterra [239].

We can create more ecologically reasonable networks by "unfolding" the network in time, creating a *Dynamic Bayesian Network* (DBN) [112]. Using this framework, the presence/absence of a child species at time $t+1$ depends on the presence or absence of its parent species at time t. Put more precisely, the probability that the child species is present at time $t+1$ is conditionally independent of all non-parent species, given the presence/absence of the parent species at the previous time step t. Since parents always predict the child in the next time step, any species may be a parent for any other without violating the DAG constraint. A species may also be its own parent, capturing the effects of autocorrelation. For a DBN, the joint probability of $\mathcal{G}$ may be given as follows:

$$P\left(X_{1,t+1}, X_{2,t+1}, \ldots, X_{n,t+1}\right) = \prod_{i=1}^{n} P\left(X_{i,t+1}|\boldsymbol{P}_{i,t}\right) \tag{5.2}$$

where $X_{i,t+1}$ is node X_i at time $t+1$, and $\boldsymbol{P}_{i,t}$ is the set of parents $\boldsymbol{P}_i$ at time t. For notational simplicity, we drop the subscript for time in future equations. The state of the parents is always given for time t, and the child or random variable is always given for time $t+1$. For an example of how an empirical network may be represented by Static and Dynamic Bayesian Networks, see Box 1.

Box 1. Empirical networks Represented as SBNs and DBNs.

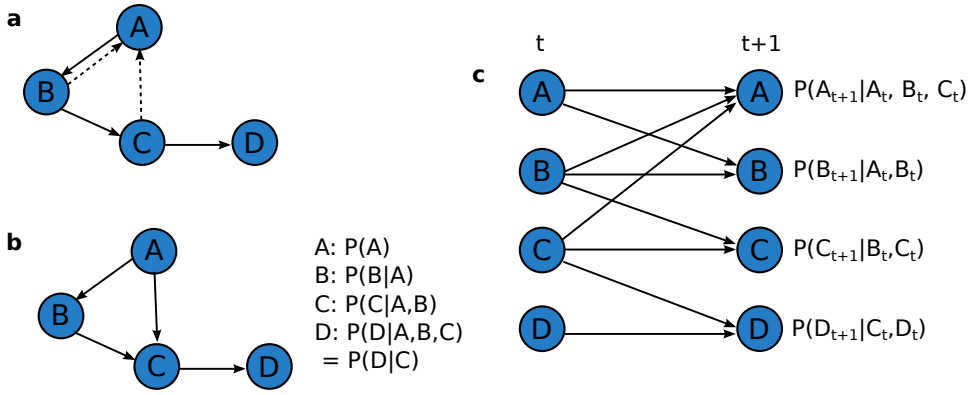

Figure 5.1: **Example of Static and Dynamic Bayesian Network structures.** (a) shows an example ecological network of dependencies between four example nodes, which contains directed cycles. Removal of the dashed edges would create a valid DAG. (b) shows a valid DAG that could be used as an SBN structure for these nodes. Because of the DAG constraint, the structure does not exactly match the ecological network. The local probability distributions that define the SBN are shown to the right. (c) shows a valid DBN network structure. The edges in the network exactly match those of the true ecological network, but are "unfolded" in time. Nodes at time t are on the left, nodes at time $t+1$ on the right. Self-dependencies have been included to capture autocorrelation. Local probability distributions that define the DBN are given to the right of each node.

Consider the network shown in Fig. 5.1a. This network contains two directed cycles: $A \rightarrow B \rightarrow A$ and $A \rightarrow B \rightarrow C \rightarrow A$. Since this is not a DAG, we cannot capture the true structure of this network with an SBN. By removing the $B \rightarrow A$ link, and flipping the direction of the $C \rightarrow A$ link, we get the DAG in Fig. 5.1b. If we consider the conditional independences of the nodes in the order A, B, C, D, we can show that Eqn. 5.1 holds for this structure (all other orderings are also valid, but less convenient).

To do so, we use the fact that for two independent variables X and Y, $P(X) = P(X|Y)$. We can use this property in combination with the conditional independence structure given in the Bayesian network (for example, in Fig. 5.1b, species B depends on A, but is independent of C and D). Then, taking the product of the local distribution functions for each node, we see that

$$\begin{aligned} & P(A)P(B|A)P(C|A,B)P(D|C) \\ =& P(A,B)P(C|A,B)P(D|C) \\ =& P(A,B,C)P(D|C) \\ =& P(A,B,C)P(D|A,B,C) \\ =& P(A,B,C,D) \end{aligned} \tag{5.3}$$

It is convenient to specify the joint probability function over the graph $\mathcal{G}$ using of these comparatively simple conditional probabilities. The DAG constraint ensures that it is possible to calculate the joint probabilities from the conditional probabilities in this way.

Using the DBN structure given in Fig. 5.1c, we can express the true network structure (Fig. 5.1a). Since all local distribution functions for nodes at time $t+1$ are conditioned on the parents at time t, all nodes at time $t+1$ are conditionally independent of each other (*i.e.*, $P(X_{t+1}) = P(X_{t+1}|Y_{t+1})$ for any species X and Y). Therefore, Eqn. 5.1 still holds:

$$
\begin{aligned}
& P(A_{t+1}|A_t, B_t, C_t)P(B_{t+1}|A_t, B_t)P(C_{t+1}|A_t, C_t)P(D_{t+1}|C_t, D_t) \\
=& P(A_{t+1}|A_t, B_t, C_t, D_t)\cdot \\
& P(B_{t+1}|A_t, B_t, C_t, D_t, A_{t+1})\cdot \\
& P(C_{t+1}|A_t, B_t, C_t, D_t, A_{t+1}, B_{t+1})\cdot \\
& P(D_{t+1}|A_t, B_t, C_t, D_t, A_{t+1}, B_{t+1}, C_{t+1}) \\
=& P(A_{t+1}, B_{t+1}, C_{t+1}, D_{t+1}|A_t, B_t, C_t, D_t)
\end{aligned}
\tag{5.4}
$$

Structure Learning

The model as described above assumes that the structure $\mathcal{G}$ is known. Since we are trying to infer a network structure from our data, we also need to learn the structure of $\mathcal{G}$. In theory, this can be done in a fully Bayesian way, by calculating a posterior distribution over both the parameter and network structure space [182]. In practice, this approach is intractable, and it is necessary to use a search algorithm to identify high-quality network structures (which are not guaranteed to be optimal), and evaluate them using a score proportional to the posterior probability of the structure given the data. We use the commonly-used scoring metric Bayesian Dirichlet equivalence (BDe) [108]. We used the software Banjo [228] for structure learning, using simulated annealing as a search algorithm, with 100 runs for 8 hours each, for each dataset. For the Tatoosh dataset, this includes 100 8-hour runs for each of the 14 subsampled datasets. The best network from each run was saved.

Due to the computational complexity of the problem, structure learning in Banjo requires that a maximum number of parents per child node be set. As a check on the sensitivity of the results to this constraint, the structure learning process and all subsequent analyses were

repeated for 2, 3, 4, and 5 maximum parents.

Consensus networks were built from the saved networks. For the France dataset, a majority rule consensus network was built using the top 10 networks. The consensus network was used both for predicting out-of-sample data and for comparison to the empirical network structure. The process was slightly different for the Tatoosh dataset, because of the multiple subsampled datasets. For each subsampled dataset, a majority rule consensus network was built using the top 10 networks. These consensus networks were used to predict out-of-sample data points. To create a single network for comparison to the empirical network structure, a majority rule consensus network was built from the 14 subsampled consensus networks, resulting in a "consensus-consensus" network. Search algorithm convergence was examined using the percentage of links that differed between the consensus networks and each network used to compute the consensus. The percentage difference d between two networks A and B was calculated as

$$d = \frac{(A \backslash B)^L + (B \backslash A)^L}{A^L + B^L} \tag{5.5}$$

where A^L is the number of links in network A, $A \backslash B$ is the relative complement of B with respect to A (the links in A but not B), and $(A \backslash B)^L$ is the number of links in $A \backslash B$. For the France dataset, d was calculated between each consensus network and the top 10 networks used to create it. Similarly, for the Tatoosh dataset, d was calculated between each consensus network and the top 10 networks used to create it, for each of the 14 data subsamples.The distribution of d provides a sense of how similar top solutions are, and by extension, how well the search algorithm is converging on optimal or close-to-optimal solutions.

Posterior Calculation and Prediction

After inferring a structure $\mathcal{G}$ as described above, we calculated the joint posterior distribution of the DBN model given $\mathcal{G}$. To calculate the joint posterior distribution, we can use the local

distribution functions $P(X_i|\boldsymbol{P}_i, \mathcal{G})$ for each Bernoulli random variable X_i. Because each parent may take on one of two states (0, "absent", or 1, "present"), there are 2^{k_i} possible configurations of the parents, where k_i is the number of parents for species i. Thus, each local distribution is a collection of binomial distributions, one for each parent configuration $\boldsymbol{P}_i = \boldsymbol{p}_{ij}$ [107].

We can analytically calculate a posterior distribution given these local distribution functions with two additional assumptions. The first is that the data in our random sample D is complete (that is, no data are missing), which is true in both datasets we considered. The second is that all parameters θ_{ij} are mutually independent, where θ_{ij} is the probability species i is present in a Bernoulli trial given parent configuration $\boldsymbol{p}_{ij}$. The assumption of parameter independence is demonstrated for multinomial local distribution functions in Heckerman 2008 [107], which also applies to the more restricted case of binomial local distribution functions.

For any node i and parent state $\boldsymbol{p}_{ij}$, the likelihood of the observed data is given by the likelihood function for binomial sampling:

$$\mathcal{L}\left(\theta_{ij}\right) \propto \theta_{ij}^{r_{ij}}(1-\theta_{ij})^{n_{ij}-r_{ij}} \tag{5.6}$$

where r_{ij} is the number of times X_i was present in the data given state $\boldsymbol{p}_{ij}$ in the previous time step, and n_{ij} is the total number of data points with the parent state $\boldsymbol{p}_{ij}$ in the previous time step.

To make this analytically tractable, we can choose a Beta prior on the parameter θ_{ij}, such that

$$P\left(\theta_{ij}|\mathcal{G}\right) = \frac{\Gamma(\alpha_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}\theta_{ij}^{\alpha_{ij}-1}(1-\theta_{ij})^{\beta_{ij}-1} \tag{5.7}$$

where α_{ij} and β_{ij} are hyperparameters which determine the shape of the prior distribution.

We choose $\alpha_{ij} = 1$ and $\beta_{ij} = 1$ for all nodes i and configurations j. This produces a flat prior, which corresponds to a level of confidence equivalent to observing X_i to be present once and absent once. Since the Beta distribution is a conjugate prior to the binomial likelihood function, the posterior distributions are simply given by

$$P\left(\theta_{ij}|D, \boldsymbol{p}_{ij}, \mathcal{G}\right) \sim Beta\left(\theta_{ij}|1 + r_{ij}, 1 + n_{ij} - r_{ij}\right) \tag{5.8}$$

Having calculated the posterior distributions, we can then use them to make predictions about the state of the system at time $t + 1$, given the state at time t. Using data from time step t to establish the state of the parents, the expected probability that species i is present at time $t + 1$ is given by the expectation of the Beta distribution for the parent state $\boldsymbol{p}_{ij}$:

$$P(X_i = 1|D, \boldsymbol{p}_{ij}, \mathcal{G}) = \frac{1 + r_{ij}}{2 + n_{ij}} \tag{5.9}$$

Repeating this process for all species, we can predict the state of all species in the next time step. In this way, we can compare the predicted state to the observed state at each timepoint in the out-of-sample (test) data, in order to evaluate the short-term predictive power of the model.

5.3.4 Lasso Regression

Lasso regression is a generalized linear regression approach with an additional regularization parameter λ [247]. This regularization parameter acts to shrink or eliminate predictor coefficients that contribute least to the model fit, helping to prevent overfitting. The line of best fit is calculated based on the following equation:

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = argmin \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ where } \sum_j \left|\beta_j\right| \leq \lambda \tag{5.10}$$

where $\hat{\alpha}$ is the intercept for the line of best fit, $\hat{\boldsymbol{\beta}}$ is the vector of coefficients for the line of best fit, N is the number of data points in the random sample, y_i is the observed value of the dependent variable for the ith data point, and x_{ij} is the observed value of the jth predictor for the ith data point. As in a standard generalized linear model, we are minimizing the sum of squared errors, but with an additional constraint on the total magnitude of the βs. This causes the βs for relatively unimportant predictors to shrink to 0. We used this shrinkage property to construct a network. By running Lasso regression on each species, we identified the interacting species with the strongest predictive effect, and created a network from those interactions. Lasso was chosen over other regularization approaches (such as elastic net) due to its popularity for biological network inference [174, 82, 159, 138].

A separate regression was performed for each species, where the presence/absence of one species was regressed on the presence/absence of all species (including itself) at the previous time step; that is, the lag 1 of all species. Predictor species with non-zero coefficients were kept as interactions in the resulting network. We used k-fold cross-validation with the training set to identify the optimal value of λ. Using this value, we used the trained model to predict the test set. We trained two Lasso regression models: one containing only first-order (*i.e.*, pairwise) interactions, and one containing both first and second-order interactions. For the model containing second-order interactions, any pair of species included in a second-order interaction were considered to be interacting, for purposes of constructing the network.

The DBN model is a Bayesian approach, and therefore does not have a specific amount of data required to train a model. However, logistic regression models, similar to the lasso regression used here, can produce unstable results when there are too few observations in each category [110]. To avoid this issue, we excluded species with fewer than 20 observations in each class from the analysis (species which were chronically present or chronically absent). No species from the France dataset were excluded, but 8 species were excluded from the Tatoosh dataset for this reason: *Semibalanus cariosus*, *Leathesia*, *Prionitis*, *Lottia pelta*,

Lottia digitalis, *Tonicella lineata*, *Pinnotheres pisum*, and *Henricia*. These species were also excluded from the Tatoosh empirical networks when comparing them to the inferred Lasso networks.

5.3.5 Pearson's Correlation Coefficient Method

We calculated the Pearson's correlation coeffient for each species with the presence/absence of each other species in the previous time step (that is, we calculated the correlation between each species and the lag 1 of each other species). We then performed a permutation test with 10,000 repetitions to determine the significance of each correlation. In each permutation, the observations for each species were shuffled, such that the ratio of presences to absences was preserved for each species, but were randomized with respect to time and site. A null distribution of correlations was estimated based on correlations calculated for the permuted data. Correlations which were significantly larger in magnitude than expected by chance, for an overall false discovery rate of .05 (using the Benjamini-Hochberg FDR correction [26]), were kept in the network as interactions.

5.3.6 Null Models

Two null models were used to assess comparative fit of the network models. The first null model was a weighted coin flip for each species. That is, the presence or absence of a species at any time point was predicted by the probability of presence in the training set. This null model captured weighted presence/absence probabilities, but ignored the time component of the data.

The second null model was a disconnected DBN, where the probability that a species was present was independent of all others, conditional only on its own presence or absence in the previous time step. Essentially, the DBN structure was given by a network where each species connects only to itself. The parameters of this model were calculated as for the full

DBN model. This null model captures the autocorrelation of the data, but no interactions between species.

5.3.7 Model Comparison

The similarity between empirical networks and model networks was measured using two standard machine learning measures, precision and recall. Precision measures the fraction of links identified by the model which are present in the empirical web, while recall measures the fraction of empirical links which are correctly identified by the model:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{5.11}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{5.12}$$

If most of the interactions identified by a model are present in the empirical web, the model will have high precision. If a model is able to identify most of the interactions in the empirical web, it will have high recall. The ideal model would have both precision and recall near 1, but there is often a tradeoff between the two. For example, a model which predicts that all pairs of species interact would have high recall (it would identify all empirically present interactions), but low precision (it would falsely predict many empirically absent interactions). Conversely, a model which predicted only a single true positive would have high precision (all of its positives would be true positives), but low recall (it would fail to identify most empirically present interactions).

The significance of the precision and recall values was calculated using a permutation test. For each permutation, the model network was shuffled such that each species had the same number of incoming links, but the identity of those species was randomized. This constraint was added to replicate the DBN constraint on the number of parents, and the soft

constraint of regularization in the Lasso model. A side effect of this permutation procedure is that although the number of true positives will vary by randomization, the denominators of the precision and recall formulas will not. Put another way, the precision and recall values will be perfectly correlated, such that the p-values for precision and recall of a given model will always be identical. Before calculating precision and recall, all interactions between a species and itself were excluded from both empirical and inferred networks. This is because these links represent cannibalism or self-regulation in the empirical networks, but represent autocorrelation in the inferred networks.

To compare model predictive performance, fit models were used to predict out of sample data. We calculated the log likelihood of each model correctly predicting the entire test dataset. For the Tatoosh dataset, separate models were fit to 14 cross-validation sets, resulting in 14 log likelihoods per model. The likelihood distributions of the models was compared to that of the disconnected DBN null model using a Wilcoxon rank sum test, with p-values corrected for a false discovery rate of .05 across all models.

5.4 Results

5.4.1 DBN Convergence

Individual runs were relatively similar to the overall consensus networks, in both the France and Tatoosh datasets, suggesting that the algorithm was finding similar high-quality solutions in different runs. In the France dataset, d ranged from 1.05% to 10.47%, with an overall median difference of 4.46%, and a slight decrease in d as the maximum number of parents increased ($median_2 = 6.50\%$, $median_3 = 5.26\%$, $median_4 = 2.90\%$, $median_5 = 2.88\%$). In the Tatoosh dataset, d ranged from 0.00% to 11.20%, with an overall median difference of 4.67%. Median difference increased slightly with the maximum number of parents ($median_2 = 3.14\%$, $median_3 = 4.53\%$, $median_4 = 5.61\%$, $median_5 = 5.49\%$).

5.4.2 Network Structures

All models inferred networks which were significantly related to empirical networks for one of the two datasets, but no model successfully inferred empirical networks for both. In the France dataset, only the networks inferred by Lasso models had higher precision and recall of the true food web than expected by chance (Table 5.2). The 2-parent DBN model was marginally significant ($p = .06$). The Pearson's correlation coefficient network had very high recall (.90) but relatively low precision (.19), suggesting that although the correlation network had many interactions (connectance = .85), the interactions had little relation to the food web structure.

All model networks for the Tatoosh trophic web performed poorly. No models produced a network which was significantly similar to the Tatoosh food web structure. Interestingly, the inferred networks were much more similar to the network of nontrophic interactions (Table 5.2, Fig. 5.3). The 2-parent DBN and Pearson networks were more similar to the nontrophic network than random ($p = .02$ for DBN-2, $p < .01$ for Pearson). The first-order Lasso ($p = .10$) and all other DBN networks ($p = .06$ for DBN-3, 4, and 5) were marginally significant (Table S1).

5.4.3 Predictive Accuracy

Both DBN and Lasso models predicted out-of-sample time points more accurately than null models. The disconnected DBN null model performed dramatically better than the coin flip null model in both datasets, suggesting that autocorrelation accounted for much of the trend in the data. In the France dataset, Lasso models performed better than null models, and similarly to each other. DBN models performed better than both Lasso and null models (Table 5.1). Due to the limited size of the Tatoosh dataset, models were trained on 14 data subsamples, each of which was used to predict the remaining out-of-sample points. This resulted in a distribution of likelihood values, rather than a single number (Fig. 5.2,

Table S2). Although there was some overlap between the likelihood distributions (with the exception of the coin flip null), Wilcoxon signed-rank tests between the models and the disconnected null found some meaningful differences. The DBN models all performed better than the disconnected model ($p = .003$, $.003$, $< .001$, and $.001$ for 2,3,4, and 5 max parents, respectively). In contrast, neither Lasso model had a median that significantly differed from the disconnected null ($p = .162$ and $.426$ for the first and second-order models, respectively).

The predictive accuracy of DBN models did not appear to be limited by the constraint on the number of parents. In the France dataset, the DBN model with 2 parents performed worse than the others, but the model with 3 parents had the best predictive accuracy overall (Table 5.1). The 2 parent DBN performed best of all the Tatoosh models (Fig. 5.2).

5.5 Discussion

Using three methods taken from microbial and gene network inference, we found mixed success for all methods. DBN-2 and Pearson significantly captured the nontrophic structure of the Tatoosh network, but only the Lasso regression variants significantly captured the French piscivory network. This stands in contrast to Milns *et al.* [174], who found that Bayesian networks strongly outperformed Lasso regression in inferring networks for avian communities and habitat features. It is worth noting that many models are marginally significant for both the France and Tatoosh nontrophic networks. Although we should be cautious not to overstate marginally significant results, this does point to some level of generality in the performance of the models. The three methods are uneven in performance between the two datasets, but DBN and Lasso perform moderately well on both. A comparison of methods for gene network inference also found a good deal of variability in performance across datasets [159], and no model was the best performer on all datasets. This resulted from the fact that each model had specific biases, which had more or less effect, depending on the specifics of the dataset. The most robust approach was to combine the results of multiple

methods [159]. This is a promising future avenue of research in ecological network inference.

Interestingly, some models capture the trophic structure of the France piscivory network, but no models capture the trophic structure of Tatoosh, although they perform relatively well on the nontrophic structure. Unfortunately, not enough information was available to construct a reliable nontrophic network for the France system, so we can only speculate if these methods are better at capturing nontrophic information in general, or if there is something specific to the Tatoosh system or dataset. Co-occurrence networks are generally used to infer competitive interactions specifically (*e.g.*, [59, 50, 128]). Because we have dynamical data on the dynamics of the system through time, we might naively expect to identify the effects of both trophic and nontrophic interactions, and at least in the French network, this was the case. However, it is possible that presence-absence data is too blunt an instrument to capture predator-prey dynamics in some systems. In a simple Lotka-Volterra predator-prey interaction, neither species is expected to go extinct, so if the sampling effort is high enough, both species would simply show up as present for long periods of time. In contrast, competitive exclusion by definition results in the local extinction of one species, and therefore might be easier to capture in this type of data. In this way, the detection threshold of the data may influence the interactions the model can infer.

The many differences between these datasets allow us to see the variability in performance from the network inference algorithms, but make it difficult to identify specific factors underlying those differences. However, one likely factor is the type of competition dominant in each of these systems. In the Tatoosh system, competition for physical space strongly structures the system [196]. Most sessile species are either filter feeders or photosynthetic, and as a result, are directly competing for space on the rock or other individuals, rather than competing for a limited food resource. In this zone of the intertidal, bare rock is generally revealed only in response to disturbance, and is quickly colonized [196]. Competition for space in the Tatoosh middle intertidal is so strong that competitive interactions may drown

out the signal of feeding interactions (but see [271]), at least when the data are limited to presence-absence. In contrast, physical space in rivers is less limiting, so feeding interactions and competition for shared prey are more likely to be strong drivers in the French piscivory network.

Despite their mixed success for predicting network structures, the DBN and Lasso models were relatively good at predicting short term dynamics. The DBN models were consistently predictive in both the France and Tatoosh systems, and although the Lasso models were not as predictive of the Tatoosh system, they performed well on the France dataset. DBNs and Lasso are fundamentally predictive models, with the network structures produced more as a side effect than an end in itself. These models are clearly effectively finding predictive patterns in the data; however, these patterns may not always correspond to ecological interactions. Even in the model networks which corresponded well to the empirical networks, there were many false positives and false negatives. False negatives are not especially surprising, since empirical networks are likely to contain relatively unimportant interactions which may not have much influence on the community dynamics. The false positives demonstrate that species were predicting each other's dynamics without directly interacting. Some of these predicted interactions likely result from indirect effects: causal effects that occur without direct interaction between the two species. Indirect effects are known to drive dynamics in the Tatoosh intertidal (*e.g.*, [266, 267, 268, 271]). False positives might also occur when two species share a common driver. In the lagged models used in this study, species which share a predator, resource or environmental stressor might be inferred to "interact" if they respond at different time scales. The species which responds at close to a lag-1 time scale would, for example, predict a species which responds at close to a lag-2 time scale. It is possible that using a different lag, or including multiple lags, could improve the model, although previous work on the mussel *Mytilus californianus* in the Tatoosh intertidal found that 1-year lag models outperformed models with all lags up to four years [274].

Model	Log Likelihood	Median Probability
Coin	-26205.80	0.71
Disconnected	-11447.79	0.92
DBN-2	-10653.97	0.94
DBN-3	-10437.39	0.95
DBN-4	-10476.15	0.95
DBN-5	-10474.91	0.95
Lasso-1st	-11098.84	0.93
Lasso-2nd	-11017.07	0.94

Table 5.1: **Out-of-sample predictive accuracy for the French stream dataset.** Columns represent the model name, log likelihood of correctly predicting test data, and median probability of correctly predicting a single species at a single out-of-sample time point.

Although none of these methods was consistently successful, they are worth exploring further, both for different systems and different types of data (for example, abundance rather than presence-absence data). No method is a strictly significant predictor of network structure for both systems, but it is clear that methods like these are capable of capturing some dynamical signal of the underlying interactions. This suggests that they could reasonably be used in microbial systems, but only with great caution; for these two datasets, even the model networks which significantly match empirical structures had relatively low precision and recall. In the absence of known microbial networks [81], data for macroscopic communities can provide a useful proxy for validating the use and interpretation of network inference methods. As machine learning and correlational methods continue to grow in popularity, it is vital that we validate their results, and consider what we can and cannot learn from the networks they produce.

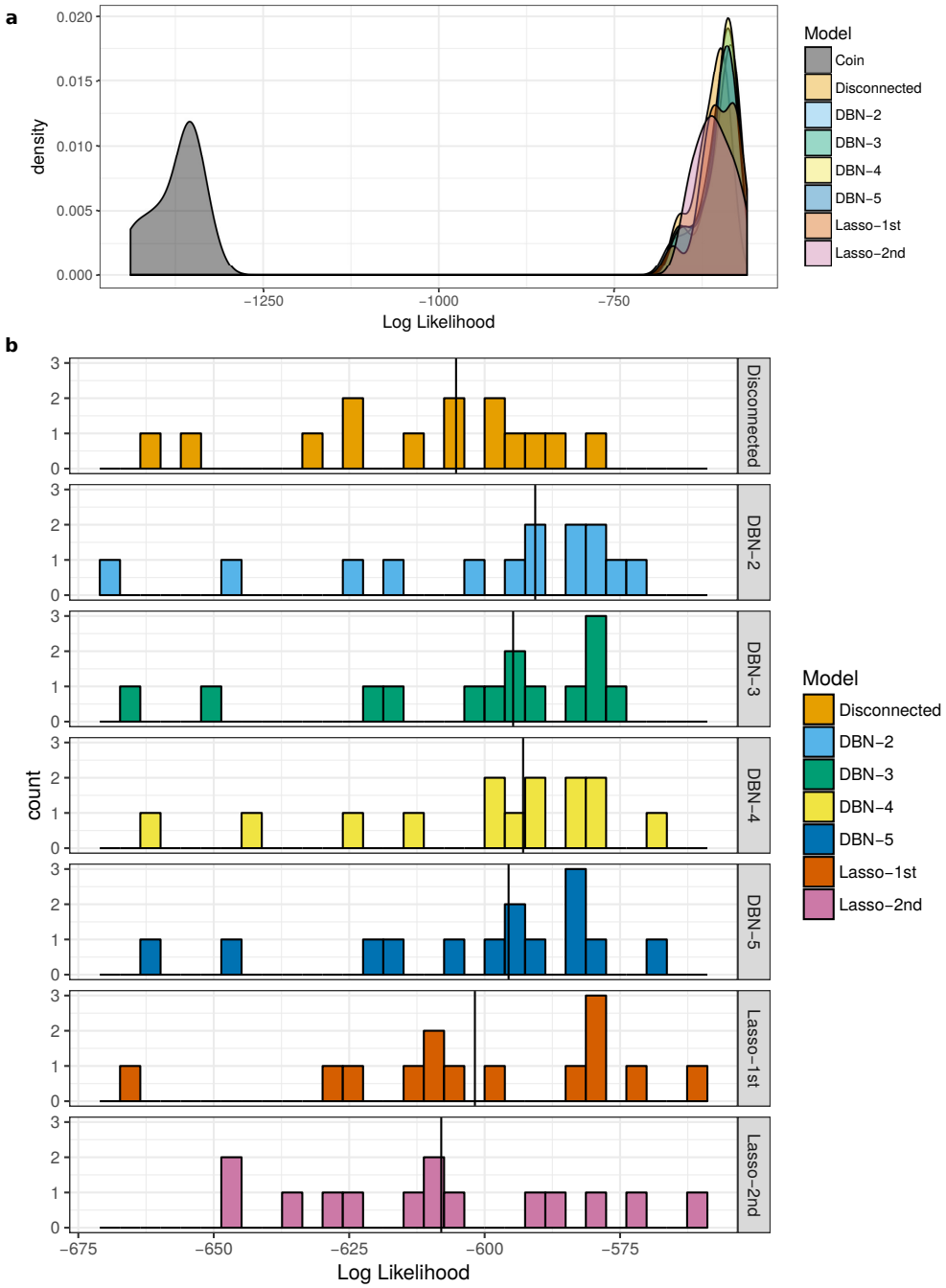

Figure 5.2: **Log likelihood distributions for Tatoosh data subsamples, (a) including the weighted coin flip null model, and (b) excluding the weighted coin flip, and zooming in on the remaining models.** Each observation represents the log likelihood of correctly predicting out-of-sample test data, given the model fit on one subsampled dataset. Black vertical lines in (b) represent the median log likelihoods for each model.

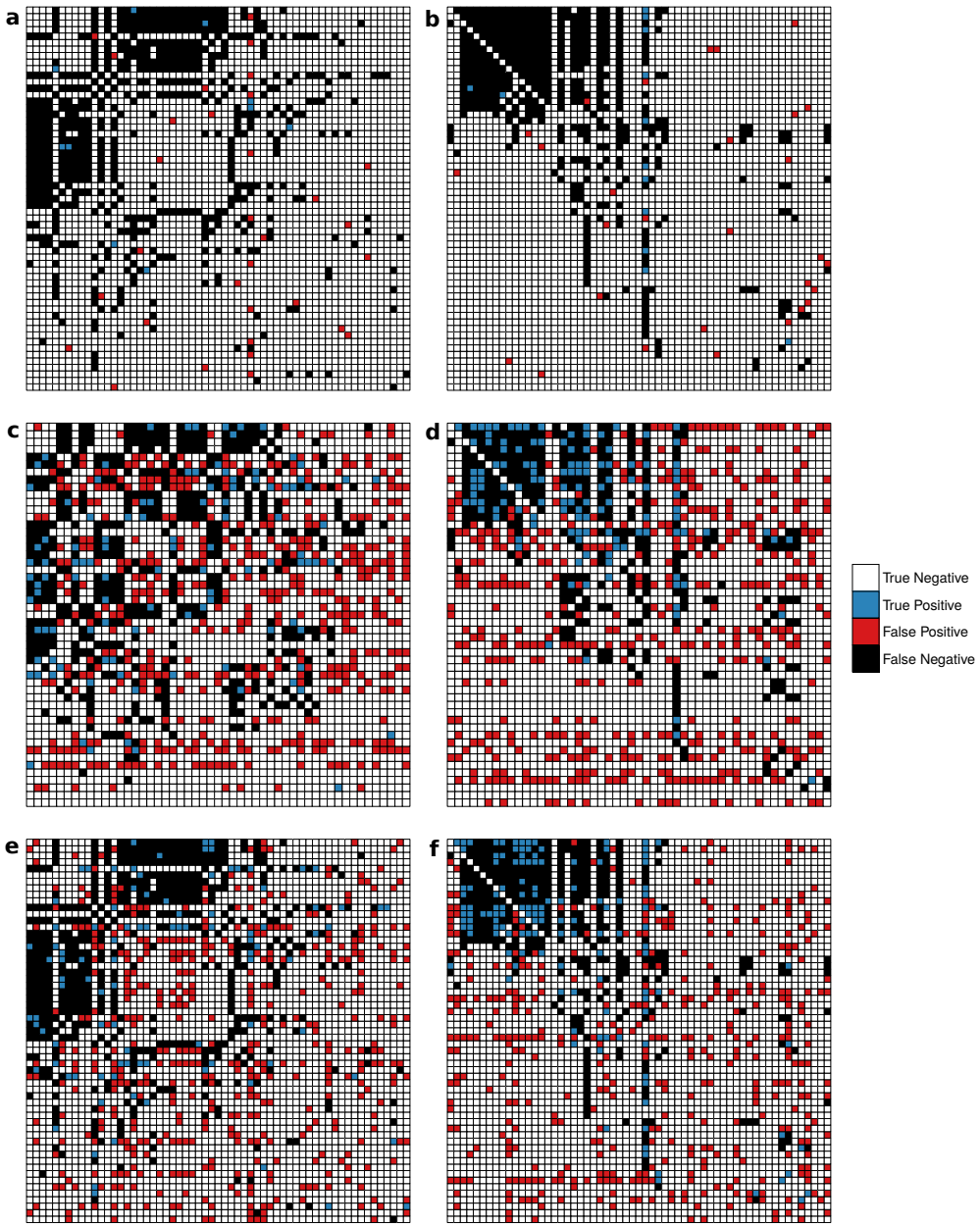

Figure 5.3: **Structural comparison between empirical and model adjacency matrices for the Tatoosh system: (a) Trophic and DBN-2, (b) Nontrophic and DBN-2, (c) Trophic and Lasso-1st, (d) Nontrophic and Lasso-1st, (e) Trophic and Pearson, and (f) Nontrophic and Pearson.** White boxes represent the absence of a link in both networks (true negative), blue represents a link present in both networks (true positive), black, a link present in the empirical but not model network (false negative), and red, a link present in the model but not empirical network (false positive).

	France		Tatoosh Trophic		Tatoosh Nontrophic	
Model	Precision	Recall	Precision	Recall	Precision	Recall
DBN-2	0.33	0.060	0.24	0.016	0.31*	0.027*
DBN-3	0.20	0.065	0.15	0.013	0.27	0.029
DBN-4	0.22	0.083	0.18	0.014	0.27	0.029
DBN-5	0.23	0.089	0.20	0.016	0.27	0.029
Lasso-1st	0.24*	0.571*	0.19	0.193	0.22	0.297
Lasso-2nd	0.23**	0.685**	0.21	0.415	0.20	0.533
Pearson	0.19	0.905	0.18	0.155	0.25***	0.278***

Table 5.2: **Precision and recall of model networks compared to empirical network structures for the French stream food web, Tatoosh trophic network, and Tatoosh nontrophic network.** Precision and recall values are given with significance based on a permutation test (<.05: *, <.01: **, <.001: ***), with p-values corrected across all models and networks for an overall false discovery rate of $\alpha = .05$.

5.6 Acknowledgements

Thanks to G. Barabás, M. Michalska-Smith, C. Pfister, M. Wang, and G. Dwyer for helpful conversations and comments. We thank the Makah Tribal Council for continued access to Tatoosh Island. ELS is supported by the NSF GRFP. JTW and long-term data collection at Tatoosh are supported by the Andrew W. Mellon Foundation, the University of Chicago, and the National Science Foundation (OCE 9711802, OCE 0117801, OCE 0452687, DEB 0919420, OCE 0928232, DEB 1148867, and DEB 1556874). SA is supported by the National Science Foundation (DEB 1148867).

5.7 Author contributions statement

ELS and SA conceived of the project. JTW collected data for the Tatoosh system. ELS wrote and analysed the models, constructed the French stream food web, and wrote the manuscript. All authors edited and reviewed the manuscript.

CHAPTER 6

APPENDIX 1: SUPPLEMENTAL INFORMATION FOR *WHAT CAN INTERACTION WEBS TELL US ABOUT SPECIES ROLES?*

6.1 Supplemental Methods

6.1.1 Supplemental Network Information

In the Norwood farm network [210], indirect interactions (i.e., parasitoids which indirectly benefitted plants by parasitizing herbivores) were excluded from the adjacency matrix. In [210], leaf-miner parasitoids were assumed to be a reasonable proxy for the interactions between plants and leaf-miners. However, since direct leaf-miner network data were not available, we chose to exclude these species altogether. Two species in this system interacted in the network in multiple ways (interacting with some species as a plant mutualist and with others as an herbivore, for example). Each of these species was listed as a single node in the adjacency matrix, but both interaction types were kept. In these cases, when species of one interaction type were removed, only the links of that interaction type were removed.

In the Doñana and Norwood webs, some plant species became disconnected from the large cluster when mutualists or herbivores were removed. However, these species were not excluded from the group model selection process, to make the single-interaction versions of the networks more directly comparable to the complete versions.

6.1.2 Taxonomic Data

Taxonomic data were obtained from the ITIS website at `//www.itis.gov/downloads/index.html` and resolved to the species, genus, or family level, whichever was the most detailed that could be obtained from this database. Taxa which could not be identified through the ITIS

database were identified by hand to the highest resolution possible using information from the Encyclopedia of Life [1], and outdated names were changed by hand where possible. Nodes that were not resolved to at least the phylum (for the Tatoosh network) or order (for the Doñana and Norwood networks) level were excluded from the taxonomy-based grouping and similarity calculations between partitions.

6.1.3 Search Algorithm

The space of possible groupings was explored using Metropolis-coupled Markov Chain Monte Carlo (MC^3), also known as parallel tempering [94, 15]. This search strategy allows multiple MCMC chains at different temperatures to run in parallel, occasionally giving chains the opportunity to swap temperatures. Here we use the term 'temperature' as it is used in simulated annealing, where the probability of accepting a bad move is a function of the temperature. We allowed our chains to take temperatures which ranged from the hottest temperature of slightly above 0 (all proposed steps accepted with equal probability) to the coldest temperature, 1 (proposed steps accepted with probability directly proportional to their likelihood). Every 20 steps, we allowed chains the opportunity to swap temperatures. These swaps proceeded in a "bucket brigade" fashion [74], such that each chain had the opportunity to swap temperatures with adjacent temperatures from hottest to coldest with probability

$$min\left[1, exp\left(\left(\frac{1}{t_i}-\frac{1}{t_{i+1}}\right)\Big(log\left(P(N|G_i)\right)-P(N|G_{i+1})\Big)\right)\right] \tag{6.1}$$

where t_i is the temperature of chain i, and temperatures are arranged such that t_i is hotter than t_{i+1}. In this way, the hottest chain may cool to become the coldest chain in a single round of swaps if it happened to find a promising part of the solution space.

Case studies have suggested that the algorithm performs best when the probability of

accepting a swap is close to 20% [74]. Based on this and our own preliminary study, we calibrated the number of chains for each network to result in a swap acceptance rate between 15 and 35 percent. The swap acceptance probability for a given number of chains tended to decrease with network size, so larger networks were generally given more chains. Temperatures were uniformly spaced, as preliminary testing indicated that alternate spacing schemes (*e.g.*, logarithmic) resulted in very high swap acceptance rates and inferior groupings.

6.1.4 Restricted Growth Function (RGF)

Groupings were stored as a vector of length equal to the number of species S. The ith element of this vector held the group membership of species i. Groups were given a number between 0 and $(S-1)$. After initializing the starting solution for each chain, and after each mutation step, the solution was run through the RGF algorithm. This algorithm renumbers the groups such that as you move from left to right through the vector, new groups are numbered sequentially. For example, the solution [2 3 1 1 4 2] would be represented as [0 1 2 2 3 0] after being run through the RGF. The renumbering does not change the solution, since the group labelling is arbitrary, but it ensures that each solution has a single consistent representation. This algorithm runs in linear time and improves the efficiency of the parallel tempering algorithm, since it spends less time rediscovering solutions that are identical but encoded differently.

6.1.5 Gibbs Sampler

The mutation step for each chain used a Gibbs sampler, which ran as follows: A single species was chosen for mutation. Then, the algorithm calculated the Bayes factor for each potential group assignment for that species, including its current group assignment and assigning it the species its own unique group. The new group assignment was chosen based on probabilities B_i^T, where B_i is the Bayes factor of the ith group, and T is the temperature of the chain.

6.1.6 Jackknife Resampling

An implicit assumption of the group model as used in this study is that the "complete" network structure, containing all available interaction data, is the "true" network structure. That is, the model assumes that there are no missing or erroneous interactions. There are many possible ways to test the robustness of the method to incomplete data. Jackknife resampling was chosen due to the relative simplicity of performing and interpreting the results. Resampling was only performed using the Tatoosh Island network, because this was the only network with fast enough convergence to be computationally feasible. Species were removed one at a time from the complete network, then were partitioned using the group model 25 times, with the algorithm running for 10000 steps with 20 chains each run. The 25 runs converged to between 2 and 6 solutions with similar marginal likelihoods, suggesting that the algorithm was able to adequately explore the solution space. The best grouping for each species removal was compared to the best grouping before species removal. The match between the grouping before and after removal was calculated as $\frac{MI}{MI_{max}}$, that is, the observed MI over the maximum MI possible for the two partitions (MI_{max} is equivalent to the minimum of the two entropies).

6.1.7 Null Model for Partition Comparison

There are multiple possible null models when comparing the observed mutual information to what is expected by chance. An intuitive null model for species grouping would be to assign each species a group identity from 1 to S. However, this allows the null to have more or fewer groups than observed, which changes the number of parameters in the model. A similar problem exists if we choose to assign each species a group identity from 1 to g (the number of groups in the observed partition). The null will have the same number of groups as observed, but will likely have different numbers of species in each group. the number of species in the different groups has a large effect on the entropy, which provides

an upper bound on the MI. Therefore, we choose to randomize the grouping by shuffling the species, such that the number of species in each group remains identical to the observed partition, but the species identities in the groups are randomized. This procedure maintains the entropies of each partition, so that the MI has the same upper bound, but allows us to get a distribution of what MI would be expected by chance, in the absence of the effect of network structure. This is the most conservative null model of the three.

6.1.8 Supplemental Figures and Tables

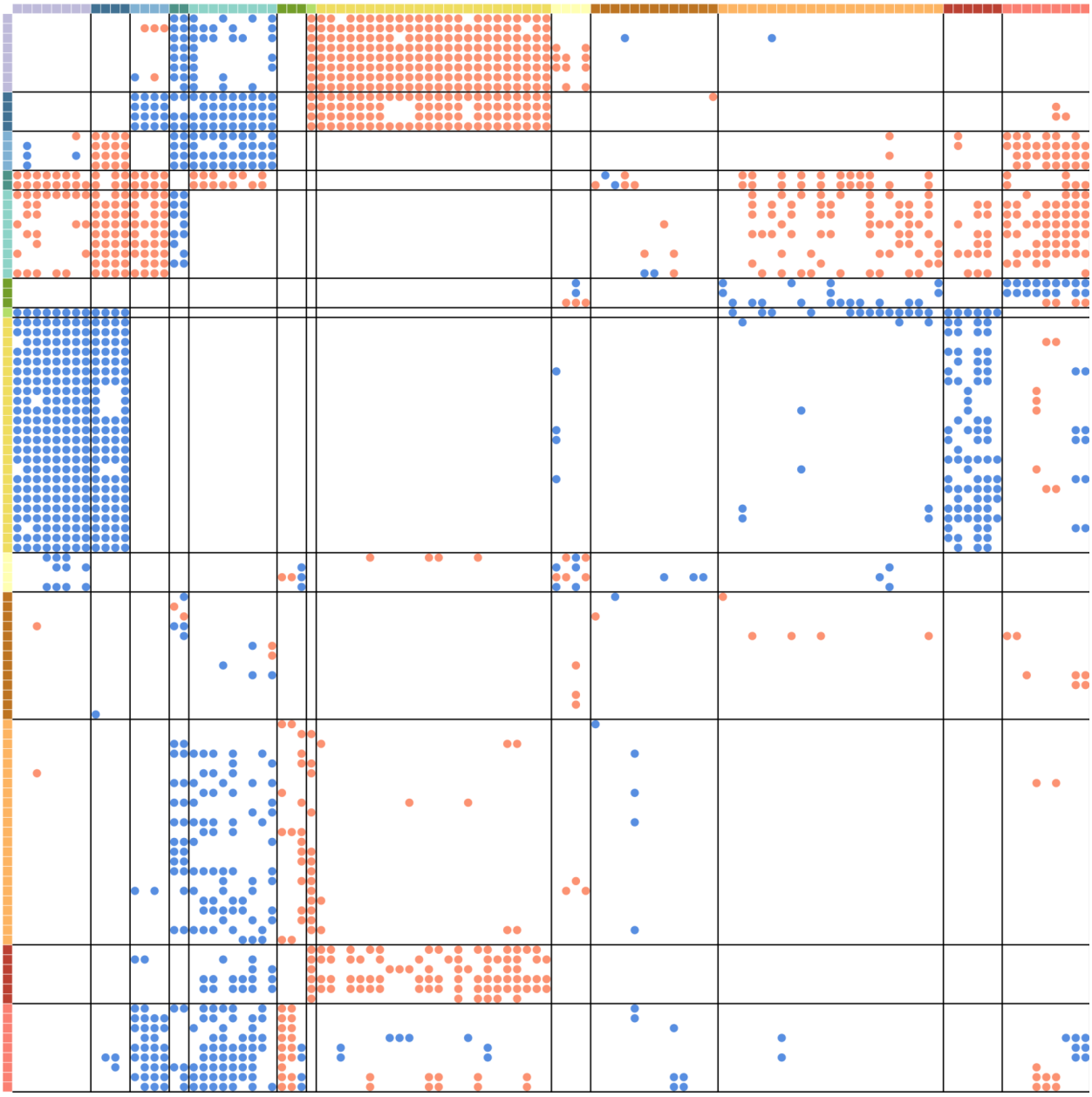

Figure 6.1: **Optimal grouping structure for the Tatoosh Island mussel bed trophic network.** Figure structured as in main text Fig. 2.1. Numbered list of taxon names is given in Table 6.1.

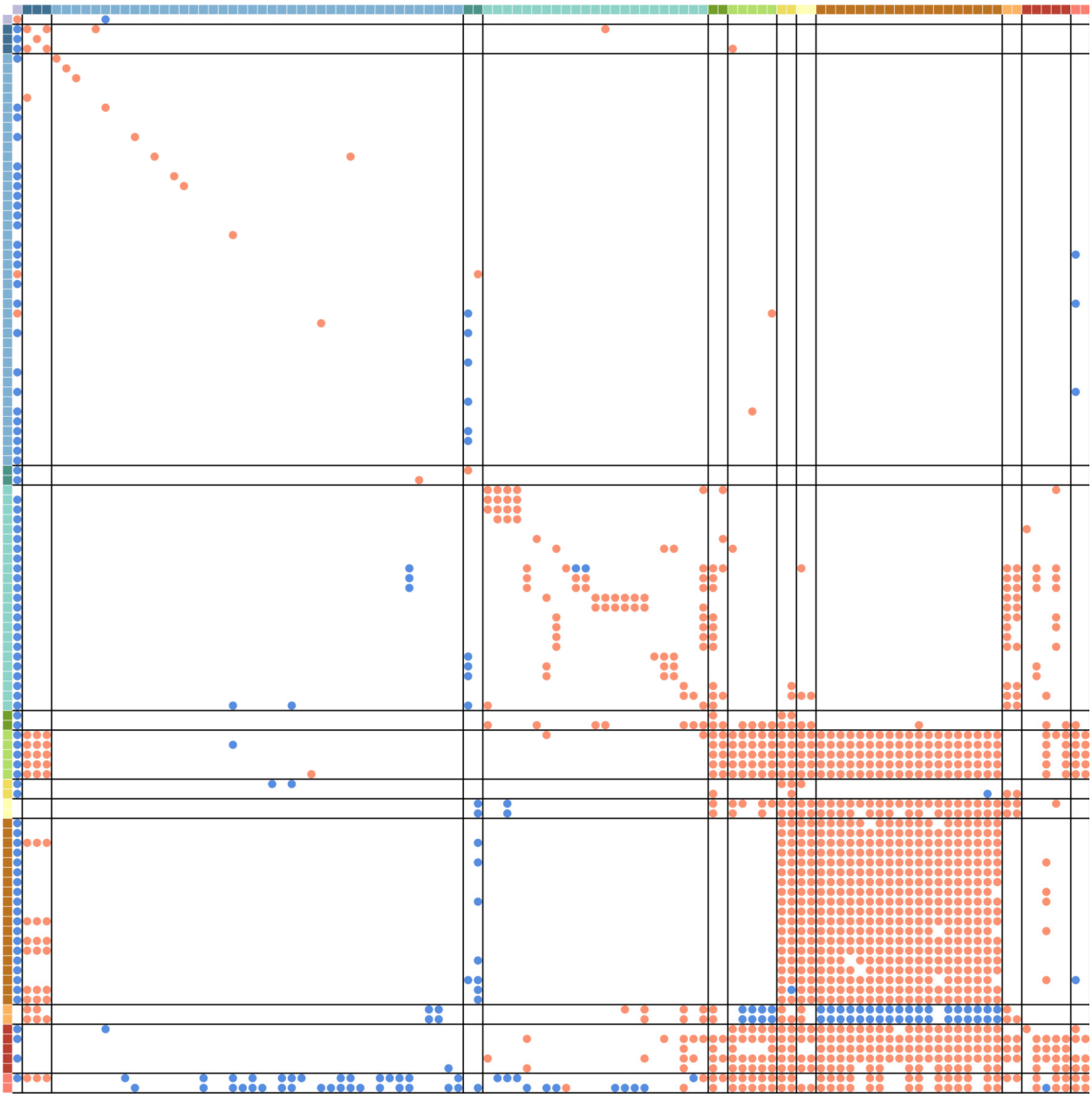

Figure 6.2: **Optimal grouping structure for the Tatoosh Island mussel bed non-trophic network.** Figure structured as in main text Fig. 2.1. Numbered list of taxon names is given in Table 6.1.

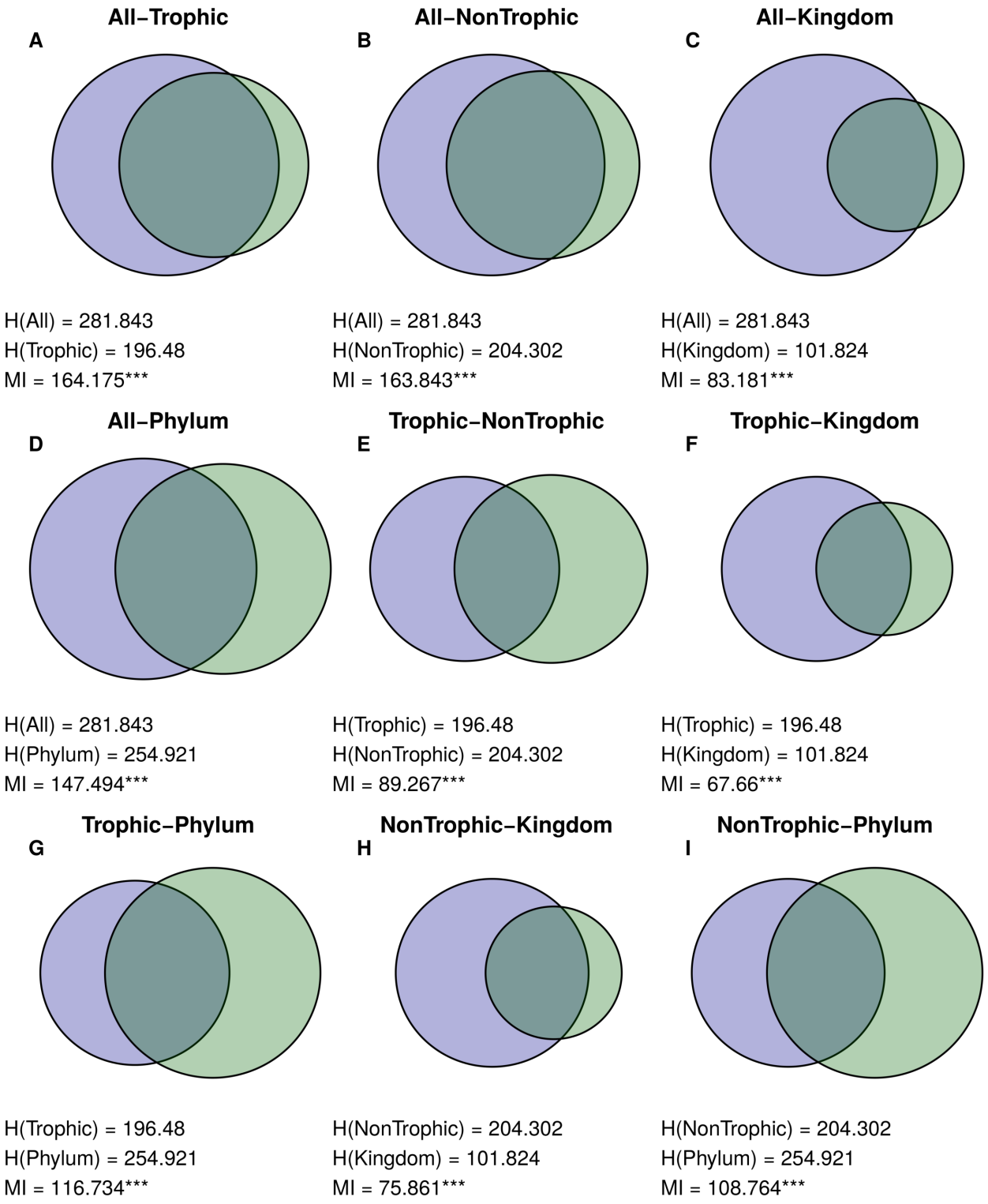

Figure 6.3: **Similarities between Tatoosh Mussel Bed partitions.** Table structured as in main text Fig. 2.4, but containing all pairwise comparisons between partitions for the Tatoosh Mussel Bed.

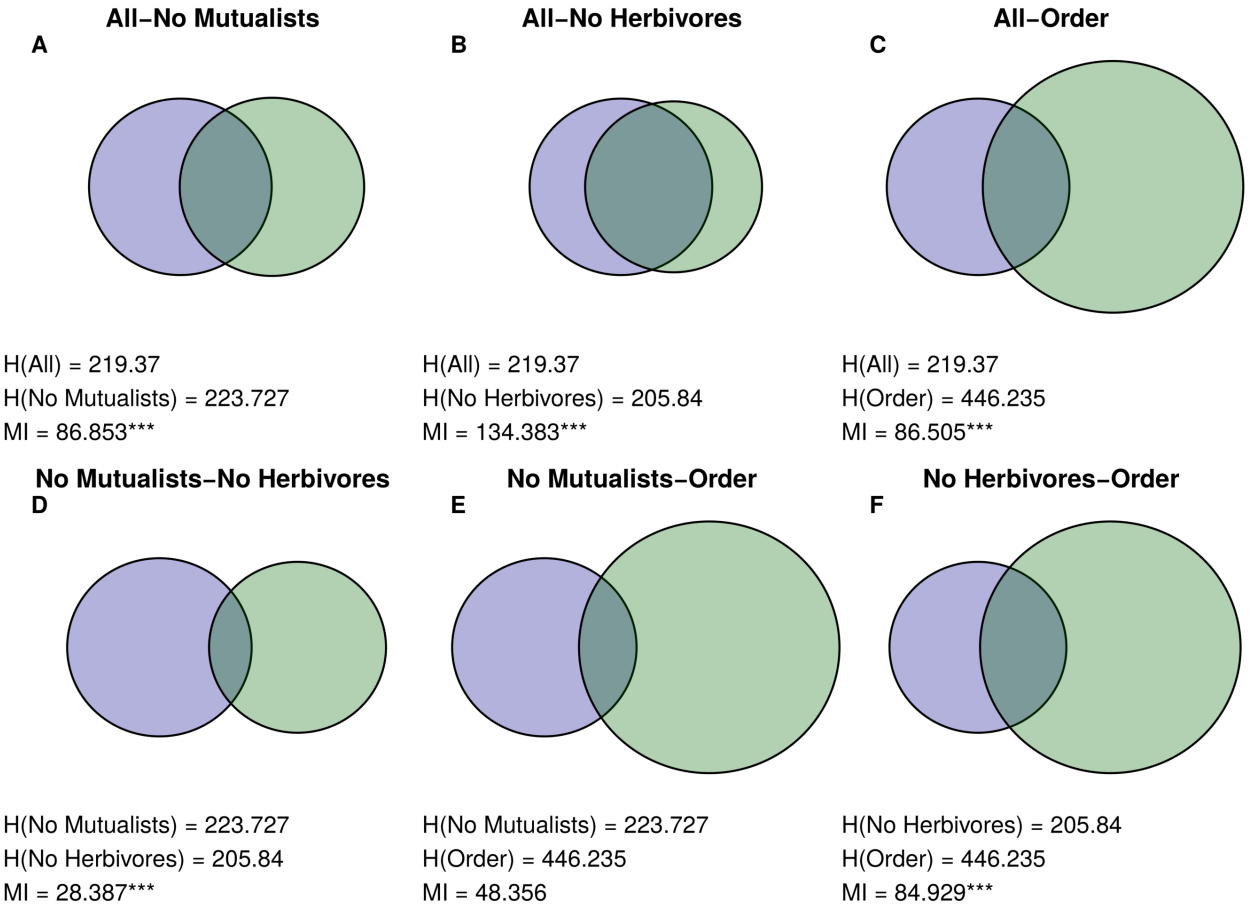

Figure 6.4: **Similarities between Doñana Biological Reserve plant partitions.** Table structured as in main text Fig. 2.7, but containing all pairwise comparisons between partitions for Doñana Biological Reserve.

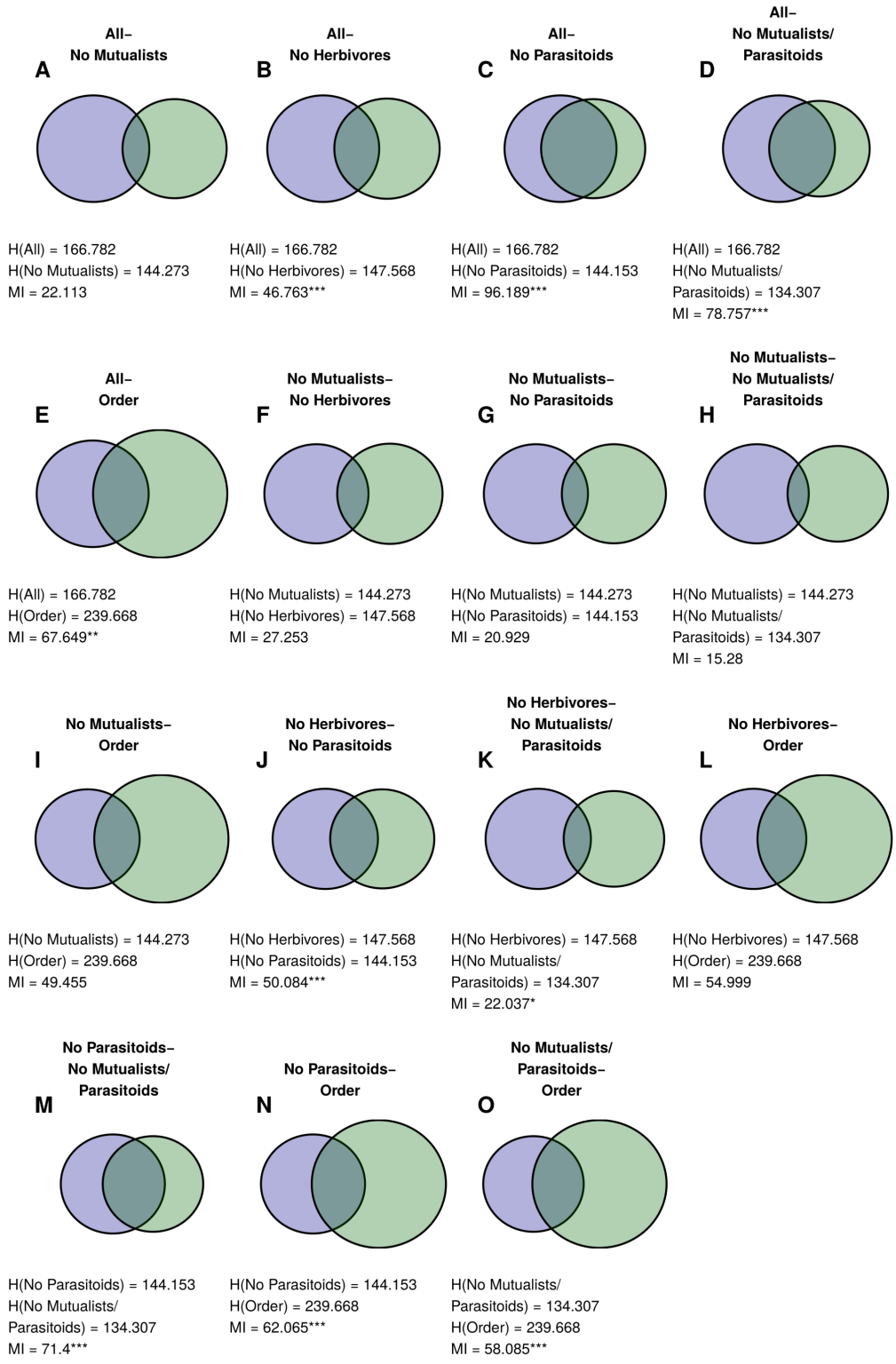

Figure 6.5: **Similarities between Norwood Farm plant partitions.** Table structured as in main text Fig. 2.9, but containing all pairwise comparisons between plant partitions for Norwood Farm.

Species Names	All	Trophic	NonTrophic
Particulate Detritus	1	6	3
Phytoplankton	1	6	12
Detritus	2	6	1
Larus glaucescens	3	4	2
Corvus caurinus	3	4	2
Haematopus bachmani	3	5	2
Dialula (Discodoris) sandiegensis	4	10	3
Archidoris montereyensis	4	10	3
Odostomia columbiana	4	10	3
Fissurelidia (Megatebennus) bimaculatus	4	10	3
Hermissenda (Phidiana) crassicornis	4	10	3
Opalia montereyensis	4	10	3
Haliaeetus leucocephalus	4	10	3
Falco peregrinus	4	10	3
Pelagic Fish	4	10	3
Pinnotheres plsum	4	10	5
Amphissa columbiana	4	11	3
Cirolana	4	11	3
Calliostoma ligatum	4	11	3
Ophiuroidea	4	11	3
Alia	4	11	3
Staphylinid Beetle	4	11	3
Small Turrid	4	11	3
Henricia	4	11	3
Chironomidae	4	11	5
sand tube worm	4	11	5
Zooplankton	4	11	5
Emplectonema gracile	5	5	3
Leptasterias hexactis	5	5	3
Notoplana	5	5	3
Pisaster ochraceus	5	5	3
Oligocottus maculosus	5	5	5
Clinocottus embryum	5	5	5
Clinocottus globiceps	5	5	5

Table 6.1: **Group identities for Tatoosh mussel bed species.** Species in the Tatoosh mussel bed network are listed in order of grouping in the complete network as shown in main text Fig. 2.6. Group identities for the trophic and nontrophic networks (as in Figs. 6.1 and 6.2, respectively) are also listed.

Species Names	All	Trophic	NonTrophic
Paranemertes peregrina	6	10	3
Amphiporus bimaculatus	6	10	3
Nereis vexillosa	6	11	3
Tonicella lineata	6	11	3
Phascolosoma agassizii	6	11	3
Cucumaria pseudocurata	6	11	3
Petrolisthes	6	11	3
Nereis sp.	6	11	3
Oedignathus inermis	6	11	3
Ceratostoma foliatum	6	11	3
Halosydna brevisetosa	6	11	3
Stichaeidae	6	11	3
Pagurid	6	11	4
Amphipoda	7	1	3
Idotea wosnesenskii	7	1	3
Chlorostoma (Tegula) funebralis	7	1	3
Katharina tunicata	7	1	5
Mopalia	7	1	5
Strongylocentrotus purpuratus	7	1	5
Strongylocentrotus droebachensis	7	1	5
Strongylocentrotus franciscanus	7	1	5
Littorina sitkana	8	12	3
Littorina scutulata	8	12	3
Lacuna vincta	8	12	3
Siphonaria thersites	8	12	3
Onchidella borealis	8	12	4
Lepidochitona	8	12	5
Lottia pelta	9	2	5
Lottia digitalis	9	2	5
Lottia paradigitalis	9	2	5
Lottia (Tectura) scutum	9	2	5
Ocenebra interfossa	10	3	3
Nucella ostrina	10	3	5
Nucella canalicalata	10	3	5
Nucella lamellosa	10	3	5
Chthamalus dalli	11	11	6
Balanus nubilus	11	13	5
Entodesma saxicola	11	13	5
Serpula vermicularis	11	13	5

Table 6.1, continued

Species Names	All	Trophic	NonTrophic
Anthopleura	12	5	6
Sacharina (Hedophyllum) sessilis	13	8	7
Laminaria	13	8	7
Alaria	13	8	7
Costaria costata	13	8	7
Postelsia palmaeformis	13	8	7
Halichondria Haliclona	14	9	12
Phyllospadix	14	9	13
Diatoms	15	7	8
Corallina vancouvriensis	16	9	9
Articulated corralines	16	9	9
Petrocelis Ralfsia	16	10	8
Halosaccion glandiforme	17	8	10
Mastocarpus (Gigartina)	17	8	10
Endocladia muricata	17	8	10
Ulva	17	8	10
Porphyra	17	8	10
Acrosiphonia (Spongomorpha) coalita	17	8	10
Fucus distichus	17	8	10
Microcladia borealis	17	8	10
Iridaea	17	8	10
Leathesia marine (difformis)	17	8	10
Polysiphonia	17	8	10
Neorhodomela	17	8	10
Prionitis	17	8	10
Petalonia fasciata	17	8	10
Callithamnion pikeanum	17	8	10
Scytosiphon lomentaria	17	8	10
Cumagloia andersonii	17	8	10
Analipus japonicus	17	8	10
Enteromorpha	17	8	10
Semibalanus cariosus	18	13	11
Balanus glandula	18	13	11
Mytilus trossulus	19	13	12
Pollicipes polymerus	19	13	12
Eudistylia	19	13	12
Mytilus californianus	19	13	13

Table 6.1, continued

CHAPTER 7

APPENDIX 2: SUPPLEMENTAL INFORMATION FOR *UNDERSTANDING THE ROLE OF PARASITES IN FOOD WEBS USING THE GROUP MODEL*

7.1 Group Model

The group model organizes species into groups such that species within a group tend to interact with other groups in the same way. Here, and throughout, we consider the directed case. The undirected case is mathematically very similar and is given a fuller treatment in [132]. In the most basic version of this model, it is possible to calculate the likelihood of obtaining the observed network A given the block structure or grouping G, where each block is treated as a separate Erdős-Rényi random graph. This likelihood is given as:

$$\mathcal{L}\left(A|G, c_{rs}; r, s \in 1, \ldots, g\right) = \prod_{r=1}^{g} \prod_{s=1}^{g} c_{rs}^{L_{rs}} \left(1 - c_{rs}\right)^{S_r S_s - L_{rs}} \tag{7.1}$$

where c_{rs} is the connectance between groups r and s (note that, since the graph is directed, c_{rs} is not necessarily equal to c_{sr}), g is the number of groups, L_{rs} is the number of edges going from group r to group s, and S_r is the number of species in group r [221]. Using a uniform prior, the Bayes factor can be calculated for model selection. For two groupings G_1 and G_2, the Bayes factor is given by

$$B = \frac{P(A|G_1)}{P(A|G_2)} \tag{7.2}$$

where $P(A|G_1)$ is the marginal likelihood

$$\int_0^1 \cdots \int_0^1 P\left(c_{rs}; r, s \in 1, \ldots, g\right) \mathcal{L}\left(A|G, c_{rs}, r, s \in 1 : g\right) \mathrm{d}c_{11} \ldots \mathrm{d}c_{gg} \tag{7.3}$$

which can be integrated to give

$$\prod_{r=1}^{g}\prod_{s=1}^{g}\frac{L_{rs}!\,(S_rS_s-L_{rs})!}{(1+L_{rs})\,(1+S_rS_s)!} \tag{7.4}$$

7.2 Group Model for Multigraphs

Here we consider the group model for multigraphs, which is consistent with the degree-corrected case. For sparse networks such as ecological networks, the possibility of multiple edges between nodes does not significantly affect the results [132]. Using the multigraph group model, the number of interactions between a species from group r and a species from group s are drawn from a Poisson distribution with rate parameter ω_{rs}, such that $E[A_{ij}] = \omega_{rs}$ when $i \in r, j \in s$. Then the likelihood of the network A given the block structure G is:

$$\mathcal{L}\left(A|G,\omega_{rs};r,s\in 1,\ldots,g\right) = C\prod_{i=1}^{S}\prod_{j=1}^{S}\left(\omega_{g_ig_j}\right)^{A_{ij}}exp\left(-\omega_{g_ig_j}\right) \tag{7.5}$$

where

$$C = \prod_{i=1}^{S}\prod_{j=1}^{S}\left(A_{ij}!\right)^{-1} \tag{7.6}$$

The likelihood of a single block rs is then

$$\prod_{i\in r}\prod_{j\in s}\left(\omega_{g_ig_j}\right)^{A_{ij}}exp\left(-\omega_{g_ig_j}\right)$$

$$= (\omega_{rs})^{\sum_{i\in r}\sum_{j\in s}A_{ij}}\,exp\left(-\omega_{rs}\right) \tag{7.7}$$

$$= \omega_{rs}^{L_{rs}}exp\left(-\omega_{rs}\right)$$

Substituting this back into the likelihood for the full network, we get

$$\mathcal{L}\left(A|G,\omega_{rs};r,s\in 1,\ldots,g\right)=C\prod_{r=1}^{g}\prod_{s=1}^{g}\omega_{rs}^{L_{rs}}exp\left(-\omega_{rs}\right) \tag{7.8}$$

Note that the fractional term C is constant with respect to the network structure, and therefore can be ignored when using Bayes factors for model selection on the same network. Note that $C=1$ when the network is not a multigraph. Using equation 2 for the Bayes factor, we need to calculate the marginal likelihood. We can then use a Gamma distribution as a (conjugate) prior for the ωs, for the marginal likelihood

$$P(A|G_1)=C\int_0^{\infty}\cdots\int_0^{\infty}\prod_{r=1}^{g}\prod_{s=1}^{g}\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\omega_{rs}^{L_{rs}+\alpha-1}exp\left(-\omega_{rs}\left(1+\beta\right)\right)\mathrm{d}\omega_{11}\ldots\mathrm{d}\omega_{gg} \tag{7.9}$$

Since the likelihood is a function, and the prior is the probability of that function, the quantity inside the integrals and products is an expectation. Since the expectation of the product is the product of expectations, this may be rewritten as

$$P(A|G_1)=C\prod_{r=1}^{g}\prod_{s=1}^{g}\int_0^{\infty}\cdots\int_0^{\infty}\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\omega_{rs}^{L_{rs}+\alpha-1}exp\left(-\omega_{rs}\left(1+\beta\right)\right)\mathrm{d}\omega_{11}\ldots\mathrm{d}\omega_{gg} \tag{7.10}$$

which can be integrated to give

$$C\left(\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\right)^{g^2}\left(1+\beta\right)^{-\alpha g^2+L}\prod_{r=1}^{g}\prod_{s=1}^{g}\Gamma\left(\alpha+L_{rs}\right) \tag{7.11}$$

where L is the total number of links in the network.

7.3 Degree-Corrected Directed Group Model

Now to incorporate degree correction into the group model for multigraphs, consider parameters θ_i and ϕ_i, $i \in 1:S$, which control the expected in- and out-degree of vertex i, respectively. Then the expected number of edges going from species j to i is given by

$$E\left[A_{ij}\right] = \theta_i \phi_j \omega_{g_i g_j} \tag{7.12}$$

Then the likelihood may be written as

$$\mathcal{L}\left(A|G,\theta,\phi,\omega\right) = C \prod_{i=1}^{S}\prod_{j=1}^{S}\left(\theta_i\phi_j\omega_{g_i g_j}\right)^{A_{ij}} exp\left(-\theta_i\phi_j\omega_{g_i g_j}\right) \tag{7.13}$$

To normalize the θs and ϕs, we impose the following constraint for all groups r:

$$\sum_{i=1}^{S}\theta_i\delta_{g_i,r} = \sum_{i=1}^{S}\phi_i\delta_{g_i,r} = 1 \tag{7.14}$$

where $\delta_{g_i,r}$ is the Kronecker delta. The likelihood can then be simplified, again by considering a single block rs:

$$\omega_{rs}^{L_{rs}}\prod_{i\in r}\prod_{j\in s}\left(\theta_i\phi_j\right)^{A_{ij}} exp\left(-\omega_{rs}\theta_i\phi_j\right) \tag{7.15}$$

$$= \omega_{rs}^{L_{rs}} exp\left[-\omega_{rs}\sum_{i\in r}\sum_{j\in s}\theta_i\phi_j\right]\prod_{i\in r}\prod_{j\in s}\theta_i^{A_{ij}}\phi_j^{A_{ij}} \tag{7.16}$$

$$= \omega_{rs}^{L_{rs}} exp\left[-\omega_{rs}\sum_{i\in r}\theta_i\sum_{j\in s}\phi_j\right]\prod_{i\in r}\theta_i^{\sum_{j\in s}A_{ij}}\prod_{j\in s}\phi_j^{\sum_{i\in r}A_{ij}} \tag{7.17}$$

then we can use the normalization constraints and plug this back into the likelihood for the full network to get

$$\mathcal{L}\left(A|G,\theta,\phi,\omega\right) = C\prod_{i=1}^{S}\theta_i^{k_i^{\text{out}}}\prod_{i=1}^{S}\phi_i^{k_i^{\text{in}}}\prod_{r=1}^{g}\prod_{s=1}^{g}\omega_{rs}^{L_{rs}}exp\left(-\omega_{rs}\right) \tag{7.18}$$

where k_i^{in} and k_i^{out} are the observed in- and out-degree of species i, respectively. Within each groups, the θs and ϕs are on a simplex (using the constraints in equation 14). Thus for each group r, we can set a prior of $Dirichlet(\vec{1})$ over the θs and ϕs in each group to get a flat prior over the simplex. As before, we use a Gamma prior over each ω_{rs}, for the marginal likelihood

$$\begin{aligned} P(A|G_1) = C\int_0^{\infty}\cdots\int_0^{\infty}\prod_{r=1}^{g}\prod_{s=1}^{g}\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\omega_{rs}^{L_{rs}+\alpha-1}exp\left(-\omega_{rs}\left(1+\beta\right)\right)\mathrm{d}\omega_{11}\ldots\mathrm{d}\omega_{gg}\times \\ \int\cdots\int_{\Delta_r,r\in 1:g}\prod_{i\in r}\theta_i^{k_i^{\text{out}}}\mathrm{d}\theta_1\ldots\mathrm{d}\theta_S\int\cdots\int_{\Delta_s,s\in 1:g}\prod_{j\in s}\phi_j^{k_j^{\text{in}}}\mathrm{d}\phi_1\ldots\mathrm{d}\phi_S \end{aligned} \tag{7.19}$$

Using the same technique as before, the products may be moved outside of the integrals, and the result may be integrated to give:

$$\begin{aligned} P(A|G_1) = C\left[\left(\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\right)^{g^2}\prod_{r=1}^{g}\prod_{s=1}^{g}\left(1+\beta\right)^{-\left(\alpha-L_{rs}\right)}\Gamma\left(\alpha+L_{rs}\right)\right]\times \\ \left[\prod_{r=1}^{g}\frac{\prod_{i=1}^{S_r}\left(k_i^{\text{in}}\right)!\left(k_i^{\text{out}}!\right)}{\Gamma\left(S_r\right)^2\Gamma\left(S_r+\sum_{i=1}^{S_r}k_i^{\text{in}}\right)\Gamma\left(S_r+\sum_{i=1}^{S_r}k_i^{\text{out}}\right)}\right] \end{aligned} \tag{7.20}$$

7.4 Mutual Information

We compared the groupings for degree-corrected and degree-uncorrected models using the information theoretic measure of mutual information (MI). MI measures the reduction in entropy of partition B when partition A is known, such that if the MI is 1, the two

partitions contain identical information. It can be thought of as the intersection between the two entropies, and is calculated as follows:

$$MI_{AB} = H(A) + H(B) - H(A,B) = \sum_{a \in A} \sum_{b \in B} p(a,b) \ln \frac{p(a,b)}{p(a)p(b)} \tag{7.21}$$

where $H(A)$ is the entropy of partition A, and $H(A,B)$ is the joint entropy of partitions A and B. The significance of MI_{AB} was calculated using a randomization test. For more details, see [221].

7.5 Search Algorithm

High-quality partitions were searched for using Metropolis-coupled Markov Chain Monte Carlo (MC^3). This algorithm uses multiple MCMC chains, run in parallel at different temperatures, with occasional opportunities for chains to swap temperatures. The temperature parameter tunes the probability of accepting a "bad" move, that is, accepting a move that reduces the marginal likelihood. At low temperatures, the chain acts as a local search, only accepting steps which improve the marginal likelihood. At high temperatures, the chain acts more like a random walk, accepting many "bad" steps, in hopes of escaping local optima to find the globally optimal solution.

The algorithm was given a maximum number of groups g. Solutions were initialized by randomly assigning each species a group assignment between 1 and g. Throughout the search, partitions were allowed to collapse down to fewer than g groups (that is, some groups were allowed to be "empty"), but were never permitted to have more than g groups.

For further implementation details, see the main text and supplement of [221]. The structure of the search algorithm used here is identical, but with the added constraint of a maximum number of groups.

7.6 Subgraph Role Analysis

Here we investigate the usefulness of a species' subgraph contributions in classifying trophic strategies. We start by enumerating how many of each of the thirty possible subgraph-role combinations (Figure 7.1) each node of the network participates in. These distributions (a vector of length 30 for each node in the network) were run through a principal component analysis to remove co-linearities and the resulting principal component coordinates were then clustered using a k-means algorithm into 2, 3, 5, or 10 groups in `R` (corresponding to the number of groups found using the group model). The k-means algorithm divides the data into a set number of groups (without the possibility of empty groups), thus it did not make sense to also repeat the $g = 100$ case in which we were looking for a natural upper-bound on the number of groups.

The groupings found by the k-means algorithm were then evaluated for imbalance with respect to trophic strategy. The results of this analysis are depicted in Section 7.9, in tables analogous to those for the group-model groupings in Section 7.8. Because this analysis is computationally intensive, we omitted the largest network (Punta Banda).

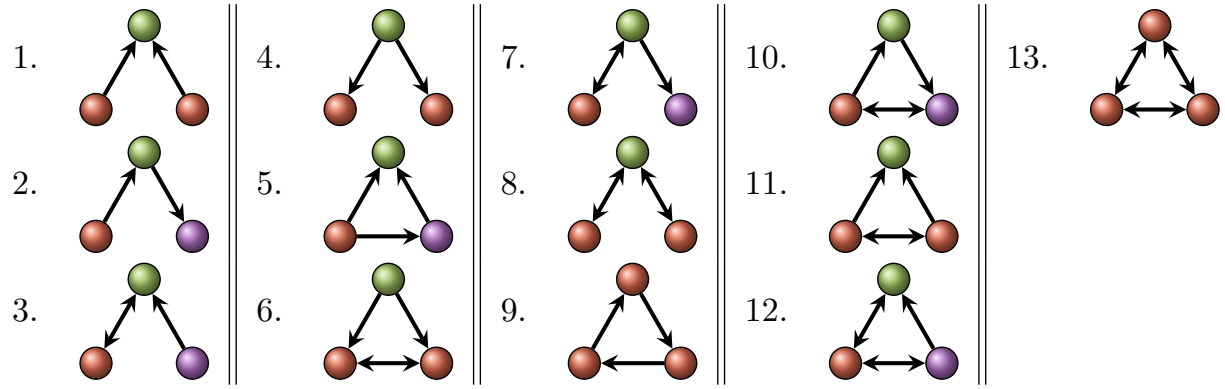

Figure 7.1: A graphical portrayal of the thirteen unique, connected, three-node subgraph structures. Unique "roles" (*i.e.* unique degree distributions within the subgraph) are depicted by differing colors. There are thirty unique subgraph-role combinations in total across these subgraphs.

7.7 Data

Name	Species	Links (with concomitant predation)	Reference
Bahia Falsa	171	2234 (3720)	[106]
Carpinteria Salt Marsh	165	2187 (3708)	[106]
Flensburg Fjord	123	968 (1406)	[276]
Otago Harbor	142	1487 (1844)	[176]
Punta Banda	214	3334 (5653)	[106]
Sylt Tidal Basin	161	1950 (3005)	[243]
Ythan Estuary	133	597 (1391)	[113]

Table 7.1: Empirical food web data used in this paper. Each row corresponds to a different food web. Columns indicate, respectively, the name of the web, its number of species, number of links without (and with) including concomitant links, and a reference for the source of the original data.

7.8 Full Imbalance Results

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.841**	0.636***	0.382	0.423	0.191**
		3/3	0.809***	0.534***	0.446***	0.406***	0.218***
		5/5	0.644*	0.717***	0.740***	0.575***	0.313***
		10/10	0.439*	0.480***	0.655***	0.459***	0.202***
		15/100	0.439***	0.329***	0.427***	0.436***	0.132***
No	Yes	2/2	0.817	0.597*	0.679***	0.494*	0.309***
		3/3	0.737	0.452	0.646***	0.576***	0.326***
		5/5	0.587	0.544***	0.568***	0.403***	0.221***
		10/10	0.450***	0.516***	0.246***	0.264***	0.115***
		38/100	0.131***	0.015***	0.012***	0.021***	0.001***
Yes	No	2/2	0.845**	0.649***	0.931***	0.544***	0.448***
		3/3	0.754***	0.580***	0.906***	0.798***	0.541***
		5/5	0.734***	0.625***	0.882***	0.897***	0.588***
		10/10	0.567***	0.645***	0.573***	0.535***	0.325***
		19/100	0.842***	0.704***	1.000	0.425***	0.358***
Yes	Yes	2/2	0.819	0.612***	0.820***	0.497*	0.422***
		3/3	0.724	0.464*	0.698***	0.473***	0.273***
		5/5	0.653***	0.425***	0.690***	0.565***	0.312***
		10/10	0.321	0.241***	0.325***	0.553***	0.091***
		43/100	0.100***	0.013***	1.000	0.067***	0.000***

Table 7.2: As Table 3.1, but for the BahiaFalsa network.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.860	0.735	0.438***	0.418	0.198
		3/3	0.836***	0.685***	0.465***	0.353***	0.195***
		5/5	0.685	0.539***	0.625***	0.576***	0.346***
		10/10	0.656***	0.587***	0.469***	0.468***	0.283***
		14/100	0.510*	0.730***	0.494***	0.312***	0.234***
No	Yes	2/2	0.865	0.750***	0.403***	0.420	0.205
		3/3	0.776	0.601	0.321***	0.498***	0.225***
		5/5	0.694	0.405	0.150***	0.289***	0.089***
		10/10	0.367	0.282***	0.147***	0.146***	0.041***
		39/100	0.234***	0.073***	0.011***	0.005***	0.001***
Yes	No	2/2	0.840	0.693	0.542***	0.420	0.236**
		3/3	0.803	0.623	0.857***	0.639***	0.358***
		5/5	0.803***	0.623***	0.823***	0.614***	0.344***
		10/10	0.769***	0.529***	0.552***	0.500***	0.259***
		17/100	0.586***	0.264***	0.517***	0.239***	0.103***
Yes	Yes	2/2	0.862	0.743	0.788***	0.608***	0.535***
		3/3	0.785	0.647	0.572***	0.352***	0.234***
		5/5	0.632	0.401	0.318***	0.502***	0.160***
		10/10	0.457	0.207	0.194***	0.369***	0.061***
		47/100	0.315***	0.055***	0.010***	0.072***	0.002***

Table 7.3: As Table 3.1, but for the Carpinteria network.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.910	0.779	0.567***	0.496***	0.425***
		3/3	0.818	0.730	0.358***	0.404***	0.281***
		5/5	0.750	0.642	0.470***	0.461***	0.312***
		10/10	0.750	0.440	0.426***	0.970***	0.365***
		13/100	0.750**	0.674***	0.557***	0.827***	0.478***
No	Yes	2/2	0.902	0.797	0.333	0.392	0.216
		3/3	0.888*	0.622	0.492***	0.728***	0.433***
		5/5	0.647	0.549	0.265***	0.789***	0.265***
		10/10	0.500	0.347	0.268***	0.404***	0.117***
		27/100	0.200	0.200***	0.023***	0.051***	0.006***
Yes	No	2/2	0.943	0.876	0.619***	0.562***	0.438***
		3/3	0.921	0.829**	0.423***	0.470***	0.291***
		5/5	0.857**	0.690**	0.315***	0.525***	0.300***
		10/10	0.615	0.680***	0.533***	0.857***	0.426***
		14/100	0.750**	0.700***	0.567***	0.814***	0.486***
Yes	Yes	2/2	0.854	0.742	0.387*	0.561***	0.356***
		3/3	0.835	0.680	0.208	0.347***	0.189***
		5/5	0.538	0.541	0.207***	0.655***	0.177***
		10/10	0.500	0.477***	0.276***	0.909***	0.237***
		32/100	0.250	0.200***	0.024***	0.240***	0.016***

Table 7.4: As Table 3.1, but for the Flensburg network.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.931	0.517***	0.504***	0.756	0.380***
		3/3	0.902	0.588***	0.343***	0.632	0.326***
		5/5	0.667	0.603***	0.263***	0.442	0.191***
		10/10	0.800	0.527***	0.351***	0.533***	0.281***
		15/100	0.667	0.277***	0.169***	0.433***	0.113***
No	Yes	2/2	0.941	0.464	0.278	0.748	0.278
		3/3	0.911	0.288	0.221*	0.619	0.182**
		5/5	0.822	0.162	0.145***	0.546***	0.083***
		10/10	0.577	0.105***	0.063***	0.412***	0.024***
		34/100	0.211	0.055***	0.007***	0.321***	0.005***
Yes	No	2/2	0.956	0.538***	0.690***	0.786**	0.452***
		3/3	0.833	0.569***	0.543***	0.589	0.371***
		5/5	0.500	0.606***	0.297***	0.358	0.165***
		10/10	1.000***	0.401***	0.465***	0.794***	0.373***
		17/100	0.831*	0.257***	0.317***	0.755***	0.217***
Yes	Yes	2/2	0.945	0.491	0.294	0.740	0.294
		3/3	0.920*	0.354***	0.178	0.613	0.178**
		5/5	0.789	0.256***	0.101*	0.490	0.086***
		10/10	0.562	0.193***	0.052***	0.572***	0.052***
		36/100	0.225	0.072***	0.008***	0.500***	0.008***

Table 7.5: As Table 3.1, but for the Otago network.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.933*	0.933*	0.491	0.579	0.491
		3/3	0.842	0.871	0.347	0.493	0.347
		5/5	0.870	0.879*	0.487***	0.660***	0.487***
		10/10	0.473	0.756	0.507***	0.543***	0.290***
		15/100	0.510	0.633	0.280***	0.514***	0.191***
No	Yes	2/2	0.925	0.927	0.494	0.605	0.494
		3/3	0.871	0.896	0.692***	0.688***	0.603***
		5/5	0.820	0.809	0.396***	0.470***	0.335***
		10/10	0.635	0.632	0.338***	0.196***	0.129***
		36/100	0.169	0.268	0.018***	0.064***	0.007***
Yes	No	2/2	0.927	0.972	0.476	0.551	0.476
		3/3	0.899	0.959	0.849***	0.812***	0.767***
		5/5	0.829	0.947	0.554***	0.718***	0.554***
		10/10	0.540	0.778	0.468***	0.675***	0.351***
		15/100	0.540	0.771	0.309***	0.452***	0.231***
Yes	Yes	2/2	0.929	0.975	0.484	0.547	0.484
		3/3	0.878	0.964	0.831***	0.728***	0.707***
		5/5	0.758	0.915	0.504***	0.644***	0.468***
		10/10	0.559	0.859	0.207***	0.476***	0.207***
		43/100	0.213	0.500	0.007***	0.042***	0.003***

Table 7.6: As Table 3.1, but for the Sylt network.

Concomitant Links	Degree Corrected	g	1° Producers	Herbivores	Consumers	Parasites	All
No	No	2/2	0.926	0.479	0.390	0.562***	0.248***
		3/3	0.899	0.533***	0.611***	0.538***	0.373***
		5/5	0.912	0.476***	0.412***	0.778***	0.371***
		8/10	0.856	0.303***	0.255***	0.713***	0.212***
		8/100	0.856	0.303***	0.255***	0.713***	0.212***
No	Yes	2/2	0.935	0.514	0.479***	0.470	0.269***
		3/3	0.914	0.754***	0.611***	0.733***	0.550***
		5/5	0.846	0.740***	0.426***	0.452***	0.344***
		10/10	0.636	0.398***	0.063***	0.079***	0.035***
		27/100	0.480	0.007***	0.004***	0.001*	0.000***
Yes	No	2/2	0.957	0.609***	0.891***	0.543***	0.457***
		3/3	0.929*	0.483***	0.753***	0.604***	0.426***
		5/5	0.819	0.476***	0.745***	0.374***	0.312***
		10/10	0.927**	0.598***	0.643***	0.519***	0.426***
		13/100	0.874**	0.326***	0.651***	0.455***	0.276***
Yes	Yes	2/2	0.940	0.497	0.706***	0.448	0.348***
		3/3	0.909	0.394***	0.593***	0.304	0.197***
		5/5	0.853	0.492***	0.699***	0.445***	0.350***
		10/10	0.764	0.362***	0.386***	0.397***	0.198***
		34/100	0.338	0.013***	0.004***	0.064***	0.001***

Table 7.7: As Table 3.1, but for the Ythan network.

7.9 Subgraph Role Imbalance Results

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.840	0.625*	0.500***	0.528**	0.236***
	3/3	0.824	0.636***	0.338	0.335	0.156***
	5/5	0.675	0.516***	0.206*	0.167	0.075***
	10/10	0.498	0.734***	0.126***	0.180***	0.072***
Yes	2/2	0.881*	0.720***	0.391	0.459	0.217**
	3/3	0.771	0.551***	0.476***	0.328	0.174***
	5/5	0.464	0.313	0.340***	0.261**	0.046***
	10/10	0.400	0.285***	0.234***	0.436***	0.081***

Table 7.8: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the BahiaFalsa network.

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.923	0.853	0.317	0.363	0.239
	3/3	0.844*	0.717**	0.273**	0.288	0.115*
	5/5	0.593	0.756***	0.192*	0.163	0.083*
	10/10	0.569	0.481***	0.026*	0.048*	0.017***
Yes	2/2	0.876	0.796***	0.774***	0.513***	0.494***
	3/3	0.918**	0.808*	0.331	0.393	0.254*
	5/5	0.734	0.539	0.275***	0.298***	0.118***
	10/10	0.476	0.313	0.282***	0.298***	0.081***

Table 7.9: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the Carpinteria network.

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.891	0.746	0.301	0.437***	0.273*
	3/3	0.902	0.762	0.450***	0.588***	0.412***
	5/5	0.890	0.809*	0.413***	0.486***	0.368***
	10/10	0.837*	0.514	0.107**	0.115**	0.057***
Yes	2/2	0.950	0.891	0.391	0.479	0.359
	3/3	0.897	0.789	0.426***	0.489***	0.362***
	5/5	0.878*	0.633	0.145*	0.150	0.057
	10/10	0.837**	0.456	0.090***	0.107***	0.042***

Table 7.10: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the Flensburg network.

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.947	0.431	0.347*	0.759**	0.309***
	3/3	0.879	0.424***	0.165	0.500	0.165
	5/5	0.930*	0.409***	0.190*	0.522	0.184***
	10/10	0.861	0.260***	0.076***	0.270	0.049***
Yes	2/2	0.963	0.615***	0.352	0.639	0.345*
	3/3	0.879	0.394*	0.309***	0.677	0.221***
	5/5	0.941	0.444***	0.139	0.370	0.132*
	10/10	0.728	0.191***	0.087***	0.299	0.043***

Table 7.11: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the Otago network.

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.650	0.962	0.477	0.778	0.477
	3/3	0.633	0.925	0.341	0.606	0.341
	5/5	0.775	0.855	0.168	0.279	0.168
	10/10	0.711	0.697	0.063	0.139	0.063
Yes	2/2	0.947	0.982	0.470	0.512	0.470
	3/3	0.924	0.971	0.366	0.417	0.366
	5/5	0.897	0.966	0.222	0.270	0.222
	10/10	0.647	0.883	0.079**	0.136**	0.073**

Table 7.12: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the Sylt network.

Concomitant Links	g	1° Producers	Herbivores	Consumers	Parasites	All
No	2/2	0.932	0.651***	0.425**	0.527***	0.333***
	3/3	0.932	0.443	0.270	0.373	0.126**
	5/5	0.874	0.545***	0.252***	0.301***	0.121***
	10/10	0.818	0.247***	0.100**	0.365***	0.055***
Yes	2/2	0.943	0.529	0.446***	0.470	0.231***
	3/3	0.876	0.377	0.261	0.278	0.102*
	5/5	0.894	0.244	0.256***	0.164	0.075***
	10/10	0.723	0.246***	0.161***	0.072*	0.028***

Table 7.13: As Table 3.1, but for groupings based on k-means clustering of the subgraph-role contributions of each node of the Ythan network.

7.10 Mutual Information Between Group Models

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
Flensburg	Yes	2	0.42	0.64	0.19***	
Flensburg	Yes	3	0.92	1.09	0.36***	
Flensburg	Yes	5	1.43	1.58	0.78***	
Flensburg	Yes	10	2.14	2.26	1.71***	
Flensburg	No	2	0.69	0.69	0.012	
Flensburg	No	3	1.08	1.06	0.23***	
Flensburg	No	5	1.45	1.58	0.85***	
Flensburg	No	10	2.02	2.26	1.50***	

Table 7.14: Overlap between partitions that are corrected for degree and those that are not. Columns list the food web (Web), whether or not concomitant predation is included (Concomitant), the maximum number of groups the network is split into (G), the entropy of the degree-corrected and non-degree-corrected partitions ($H(Corrected)$ and $H(Uncorrected)$, the mutual information shared by the two partitions (MI), and the overlap represented as a Venn diagram. The left (purple) circle corresponds to the partition found by the group model without degree correction, and the right (green) circle corresponds to the partition found by the degree-corrected model. The area of each circle is proportional to the corresponding entropy, and the area of overlap between the circles is proportional to the mutual information. Stars next to the mutual information values correspond to the level of significance ($< .05, < .01, < .001$), as calculated by a randomization test [221]. As the number of groups increases, the entropy also increases, but the two partitions become increasingly similar. Corrected and uncorrected partitions become very similar when the network is partitioned into 10 groups, although there is always some distinct information in each.

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
Carpinteria	Yes	2	0.69	0.69	0.073^{***}	
Carpinteria	Yes	3	1.08	1.09	0.39^{***}	
Carpinteria	Yes	5	1.49	1.58	0.93^{***}	
Carpinteria	Yes	10	2.06	2.27	1.49^{***}	
Carpinteria	No	2	0.69	0.69	0.03^{**}	
Carpinteria	No	3	1.07	1.08	0.37^{***}	
Carpinteria	No	5	1.58	1.60	0.83^{***}	
Carpinteria	No	10	2.11	2.28	1.30^{***}	

Table 7.14, continued

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
Otago	Yes	2	0.65	0.69	0.0021	
Otago	Yes	3	1.02	1.10	0.14***	
Otago	Yes	5	1.42	1.59	0.72***	
Otago	Yes	10	2.07	2.25	1.48***	
Otago	No	2	0.65	0.69	0.0055	
Otago	No	3	1.06	1.10	0.21***	
Otago	No	5	1.40	1.59	0.59***	
Otago	No	10	2.04	2.26	1.34***	

Table 7.14, continued

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
BahiaFalsa	Yes	2	0.64	0.67	0.46***	
BahiaFalsa	Yes	3	1.08	1.08	0.43***	
BahiaFalsa	Yes	5	1.50	1.57	1.07***	
BahiaFalsa	Yes	10	2.16	2.25	1.51***	
BahiaFalsa	No	2	0.65	0.68	0.024**	
BahiaFalsa	No	3	1.00	1.09	0.27***	
BahiaFalsa	No	5	1.53	1.58	0.79***	
BahiaFalsa	No	10	2.12	2.27	1.54***	

Table 7.14, continued

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
Ythan	Yes	2	0.62	0.69	0.30^{***}	
Ythan	Yes	3	1.08	1.10	0.25^{***}	
Ythan	Yes	5	1.52	1.61	0.86^{***}	
Ythan	Yes	10	1.93	2.28	1.22^{***}	
Ythan	No	2	0.59	0.69	0.0017	
Ythan	No	3	0.96	1.10	0.46^{***}	
Ythan	No	5	1.33	1.58	0.68^{***}	
Ythan	No	10	1.72	2.28	1.02^{***}	

Table 7.14, continued

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
PuntaBanda	Yes	2	0.65	0.69	0.44^{***}	
PuntaBanda	Yes	3	1.07	1.08	0.44^{***}	
PuntaBanda	Yes	5	1.57	1.59	0.82^{***}	
PuntaBanda	Yes	10	2.18	2.28	1.62^{***}	
PuntaBanda	No	2	0.66	0.69	0.059^{***}	
PuntaBanda	No	3	1.09	1.10	0.066^{***}	
PuntaBanda	No	5	1.58	1.59	0.67^{***}	
PuntaBanda	No	10	2.18	2.28	1.41^{***}	

Table 7.14, continued

Web	Concomitant	G	$H(Uncorrected)$	$H(Corrected)$	MI	
Sylt	Yes	2	0.64	0.69	0.018*	
Sylt	Yes	3	0.95	1.08	0.34***	
Sylt	Yes	5	1.59	1.57	0.74***	
Sylt	Yes	10	2.12	2.28	1.46***	
Sylt	No	2	0.69	0.69	0.01	
Sylt	No	3	1.06	1.06	0.16***	
Sylt	No	5	1.51	1.59	0.74***	
Sylt	No	10	2.02	2.28	1.30***	

Table 7.14, continued

7.11 Empirical Network Adjacency Matrices – Grouped by Trophic Strategy

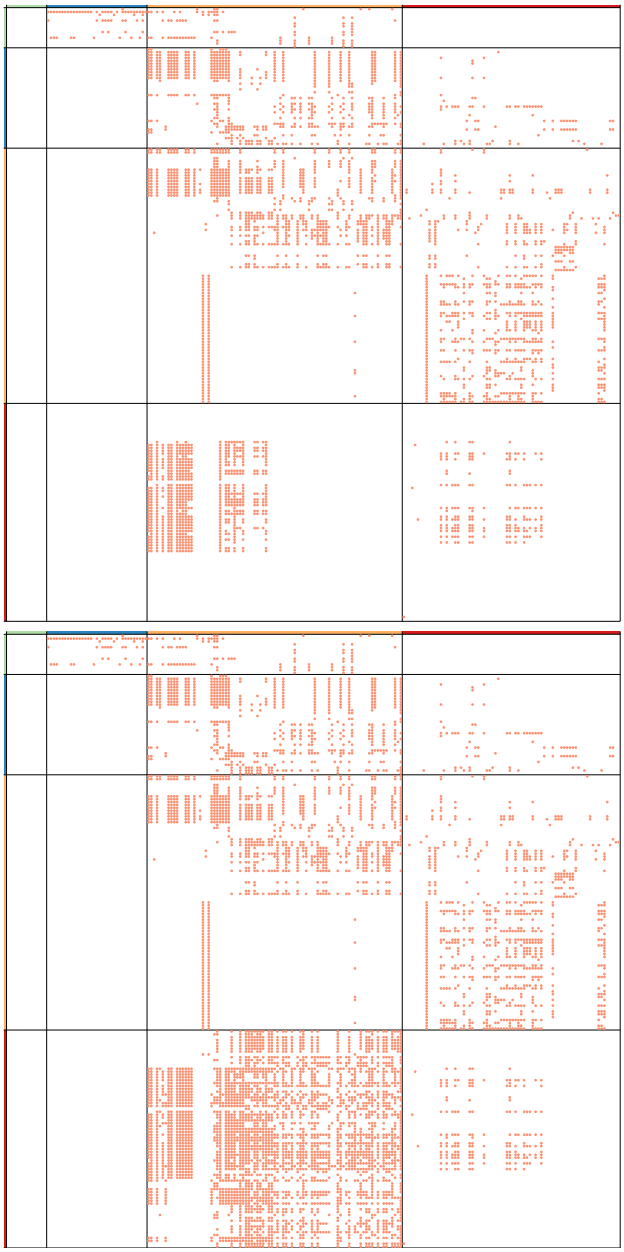

Figure 7.2: Punta Banda network structure with (top) and without (bottom) concomitant predation and species grouped by trophic strategy. Colored squares represent trophic strategy (green for primary producers, blue for herbivores, red for parasites, and yellow for other consumers), and dots represent feeding interactions wherein the column species consumes the row species. Note the addition of concomitant links increases the number of consumer-resource interactions from predators to parasites and decreases the cascade-like structure seen in the top matrix.

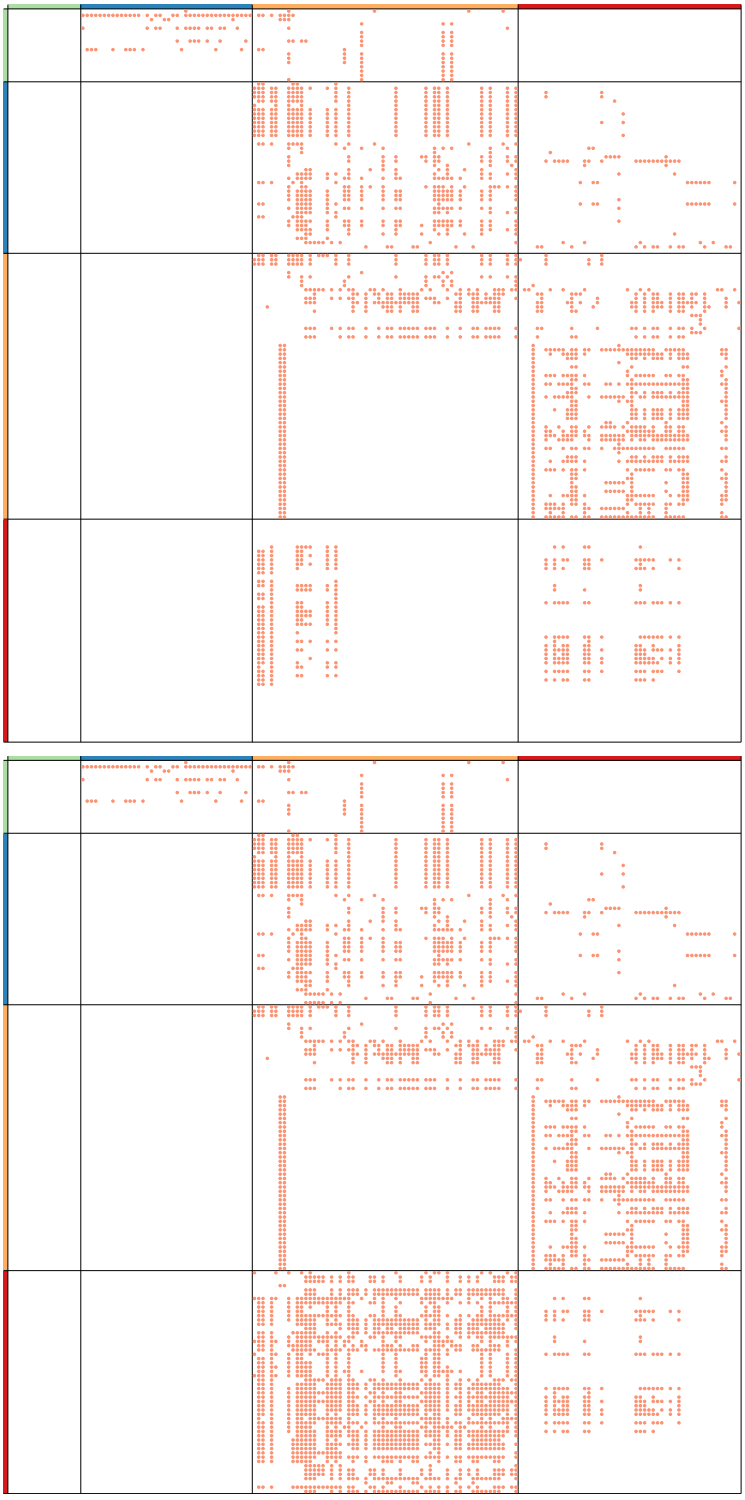

Figure 7.3: As Figure 7.2, but showing the BahiaFalsa network.

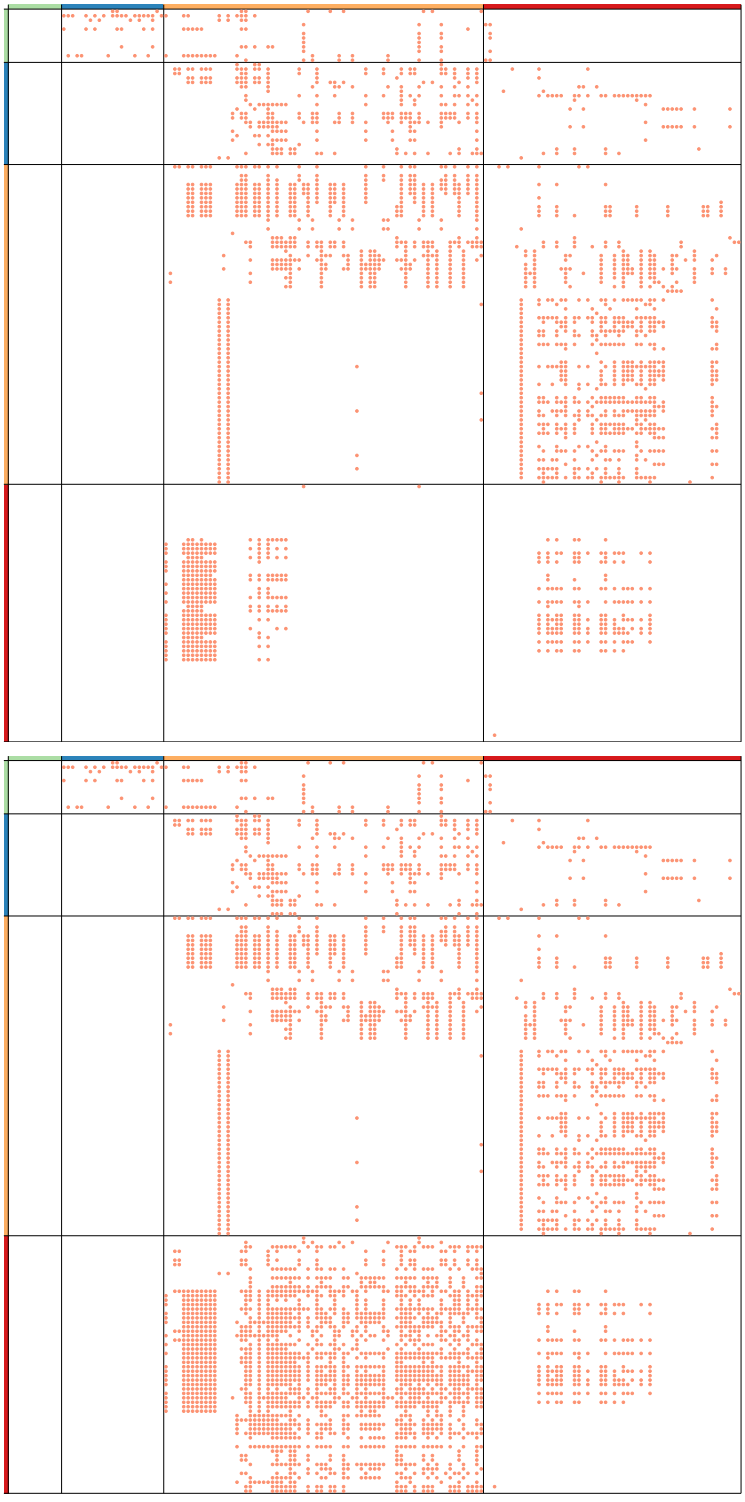

Figure 7.4: As Figure 7.2, but showing the Carpinteria network.

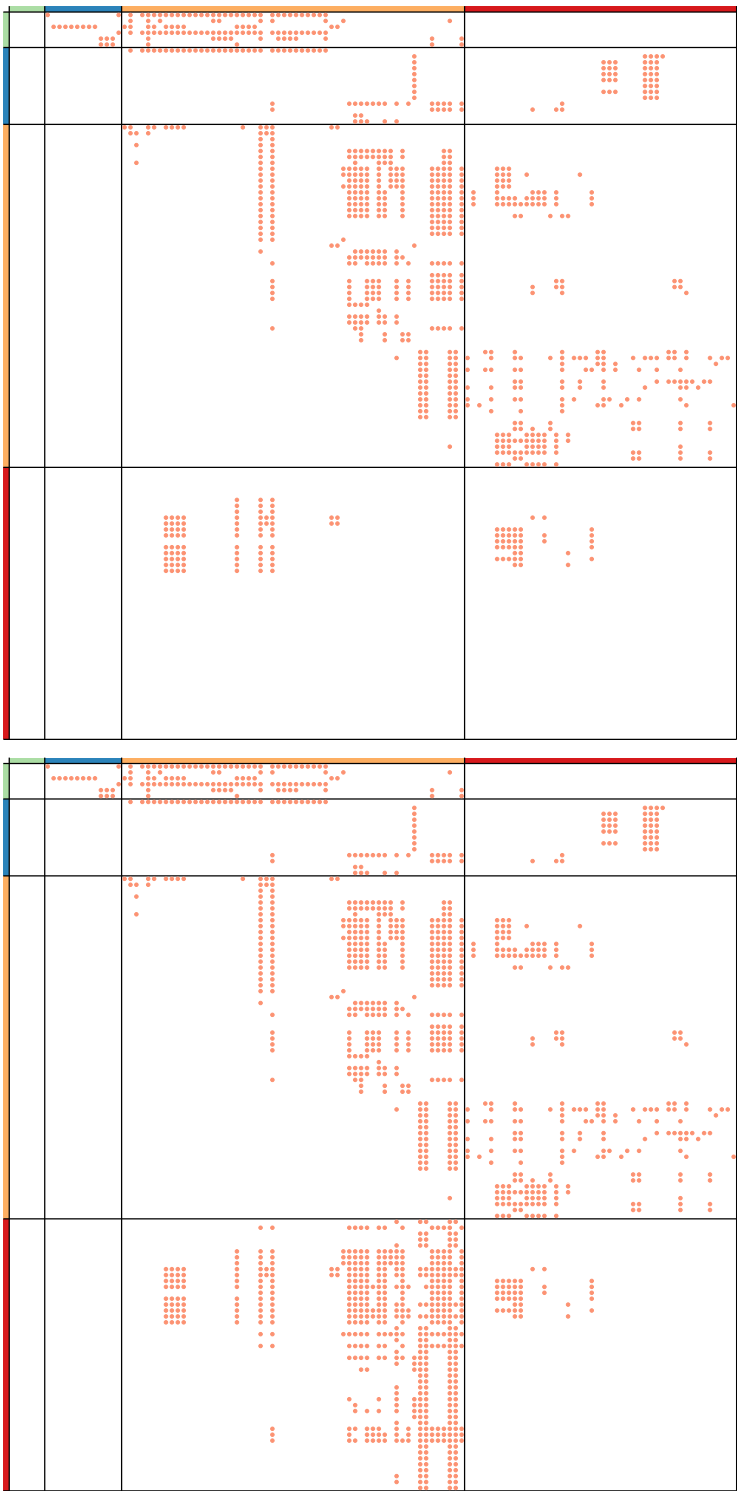

Figure 7.5: As Figure 7.2, but showing the Flensburg network.

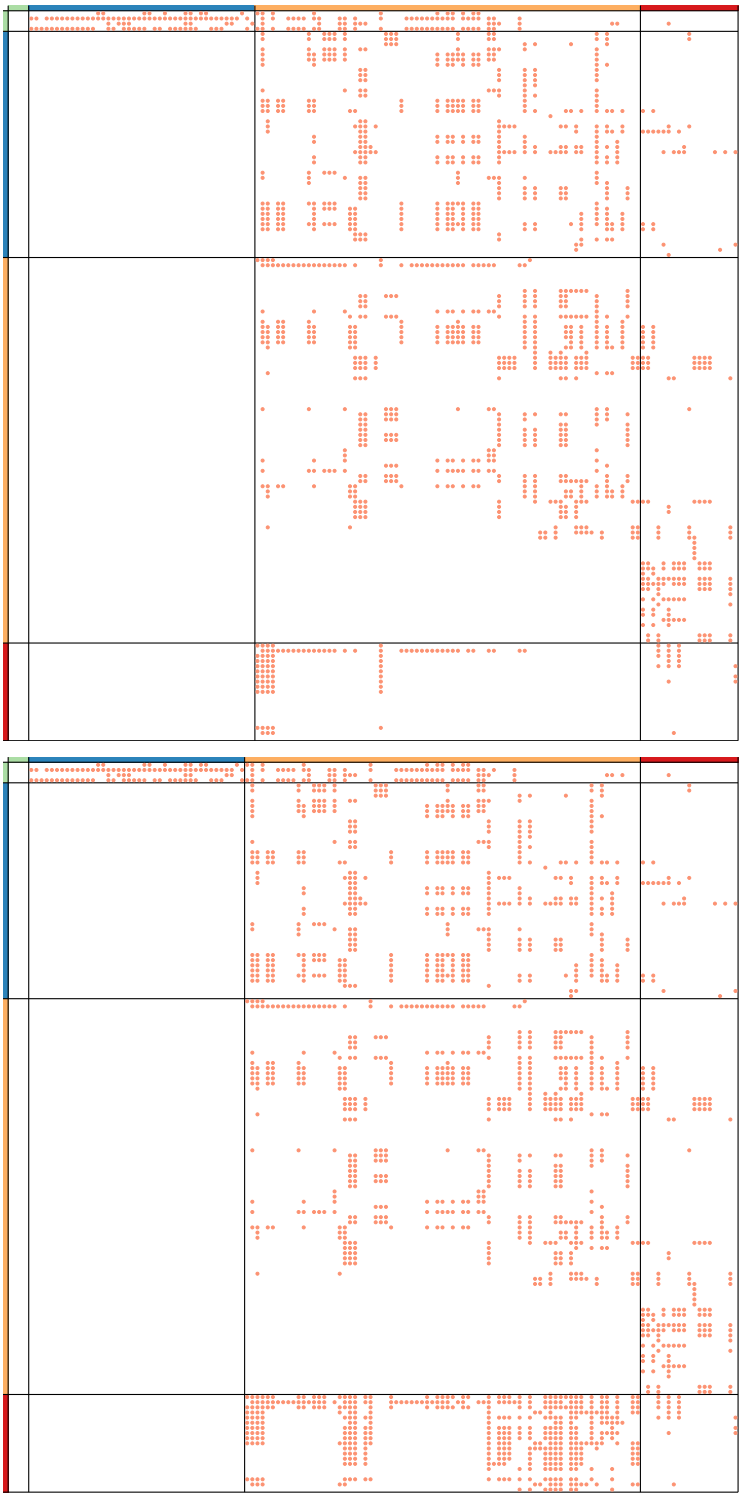

Figure 7.6: As Figure 7.2, but showing the Otago network.

Figure 7.7: As Figure 7.2, but showing the Sylt network.

Figure 7.8: As Figure 7.2, but showing the Ythan network.

7.12 Empirical Network Adjacency Matrices – Grouped by Model

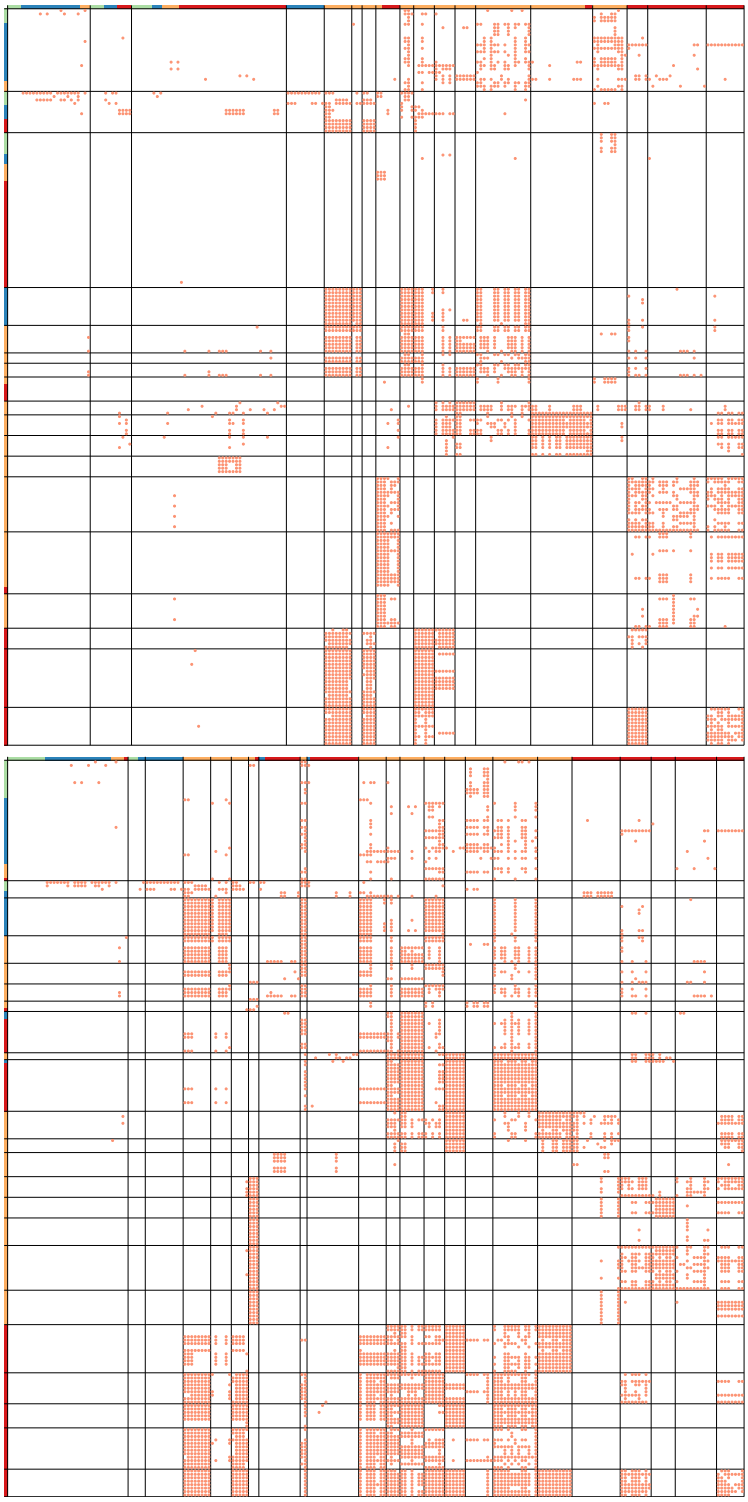

Figure 7.9: Group model results for uncorrected group model and either including concomitant links (bottom) or not (top) for the Punta Banda network.

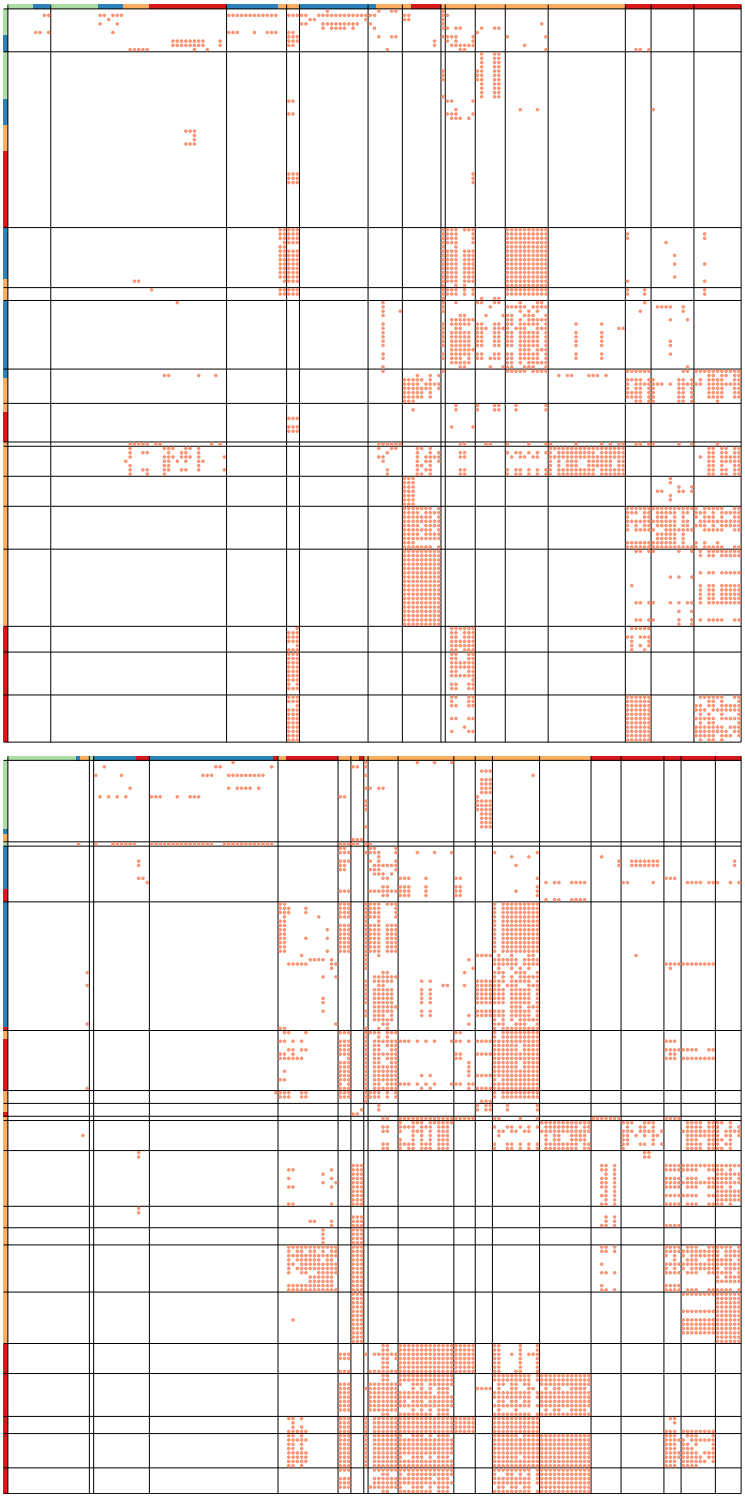

Figure 7.10: As Figure 7.9, but showing the BahiaFalsa network.

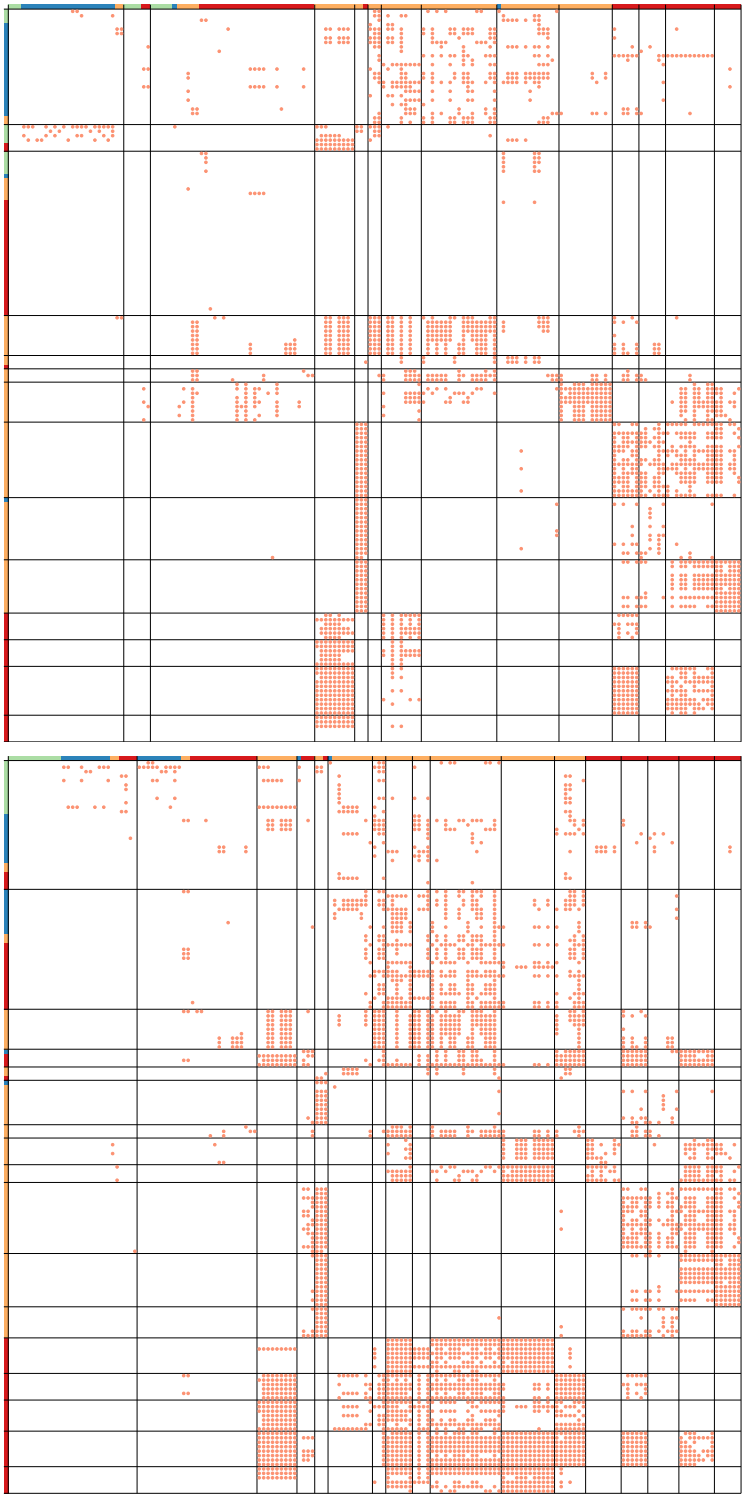

Figure 7.11: As Figure 7.9, but showing the Carpinteria network.

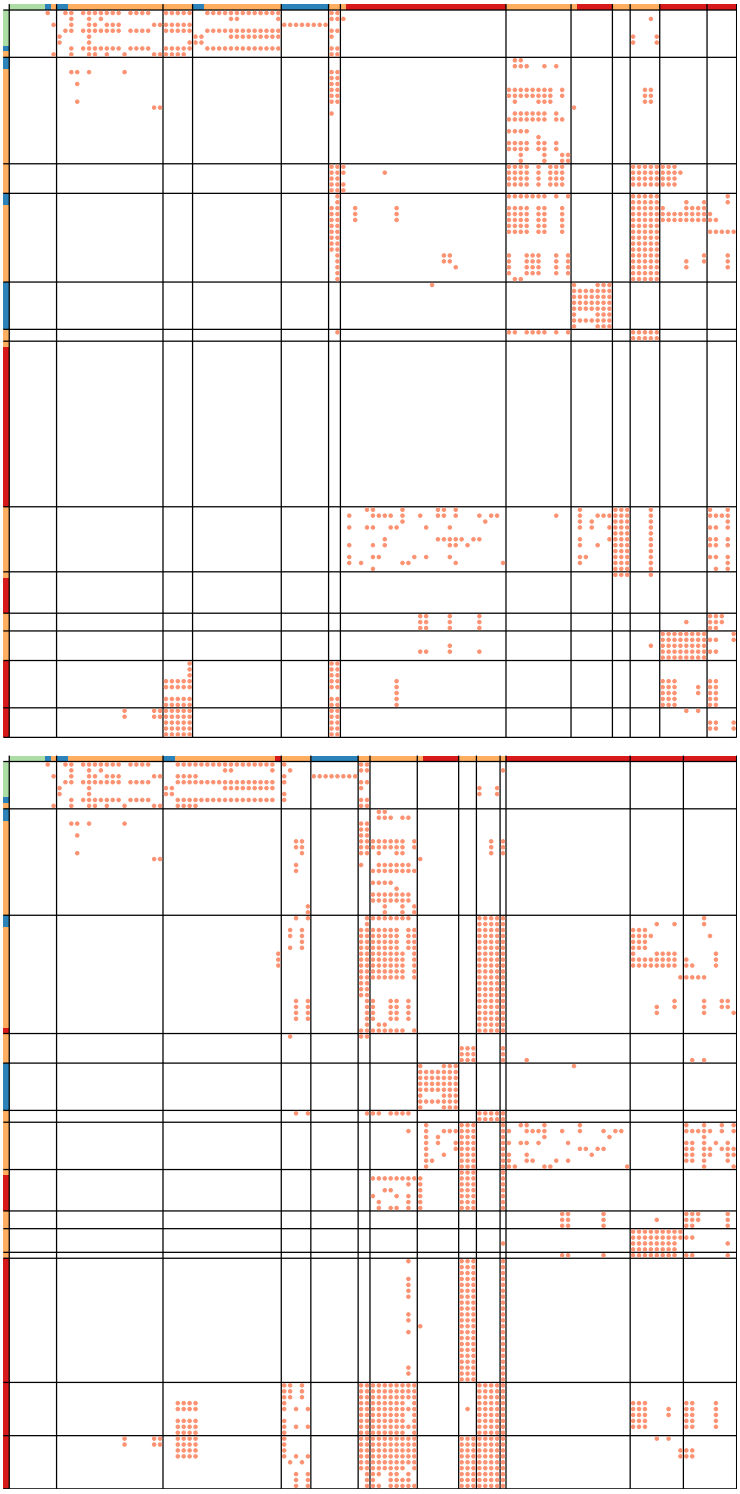

Figure 7.12: As Figure 7.9, but showing the Flensburg network.

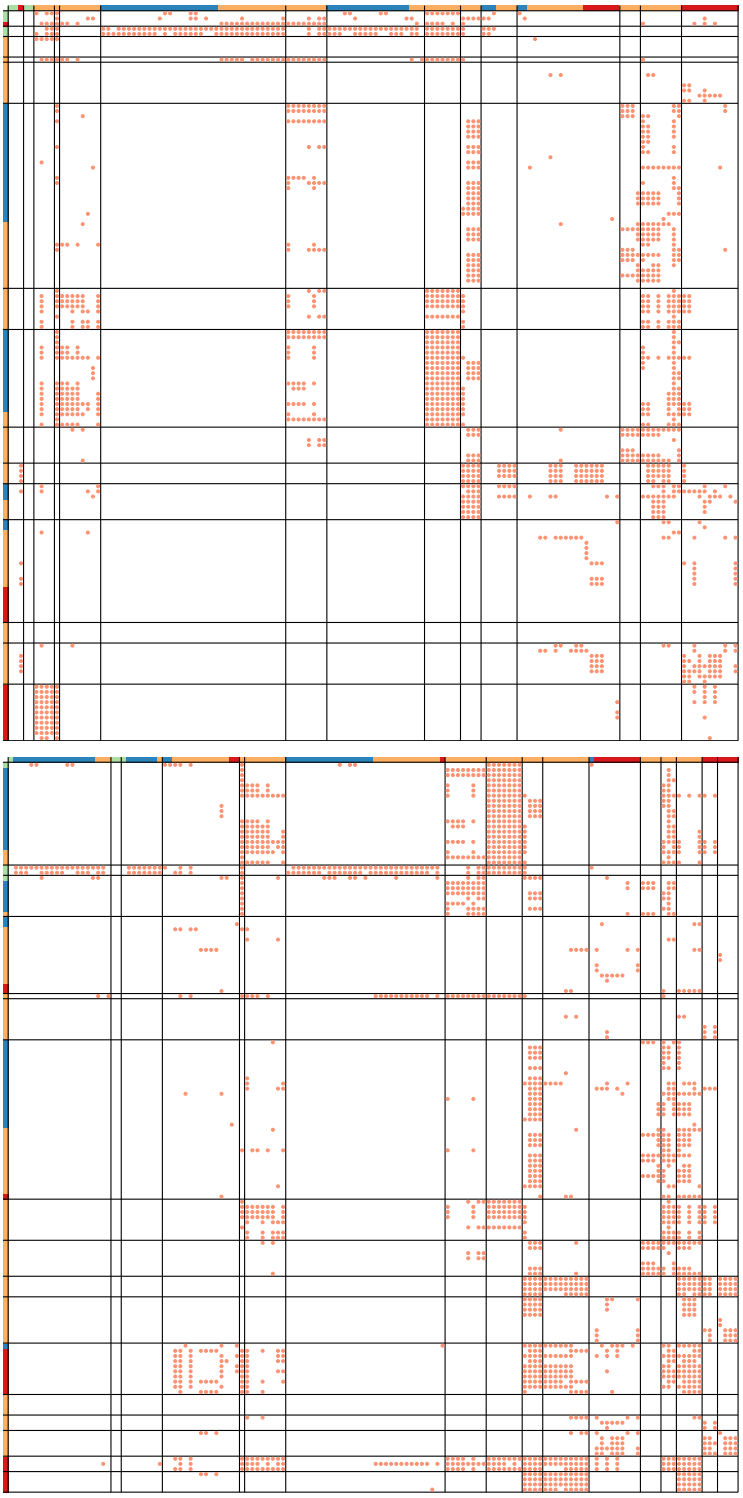

Figure 7.13: As Figure 7.9, but showing the Otago network.

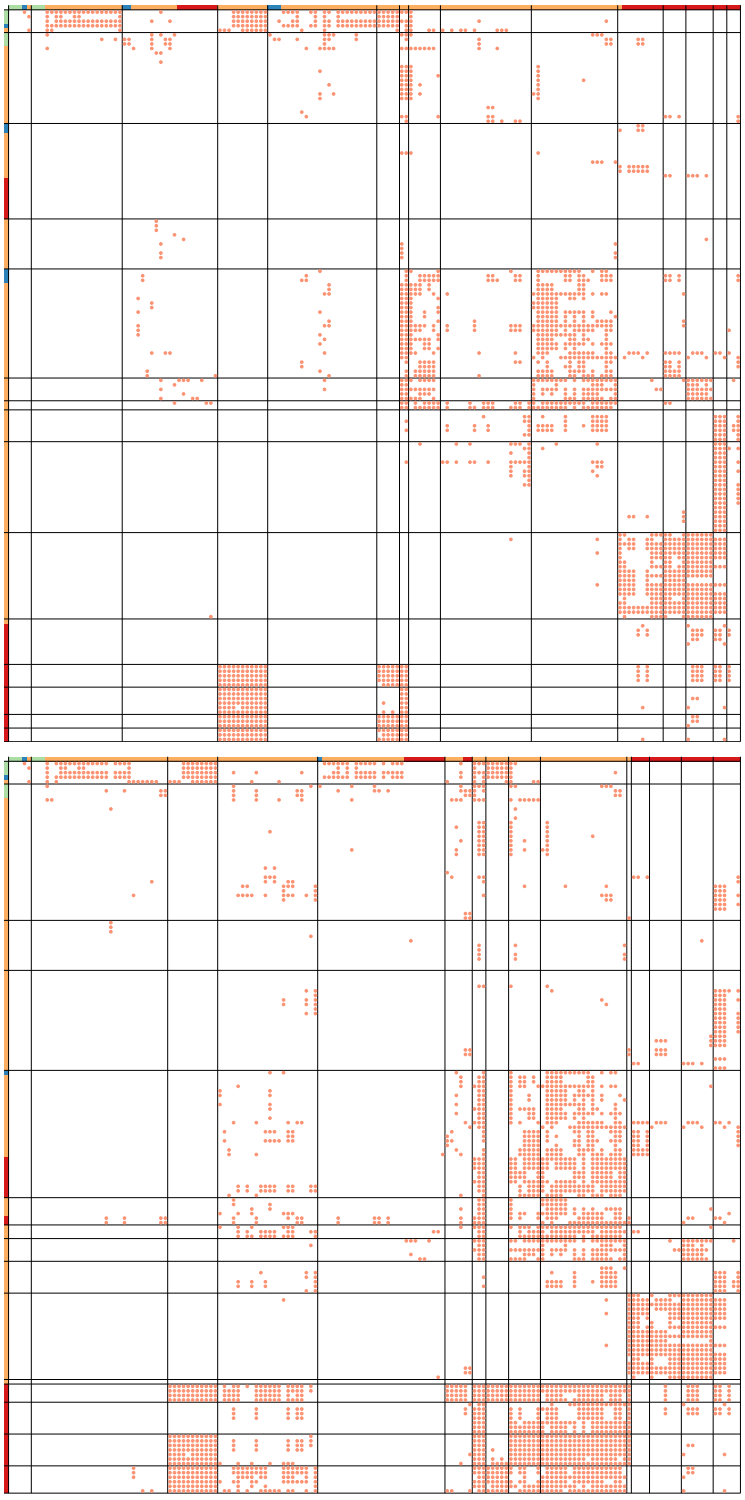

Figure 7.14: As Figure 7.9, but showing the Sylt network.

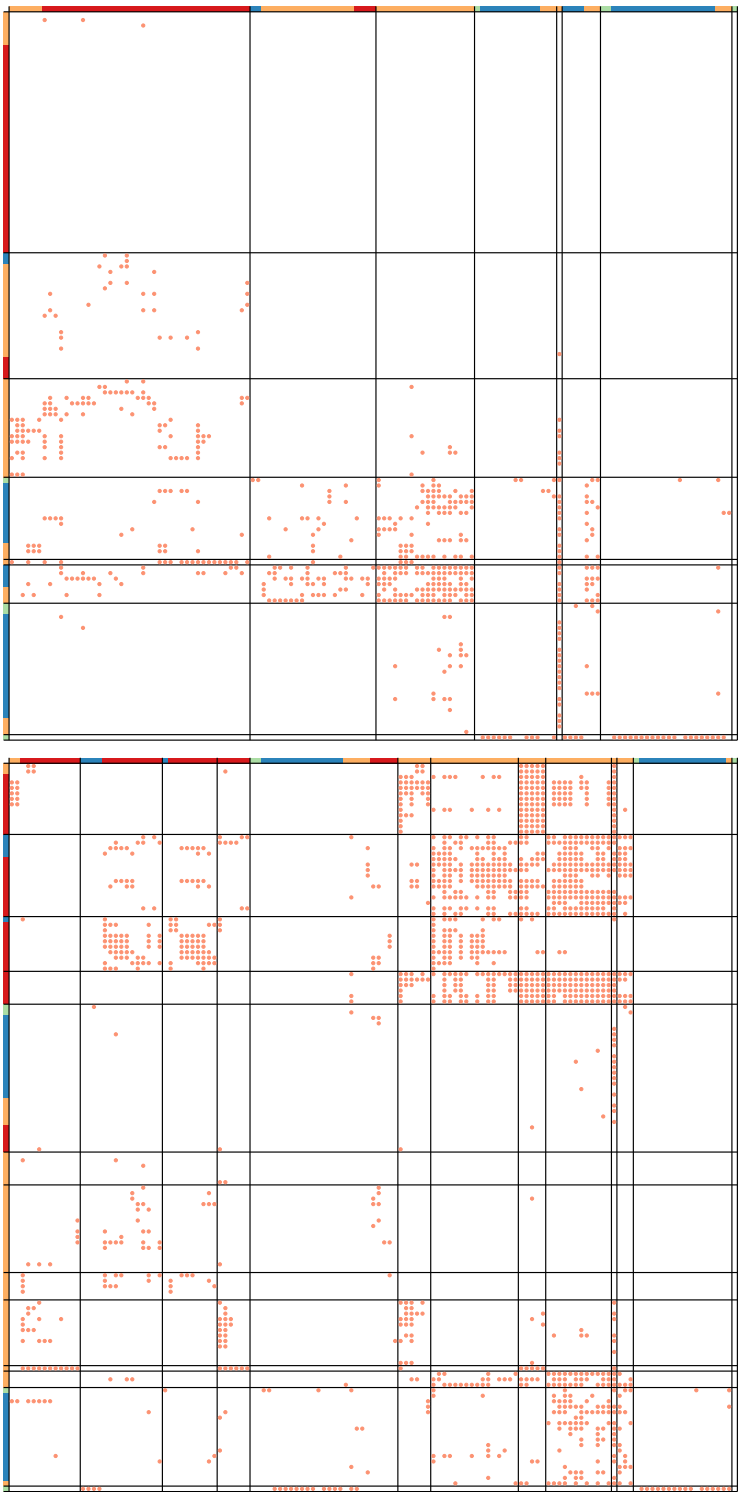

Figure 7.15: As Figure 7.9, but showing the Ythan network.

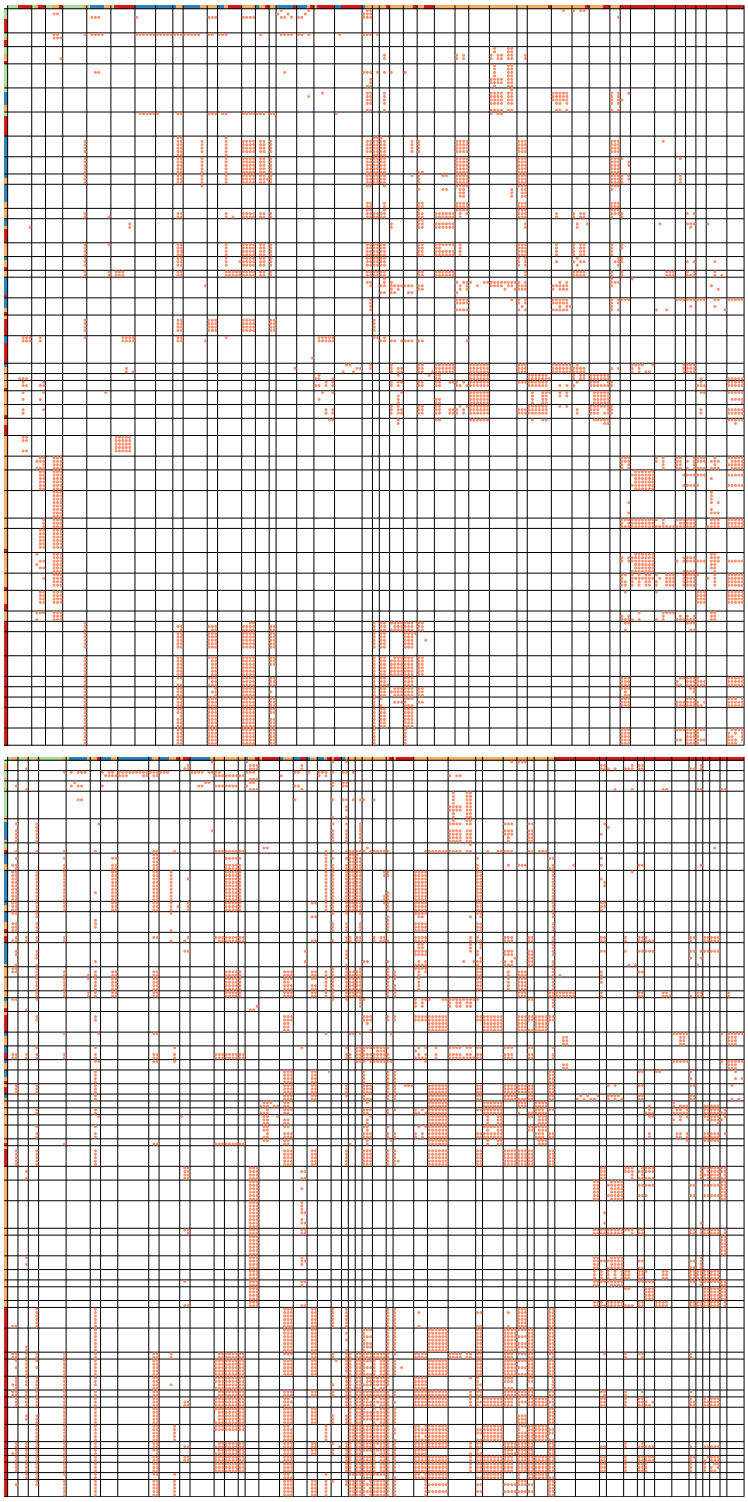

Figure 7.16: Group model results for degree corrected group model and either including concomitant links (bottom) or not (top) for the Punta Banda network.

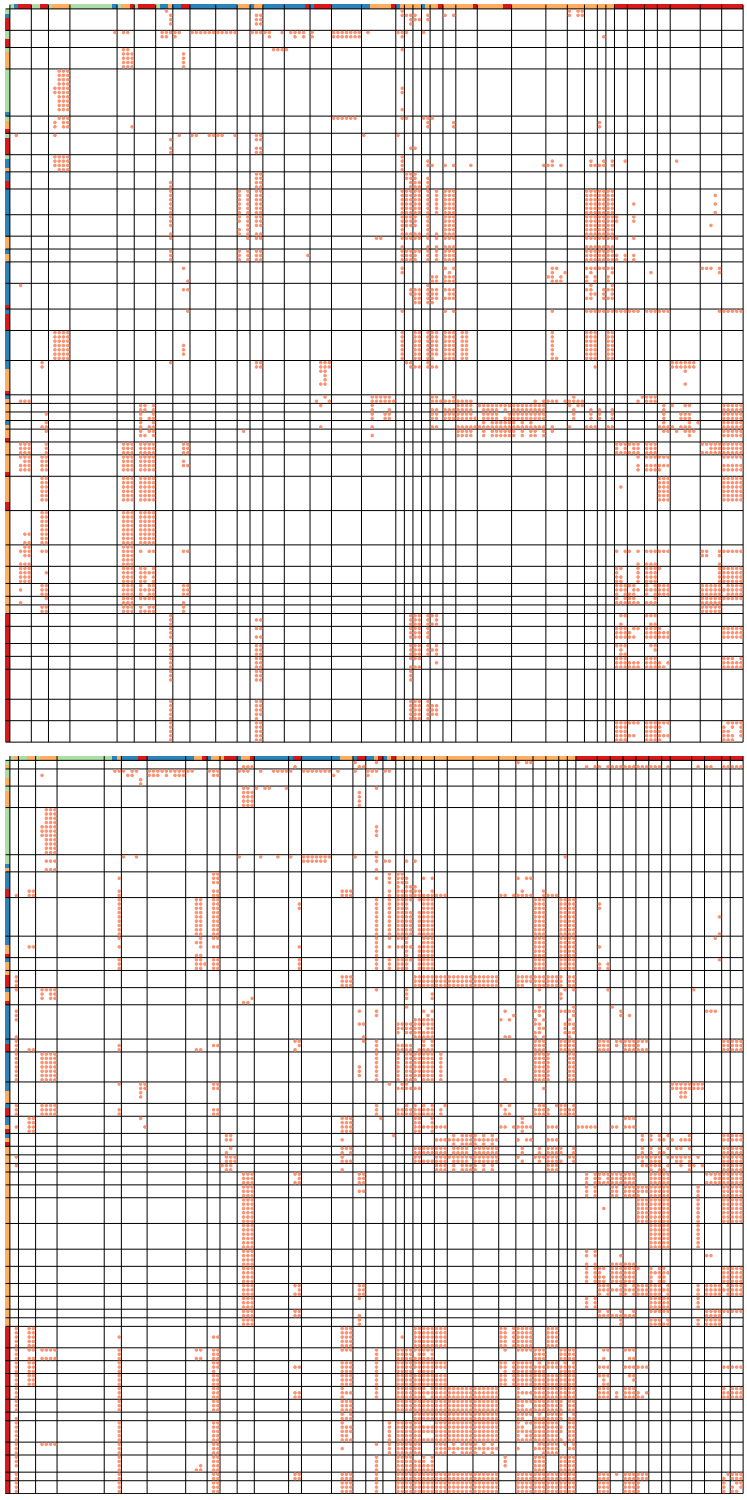

Figure 7.17: As Figure 7.16, but showing the BahiaFalsa network.

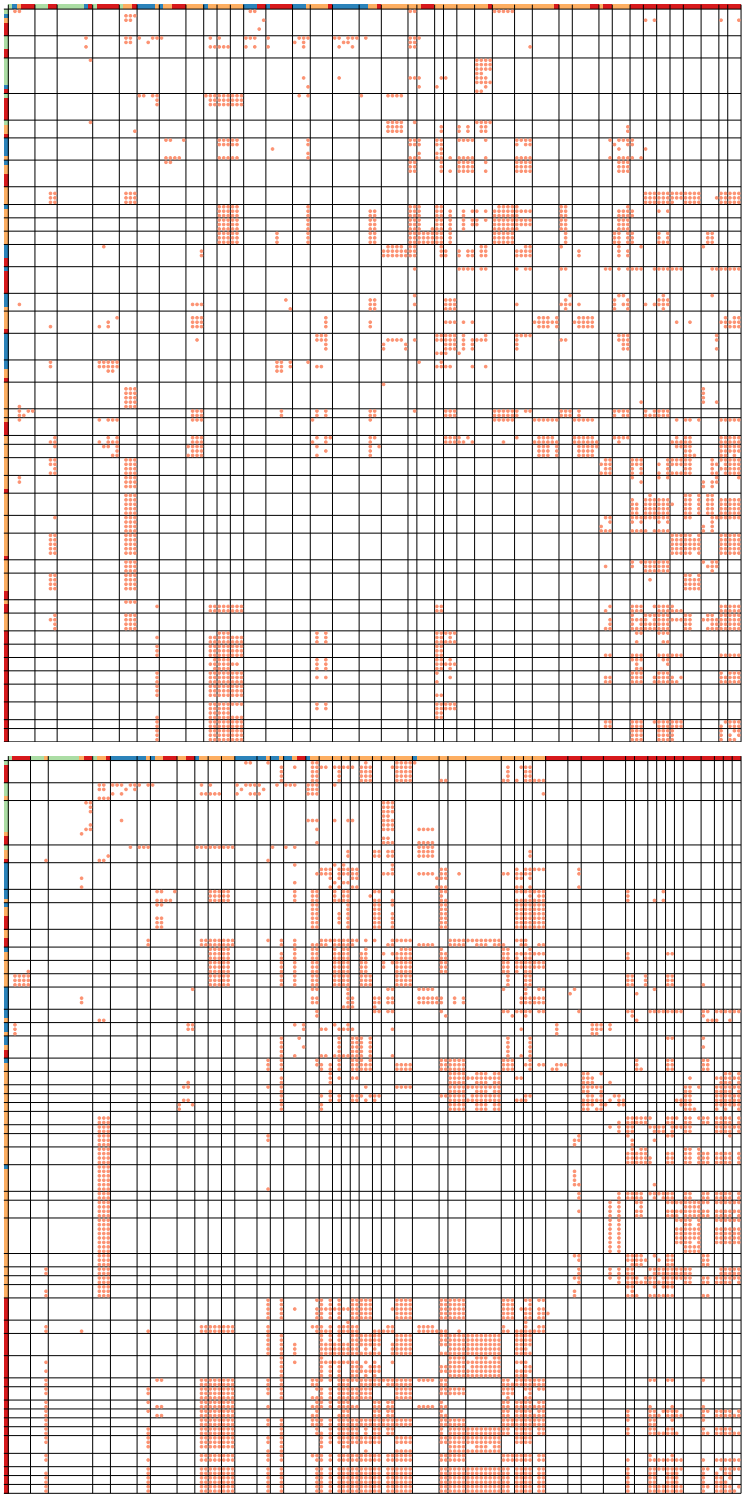

Figure 7.18: As Figure 7.16, but showing the Carpinteria network.

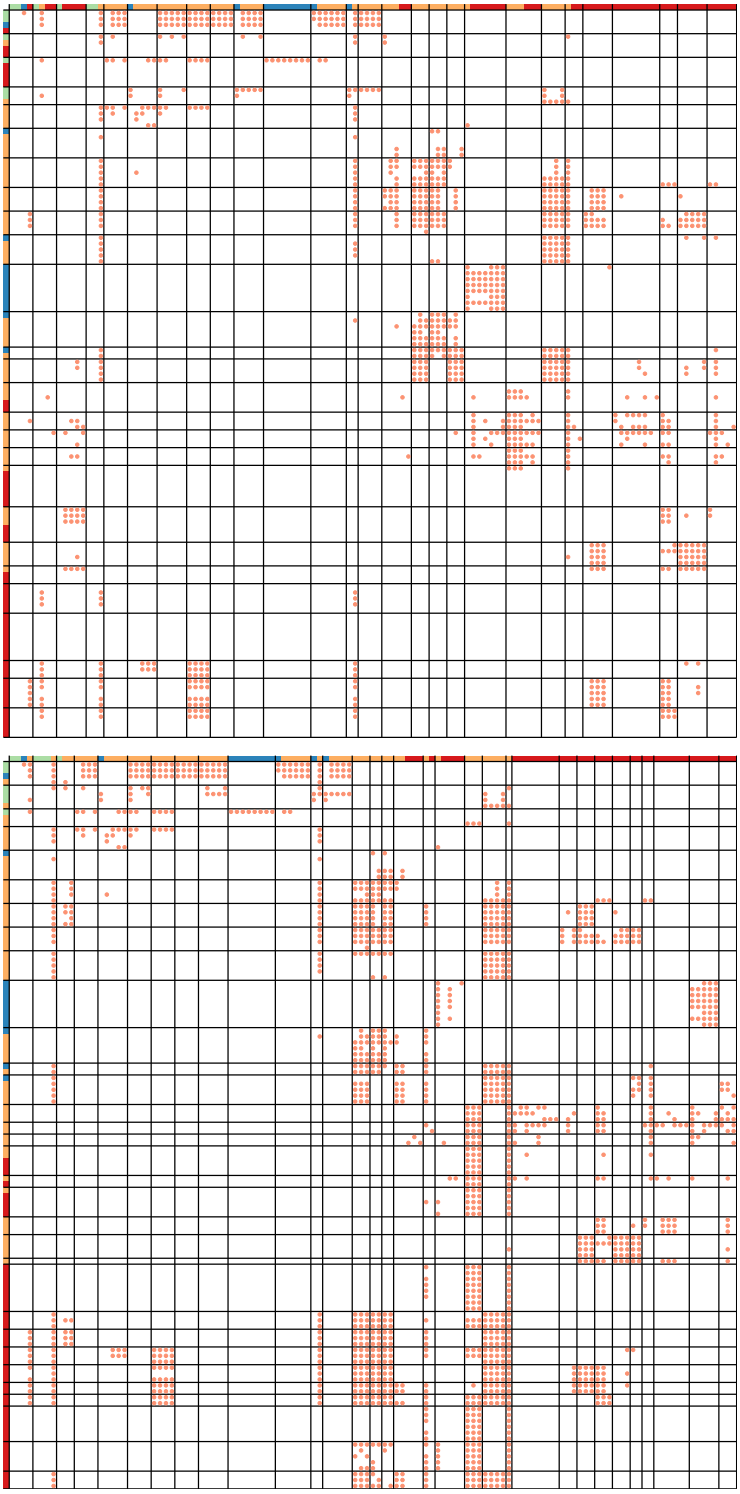

Figure 7.19: As Figure 7.16, but showing the Flensburg network.

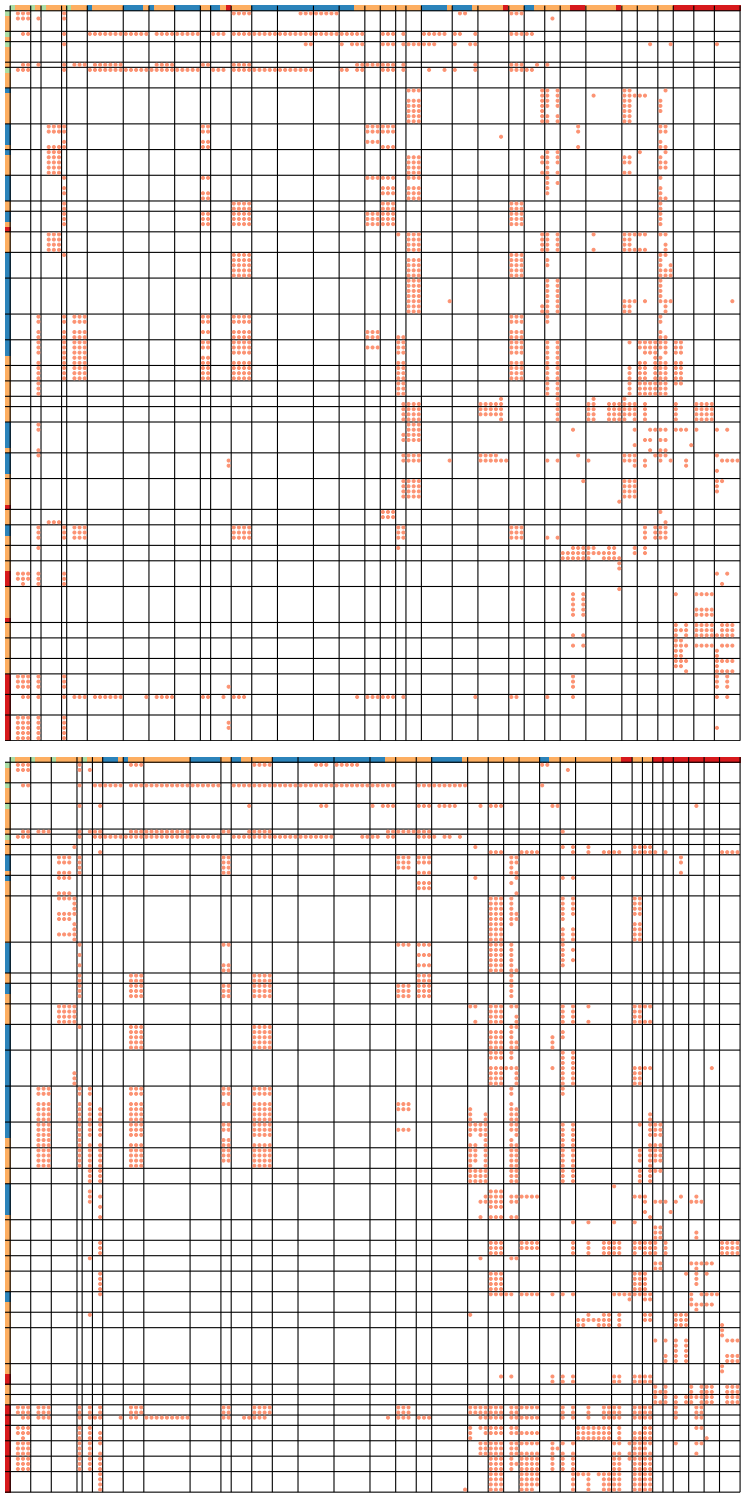

Figure 7.20: As Figure 7.16, but showing the Otago network.

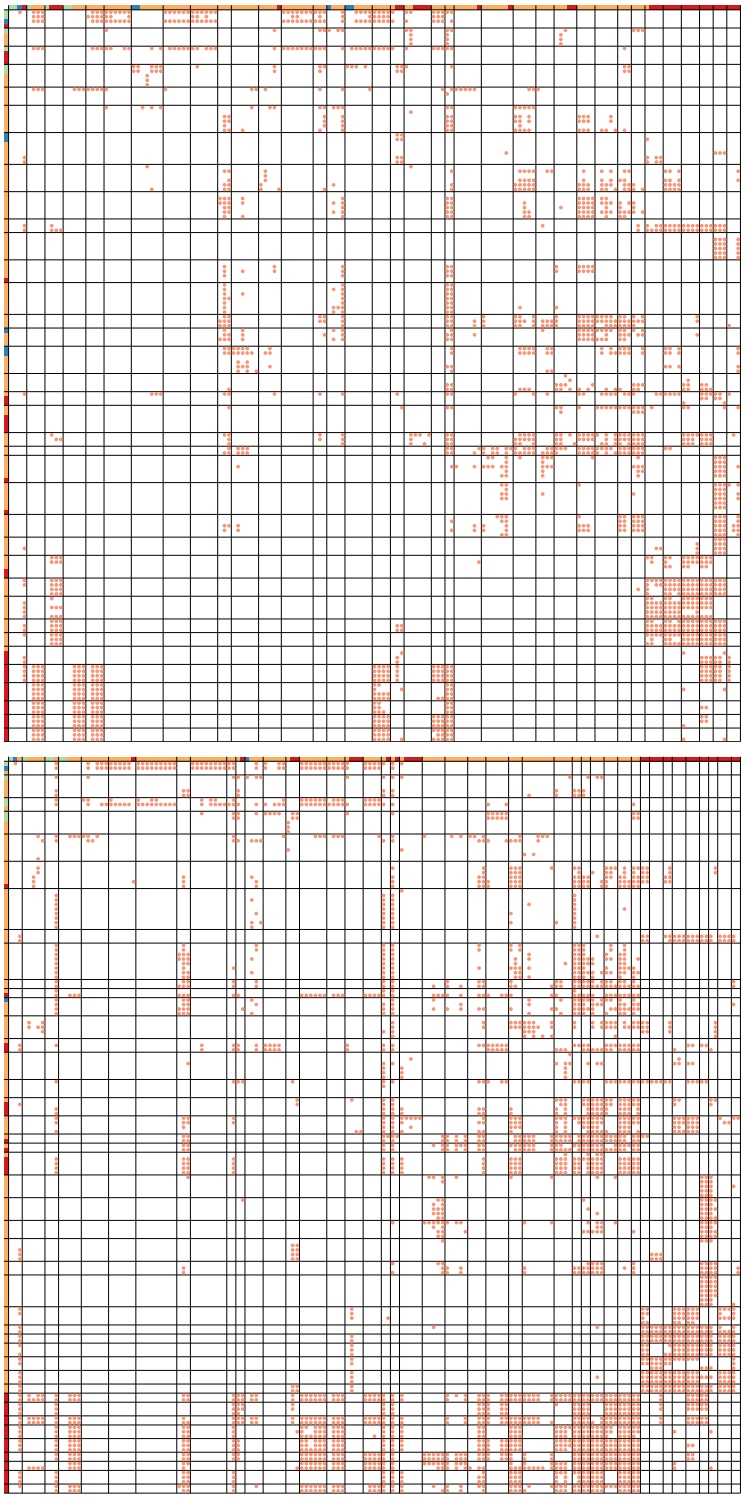

Figure 7.21: As Figure 7.16, but showing the Sylt network.

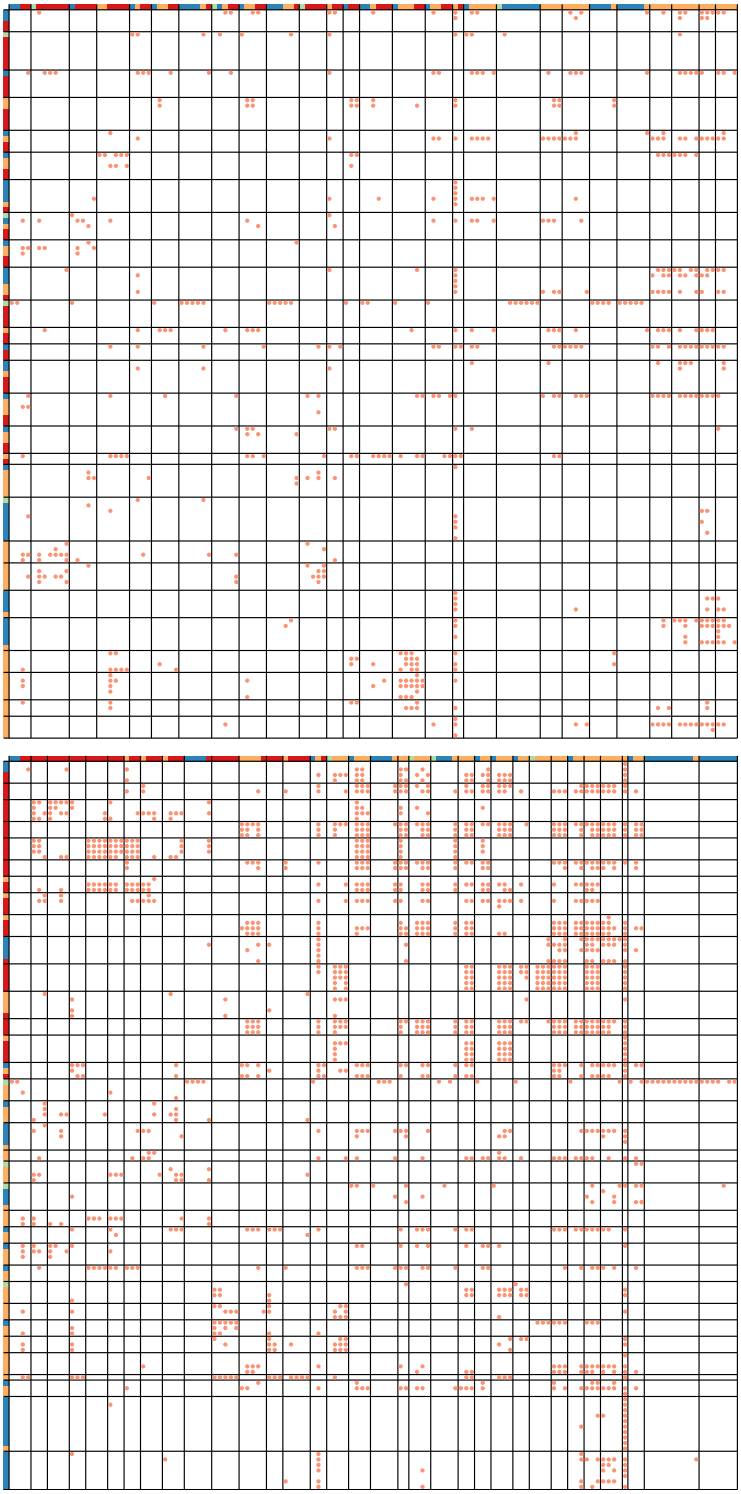

Figure 7.22: As Figure 7.16, but showing the Ythan network.

7.13 Imbalance Sampling Convergence

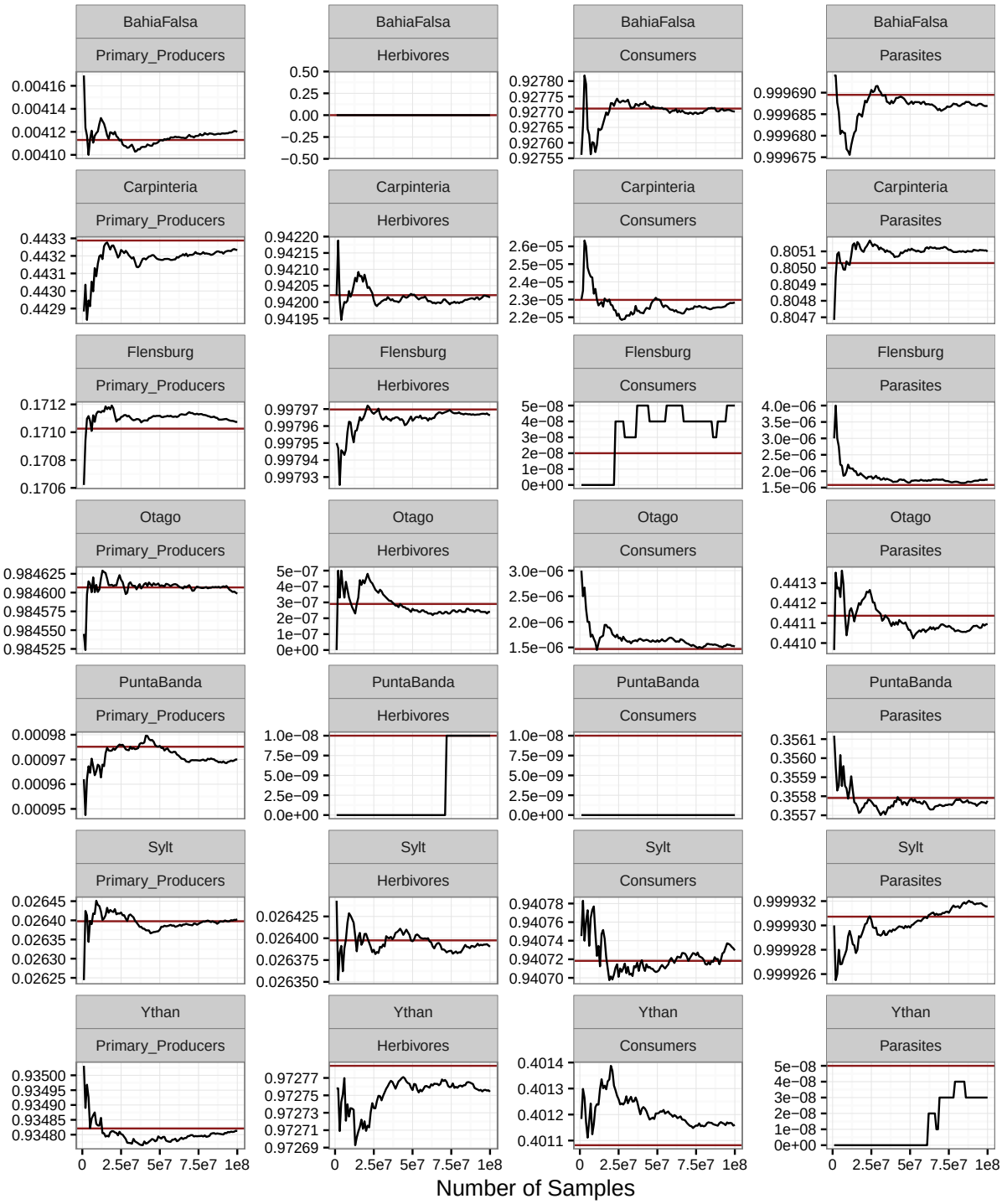

Figure 7.23: Convergence of sampling routine to analytically calculated p-values. Horizontal red line is the analytical value, while the black lines are the sampled value as the number of samples increases. Blank plots indicate computationally infeasible analytical values.

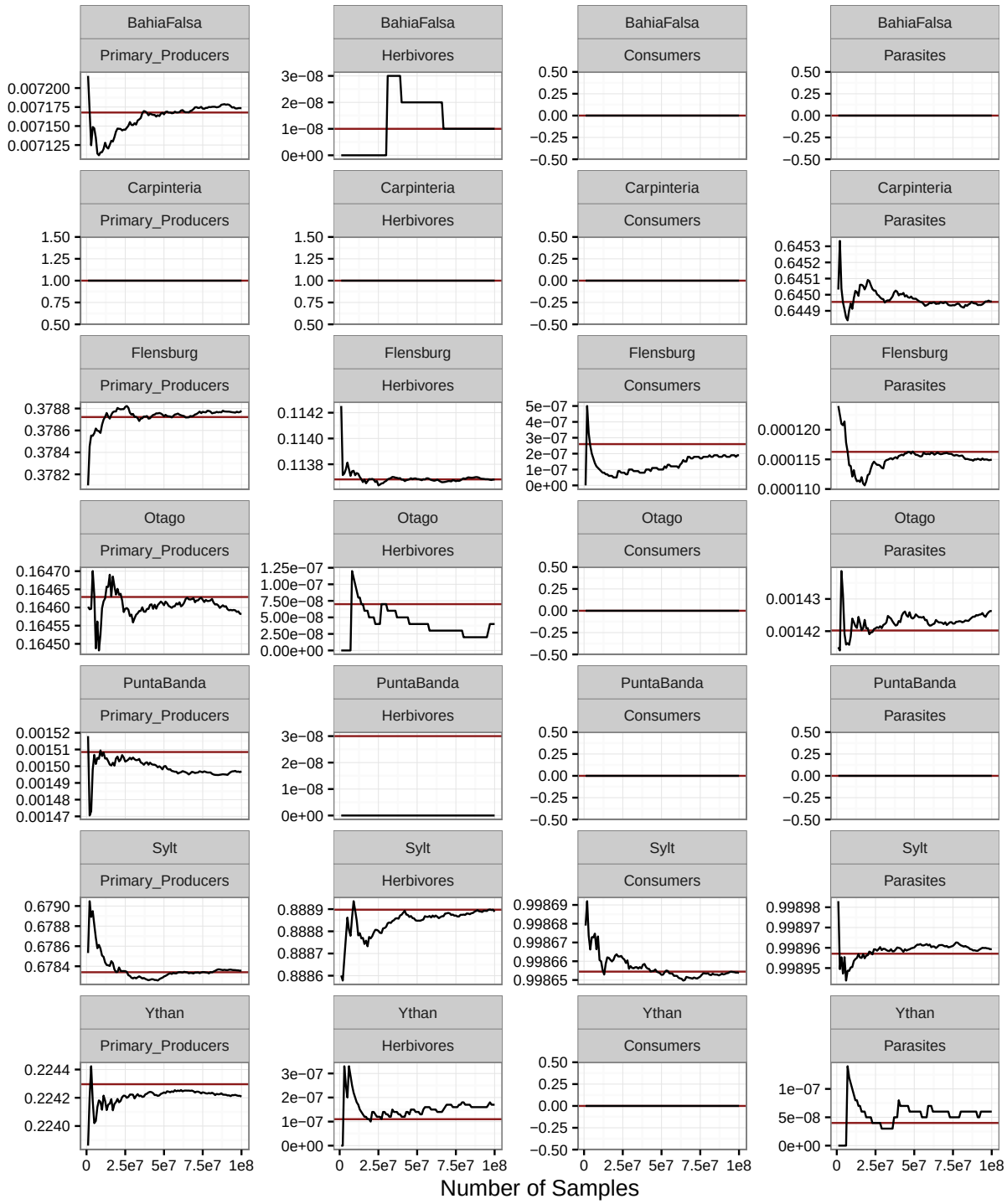

Figure 7.23, continued

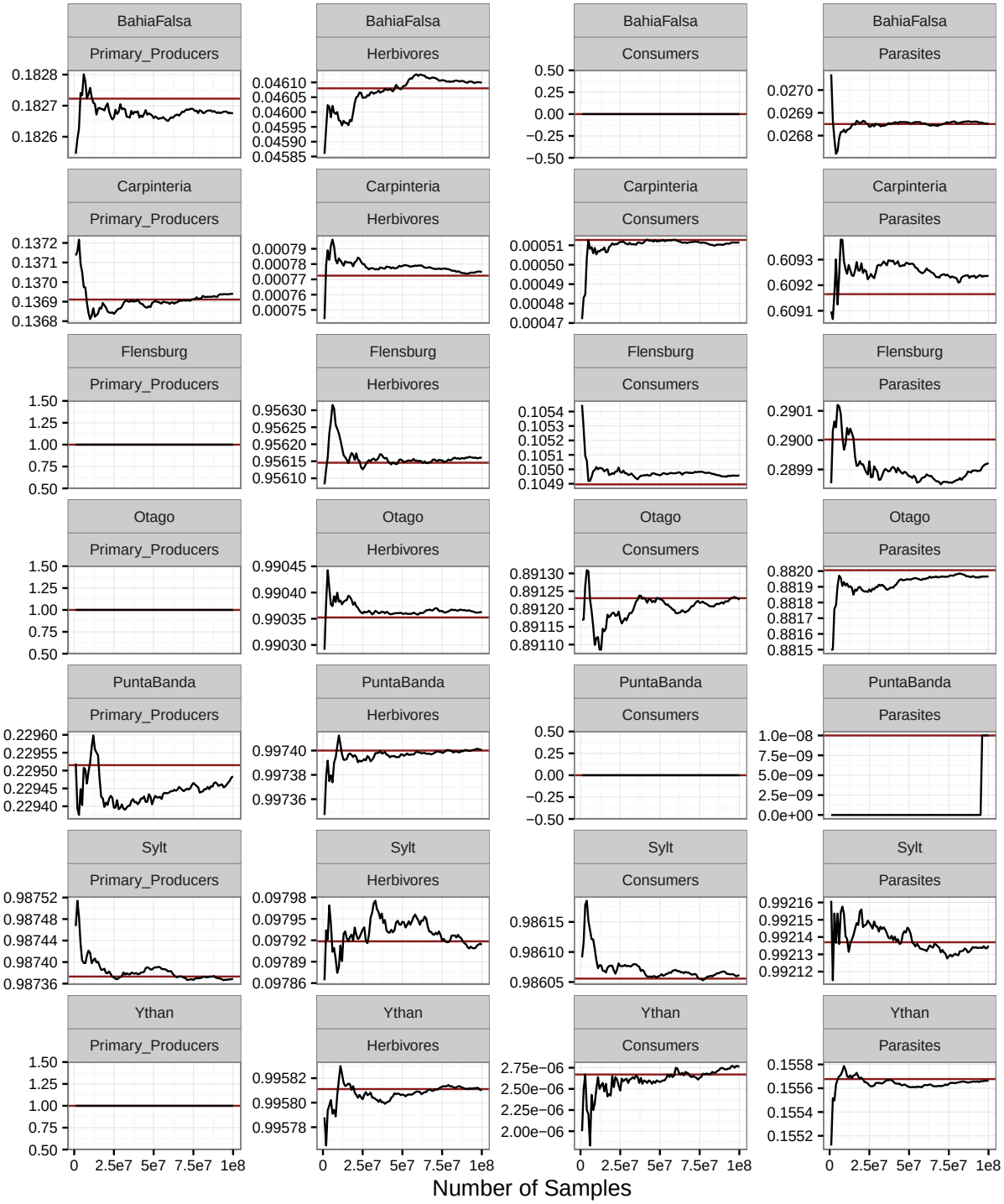

Figure 7.23, continued

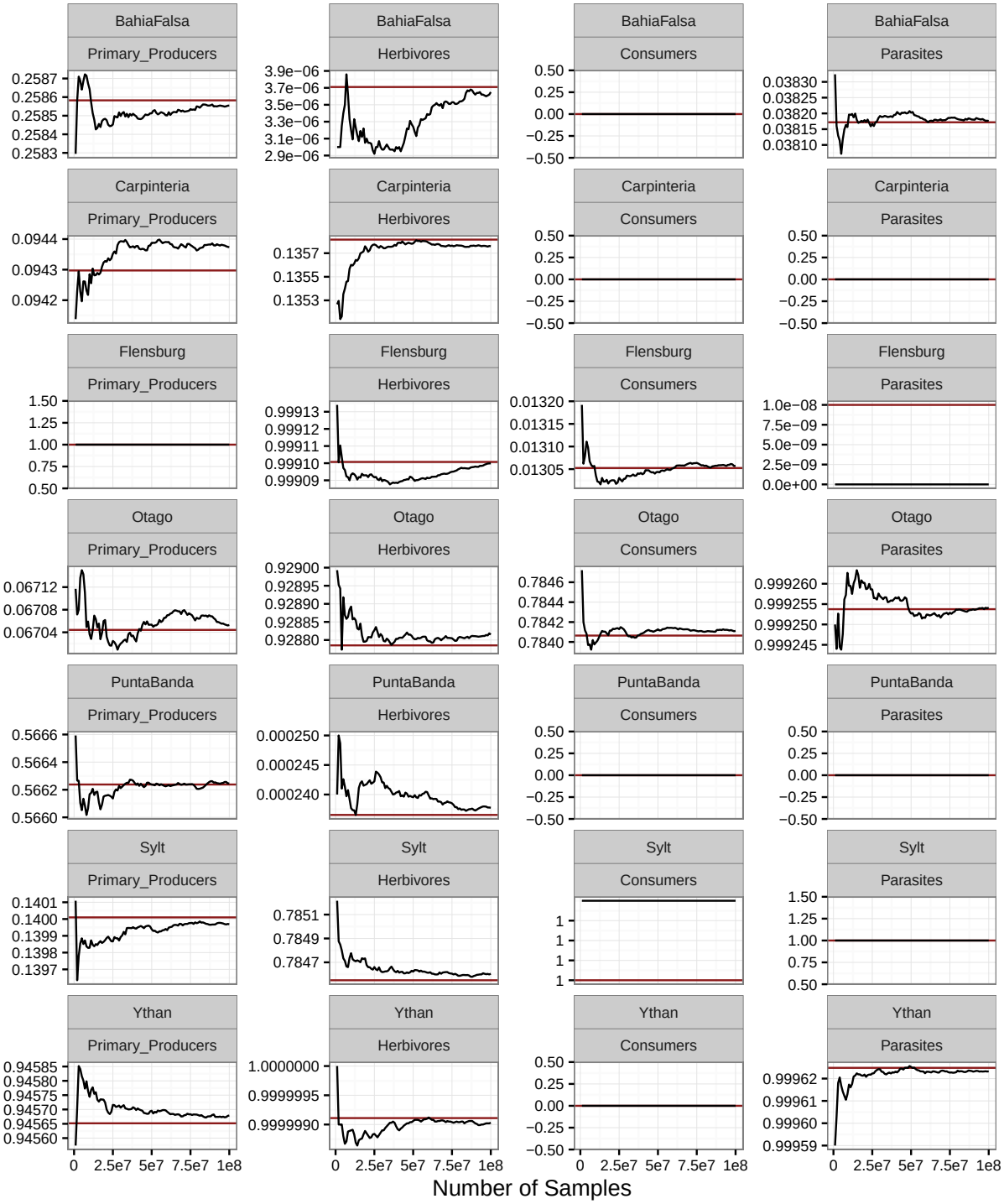

Figure 7.23, continued

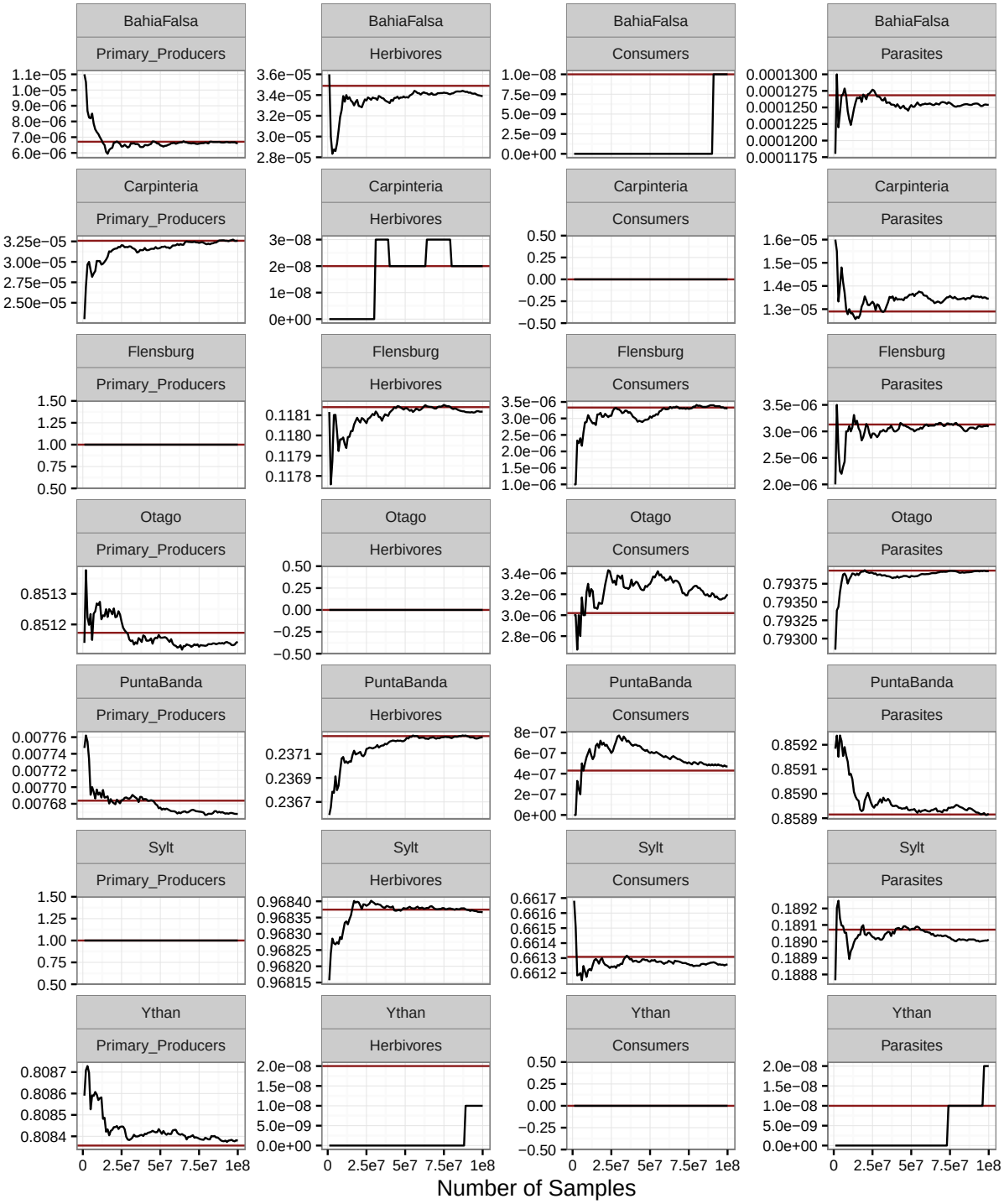

Figure 7.23, continued

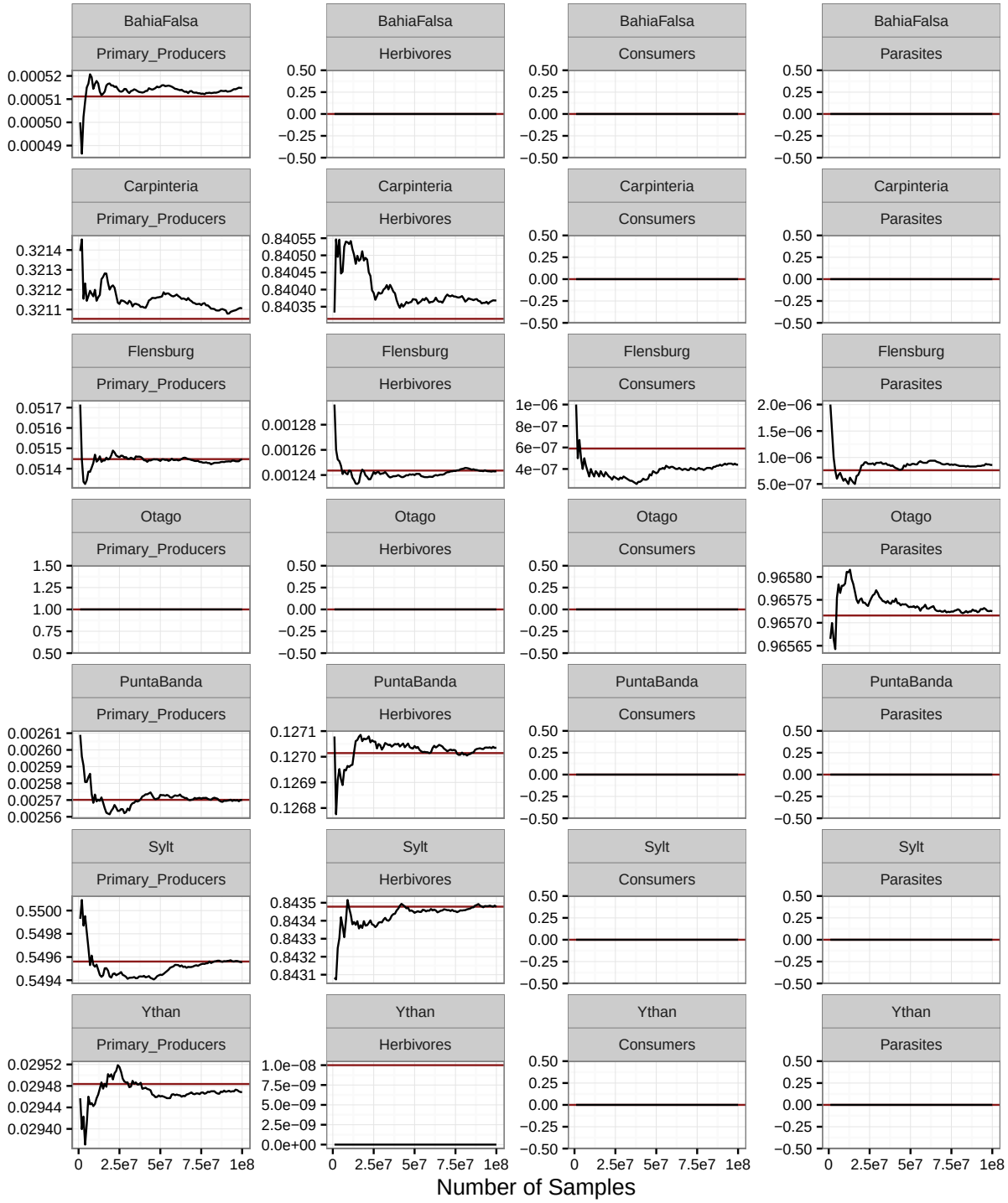

Figure 7.23, continued

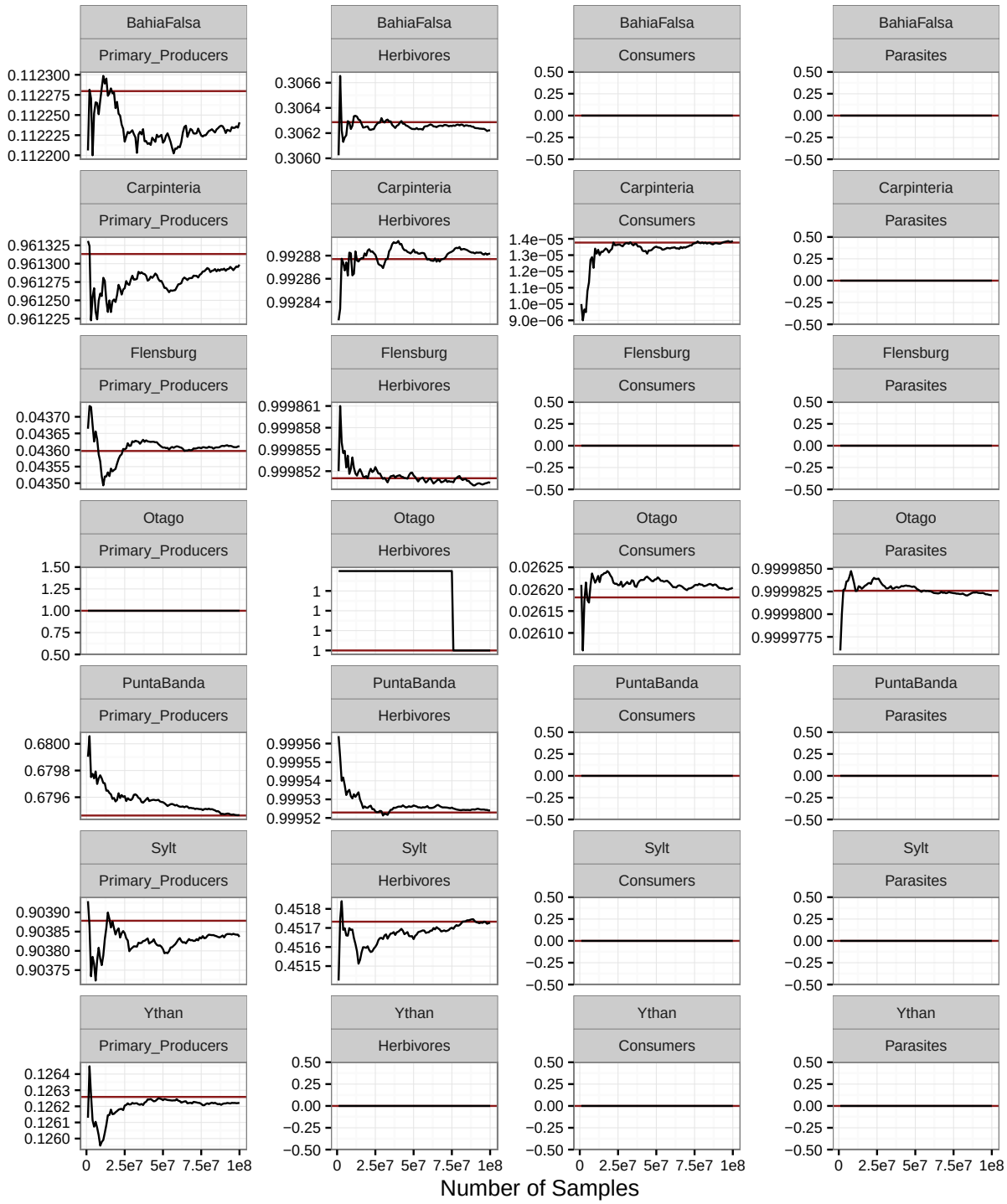

Figure 7.23, continued

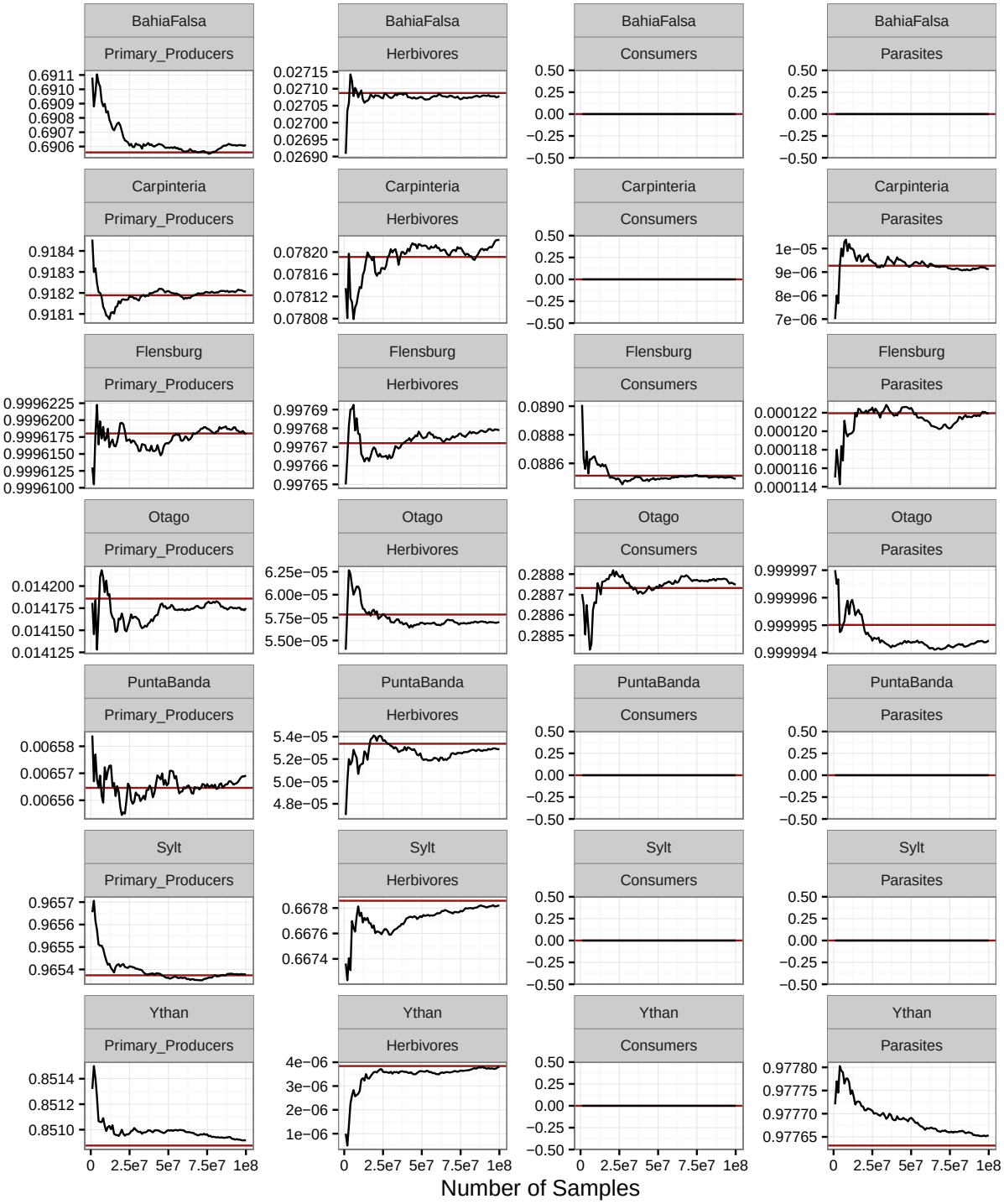

Figure 7.23, continued

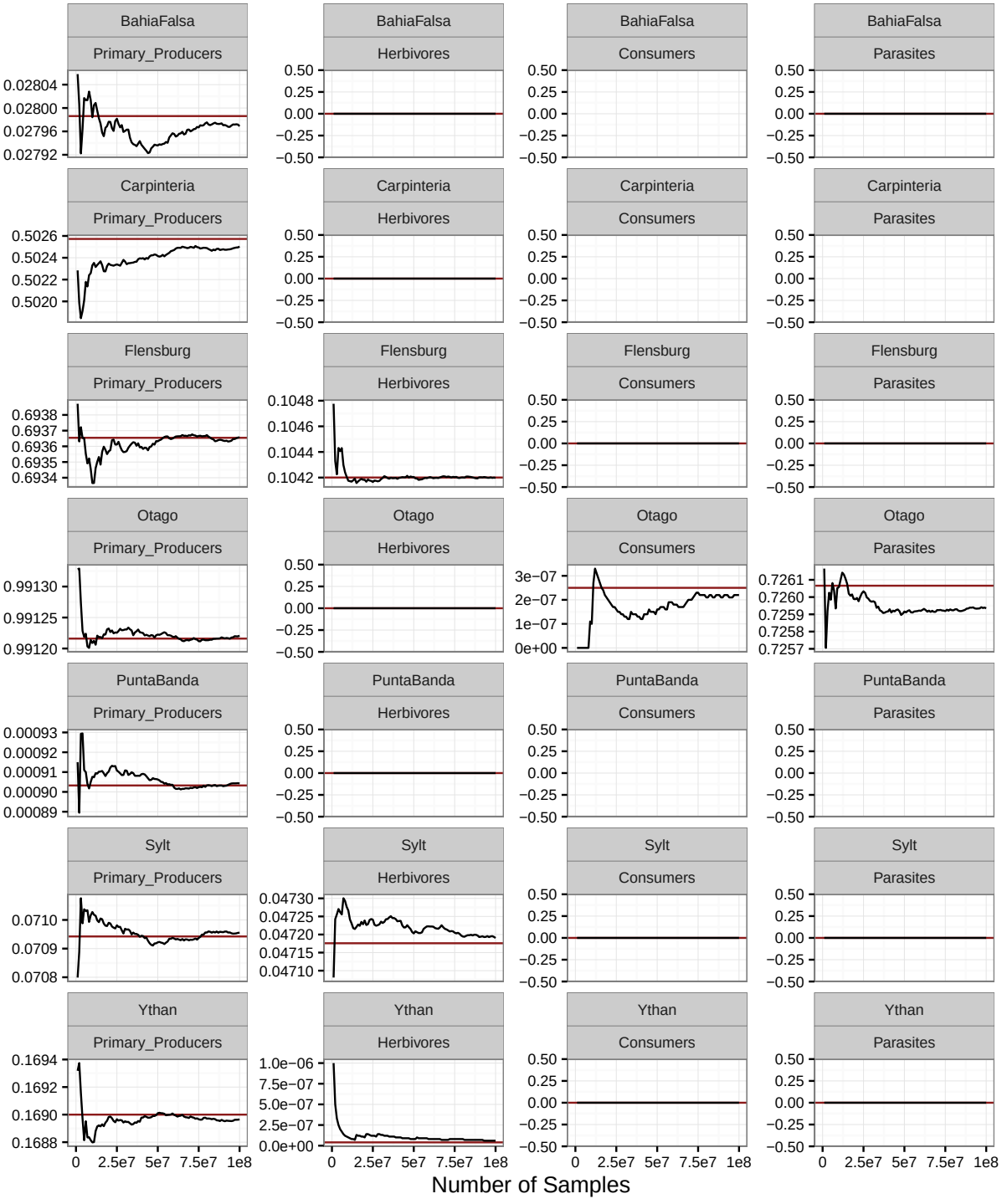

Figure 7.23, continued

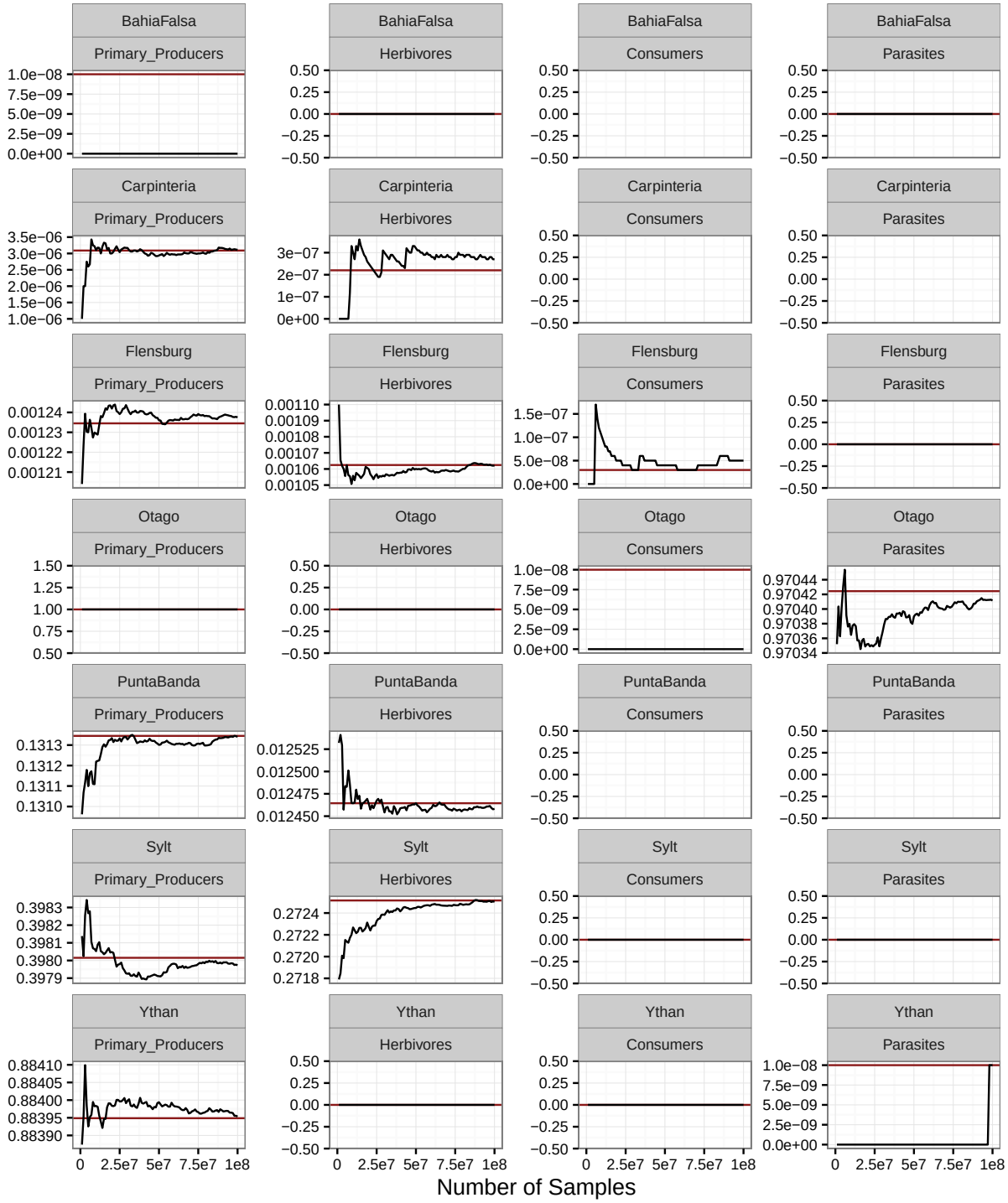

Figure 7.23, continued

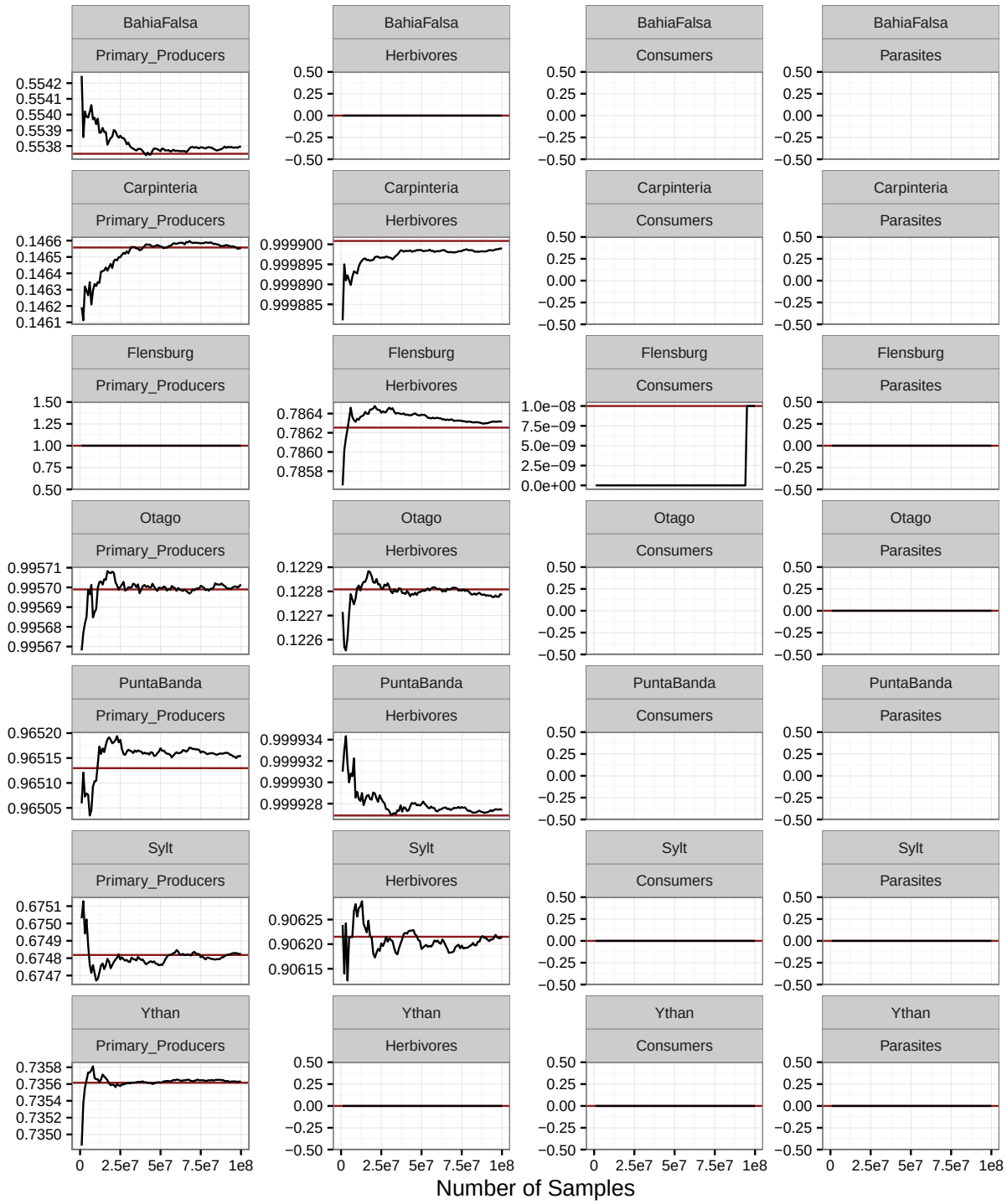

Figure 7.23, continued

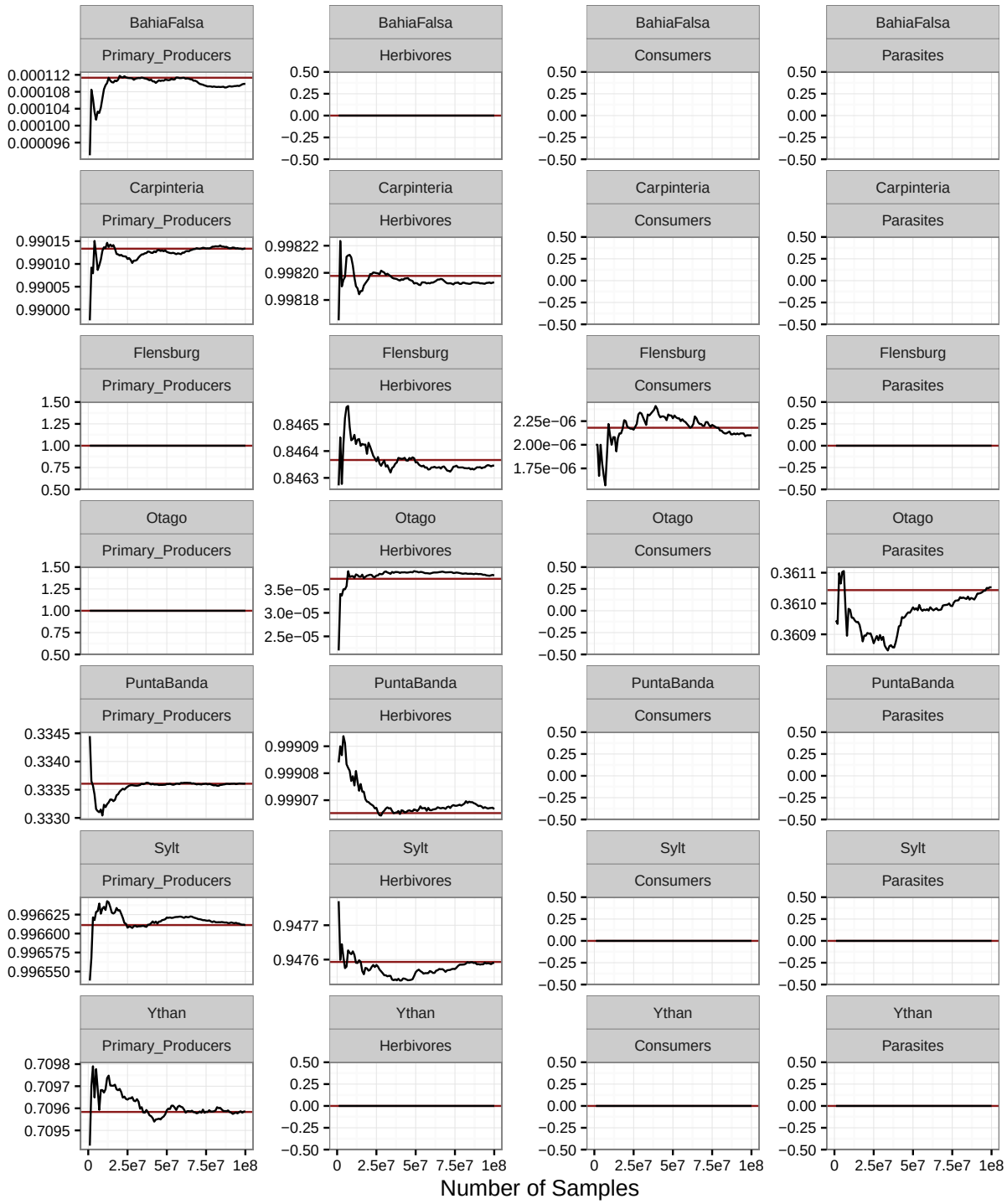

Figure 7.23, continued

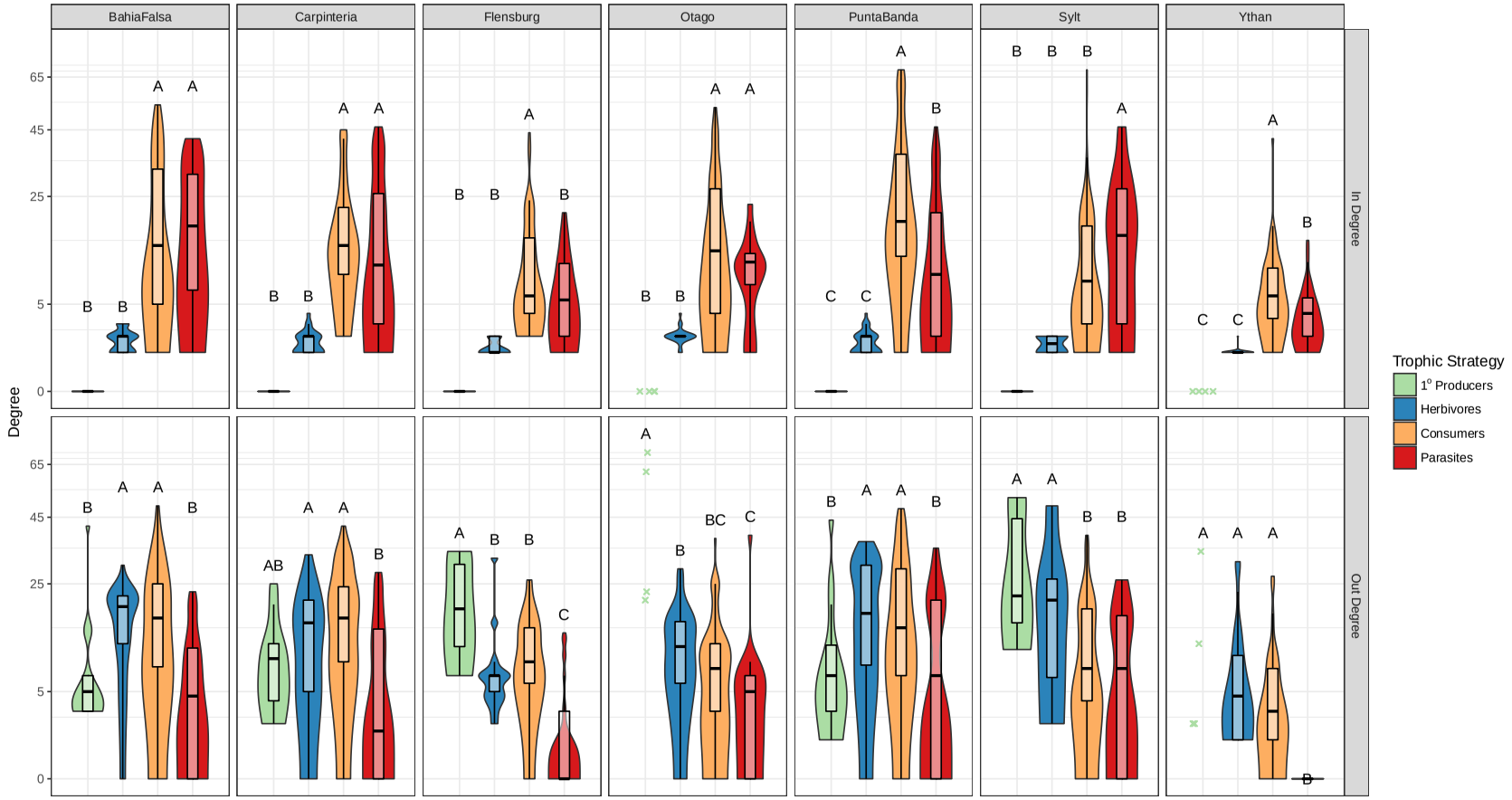

Figure 7.24: Violin and boxplots of in-degree (number of prey) and out-degree (number of predators) for different trophic strategies in all networks *excluding* concomitant predation. Degree is plotted on a square root scale. Boxes indicate the traditional 25^{th} , 50^{th}, and 75^{th} quartiles, with wiskers extending to 1.5 times the inter-quartile range. Above each violin are grouping letters as indicated by a Tukeys HSD (honest significant difference) test.

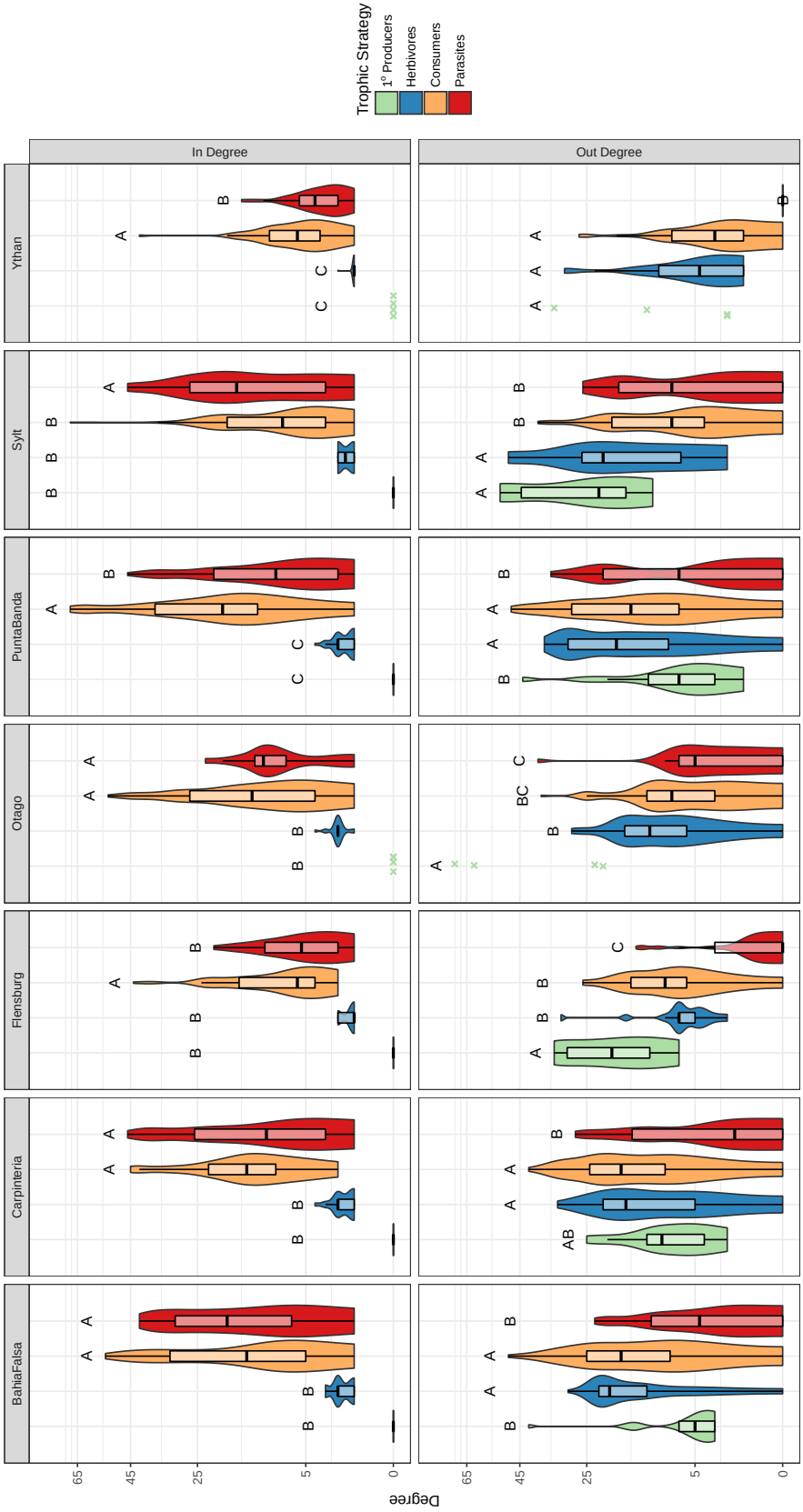

Figure 7.25: As Figure 7.24 but for networks *including* concomitant predation.

7.15 Condensed Network Diagrams

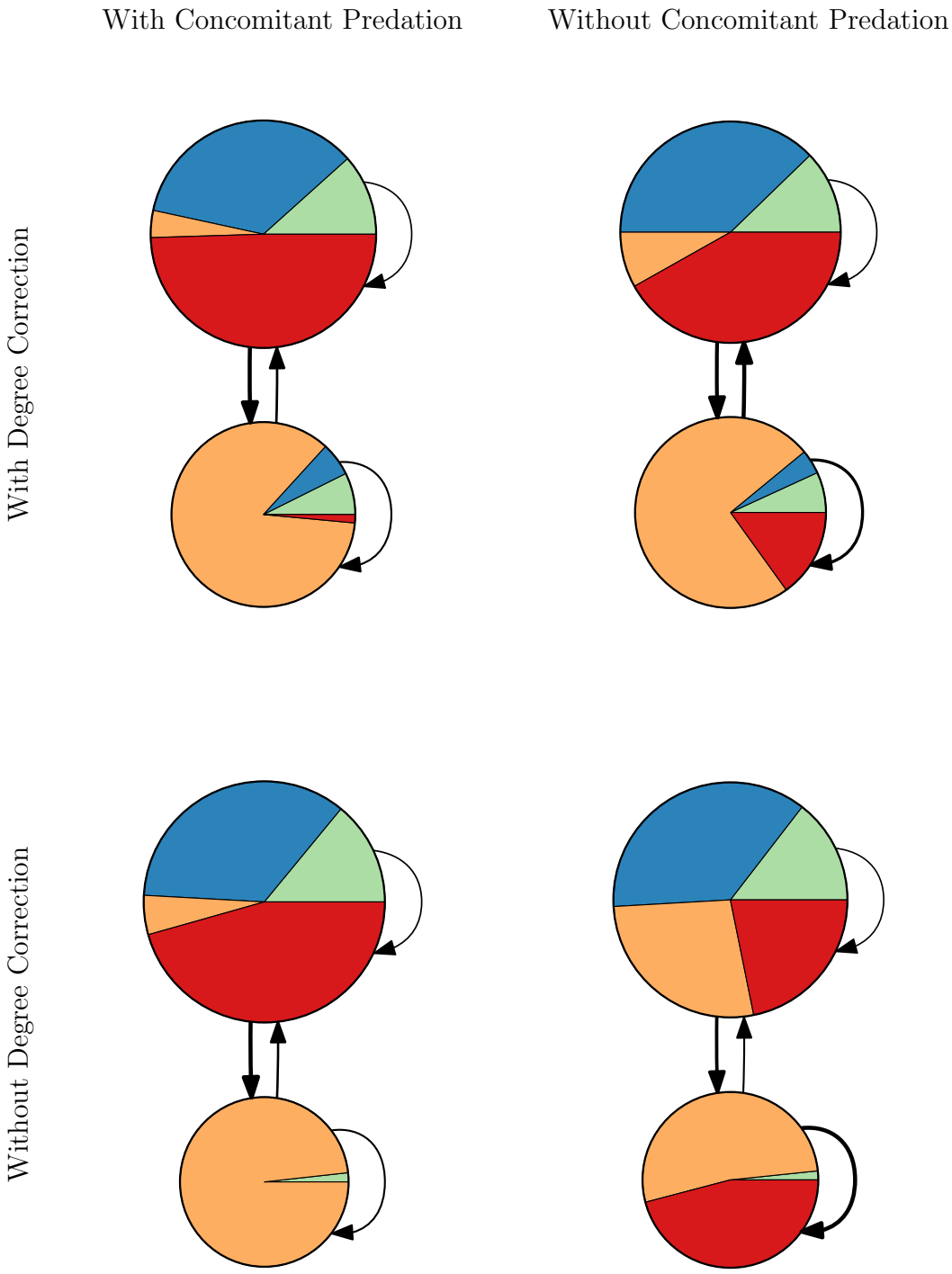

Figure 7.26: As Figure 3.3, but for the BahiaFalsa network and $g = 2$.

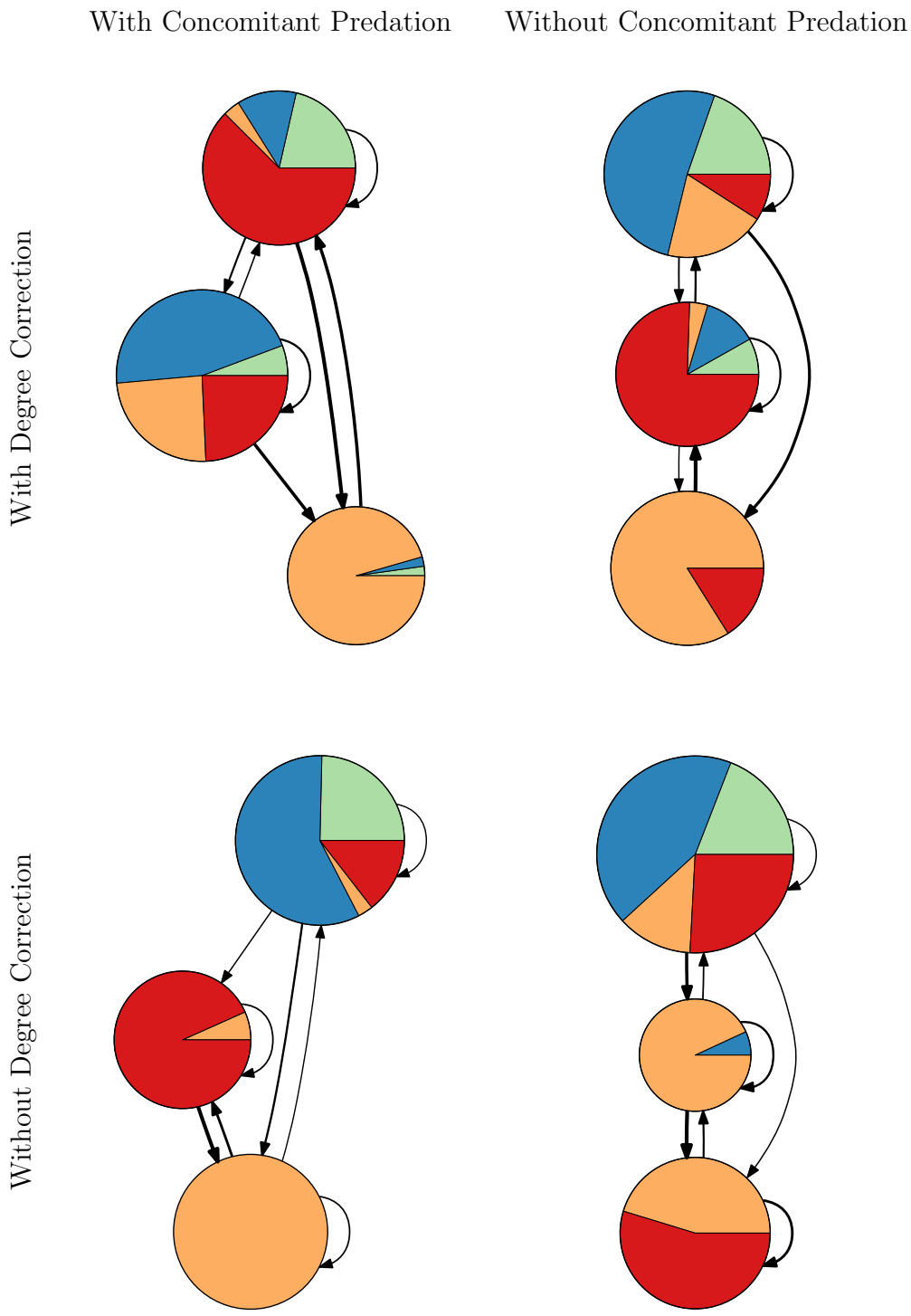

Figure 7.27: As Figure 3.3, but for the BahiaFalsa network and $g = 3$.

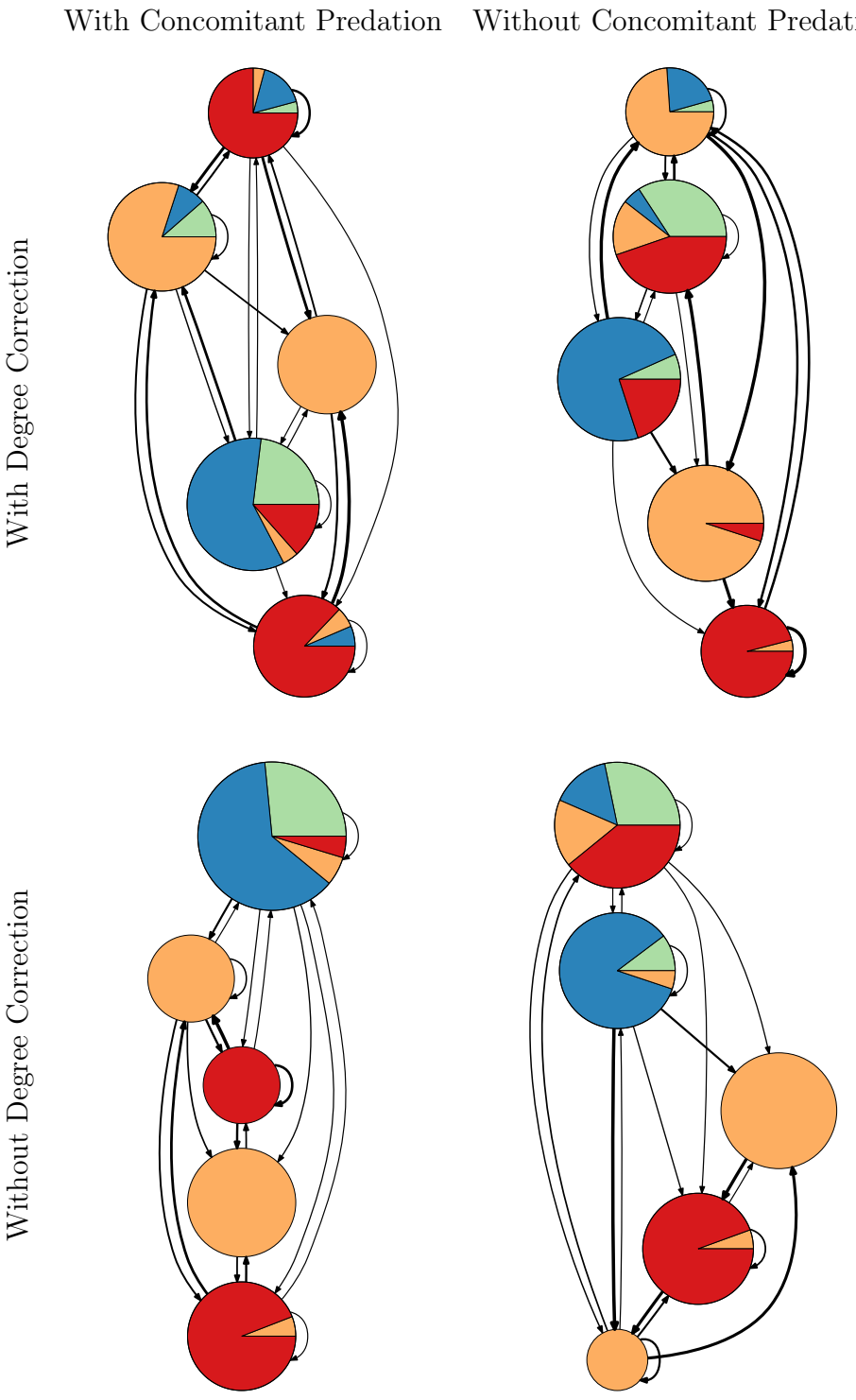

Figure 7.28: As Figure 3.3, but for the BahiaFalsa network and $g = 5$.

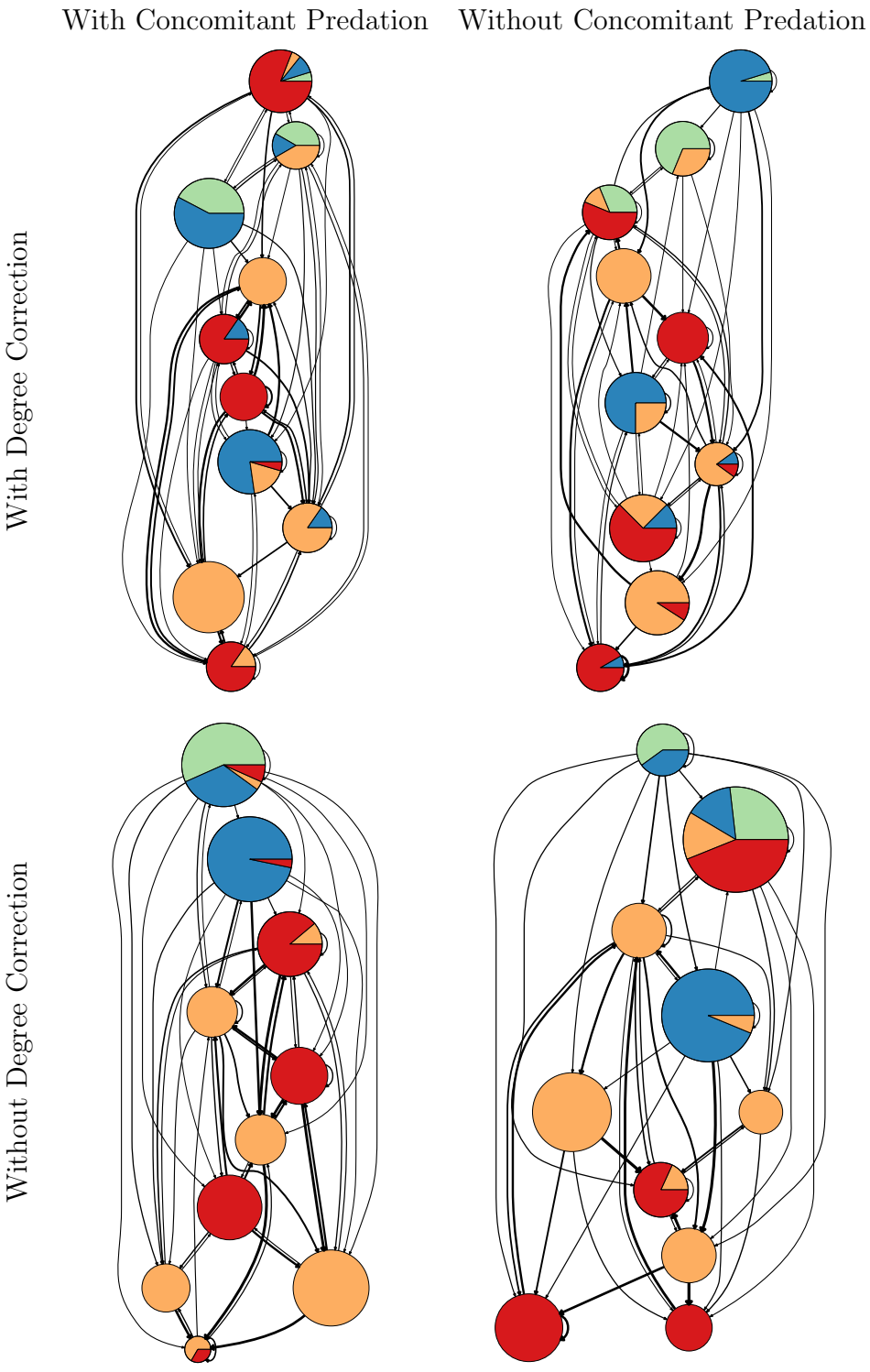

Figure 7.29: As Figure 3.3, but for the BahiaFalsa network and $g = 10$.

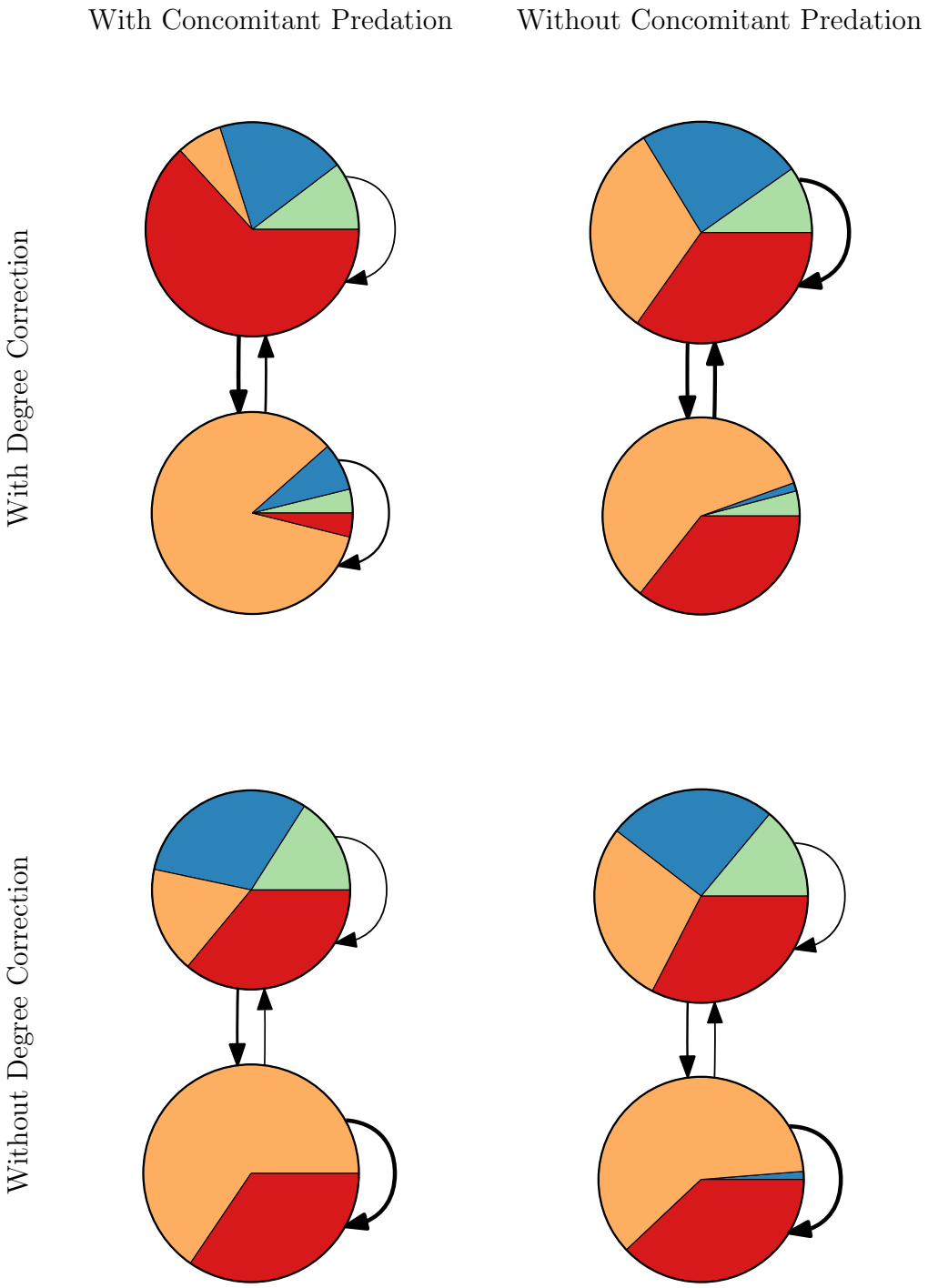

Figure 7.30: As Figure 3.3, but for the Carpinteria network and $g = 2$.

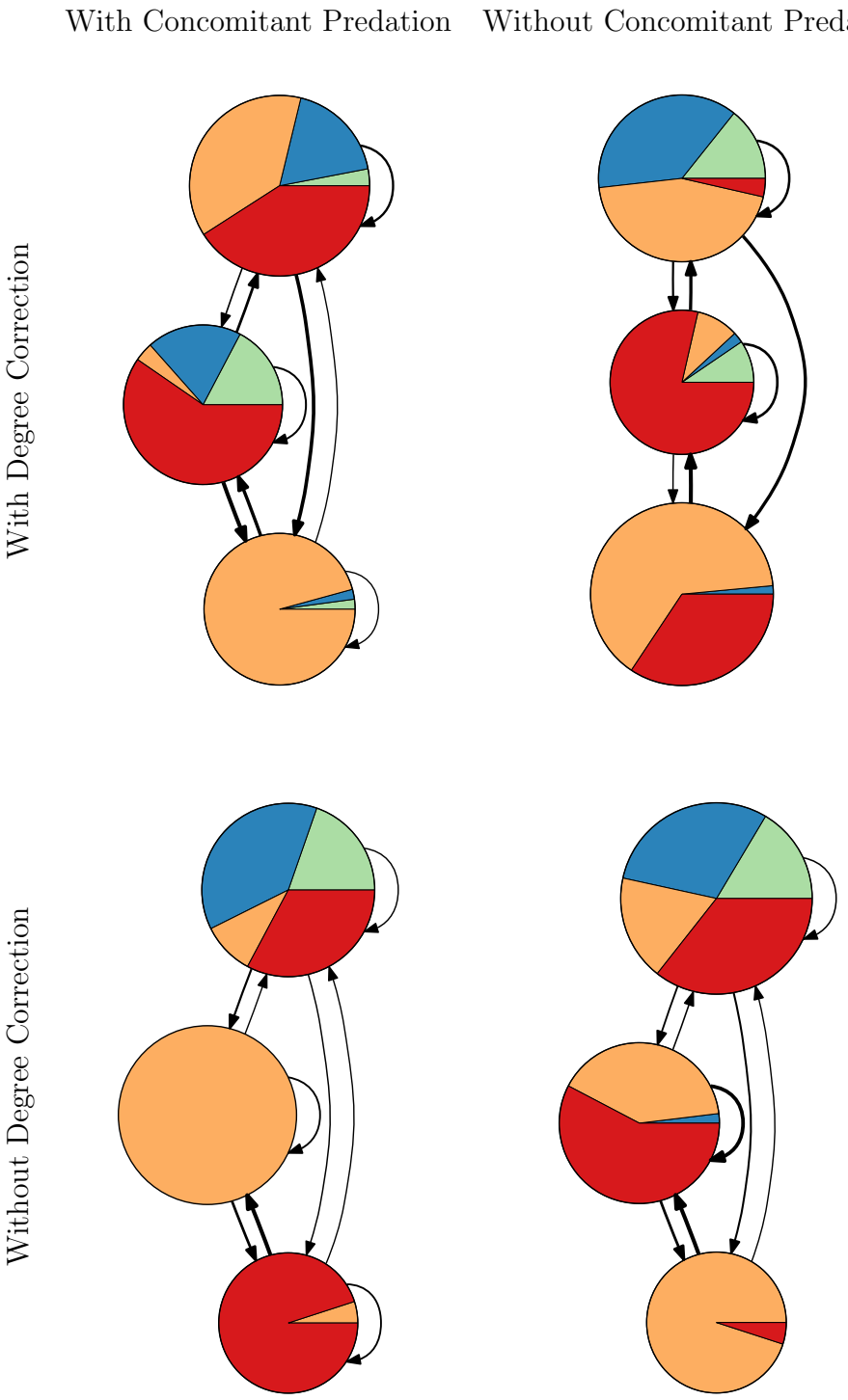

Figure 7.31: As Figure 3.3, but for the Carpinteria network and $g = 3$.

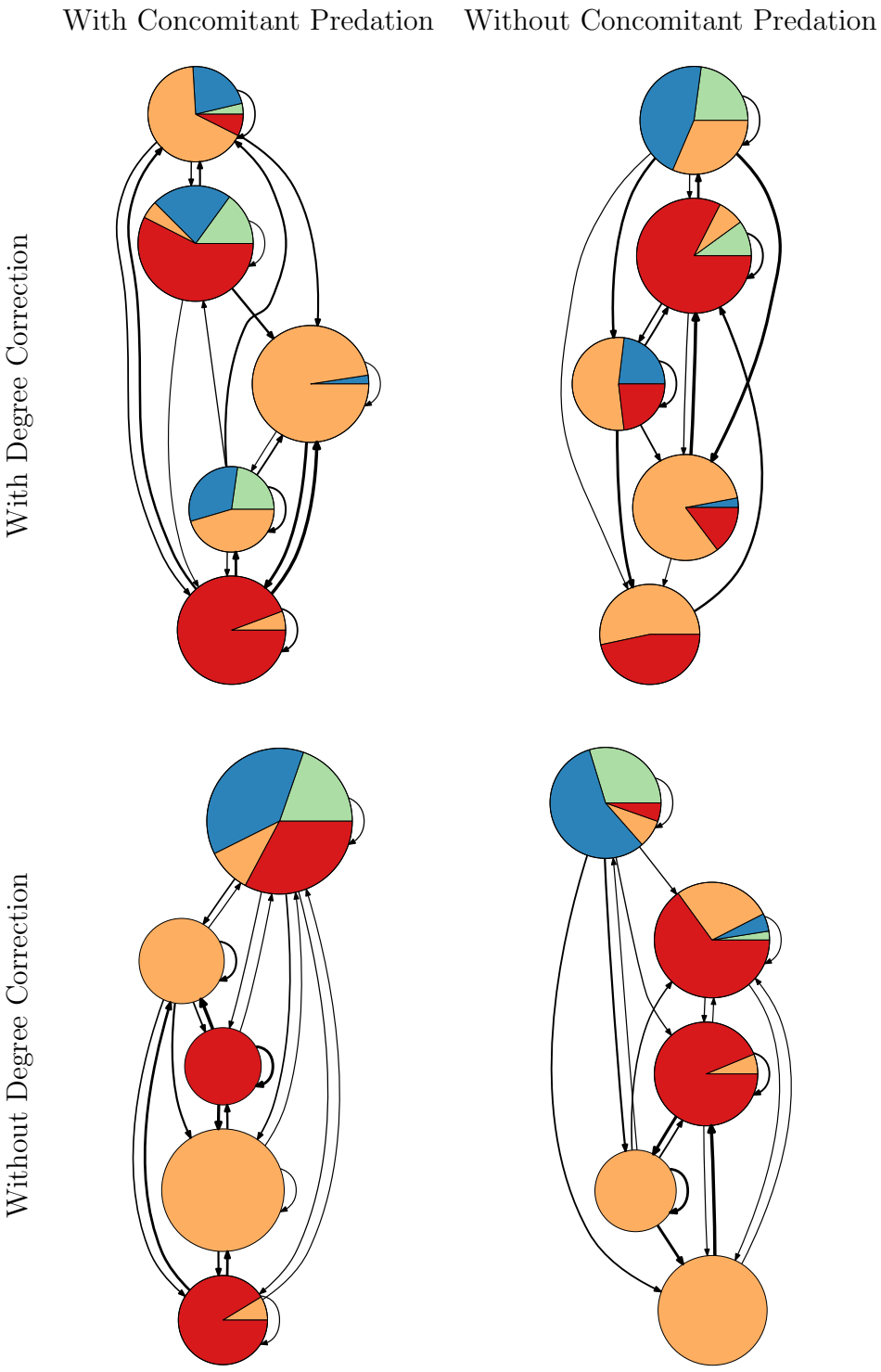

Figure 7.32: As Figure 3.3, but for the Carpinteria network and $g = 5$.

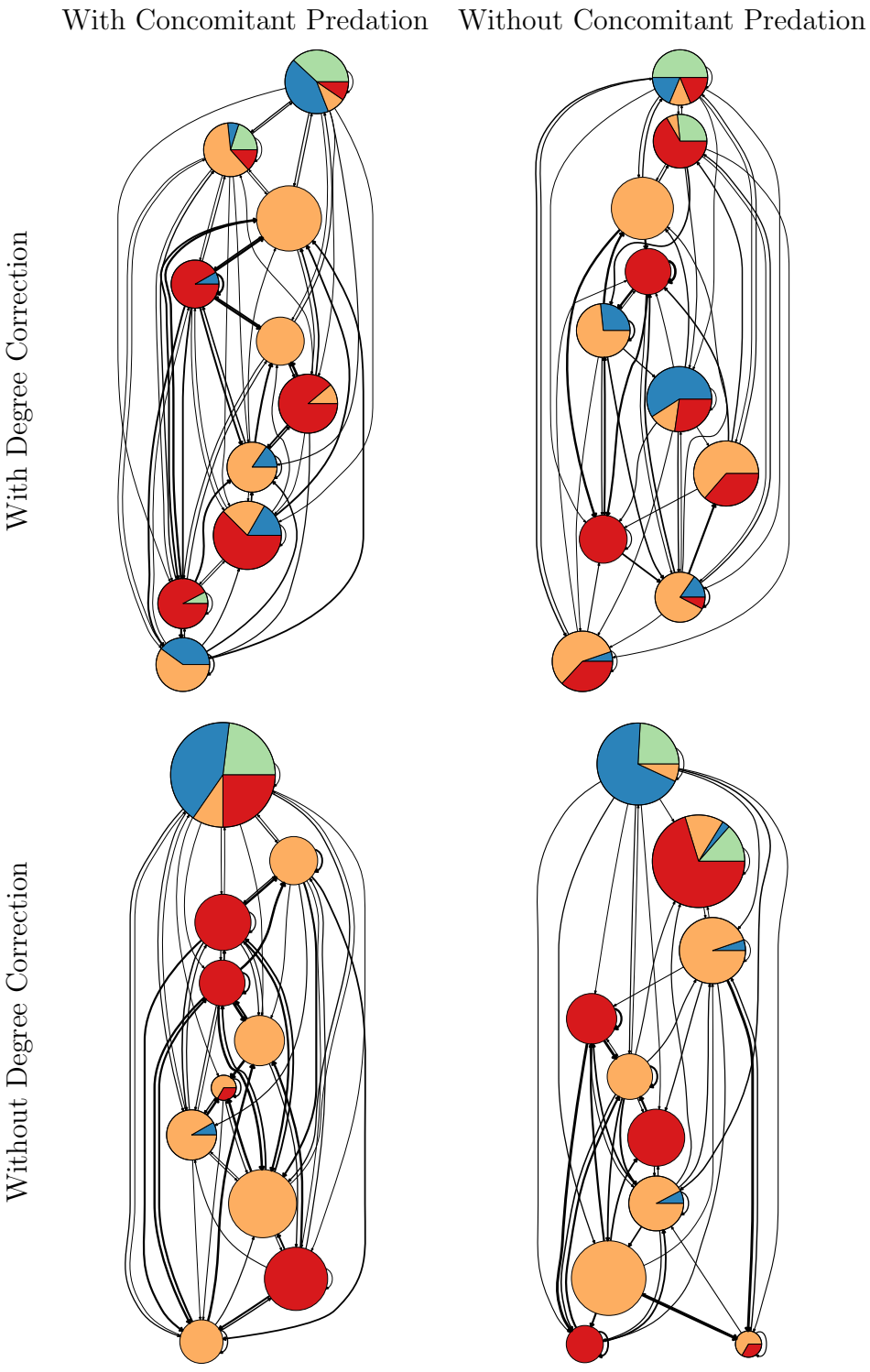

Figure 7.33: As Figure 3.3, but for the Carpinteria network and $g = 10$.

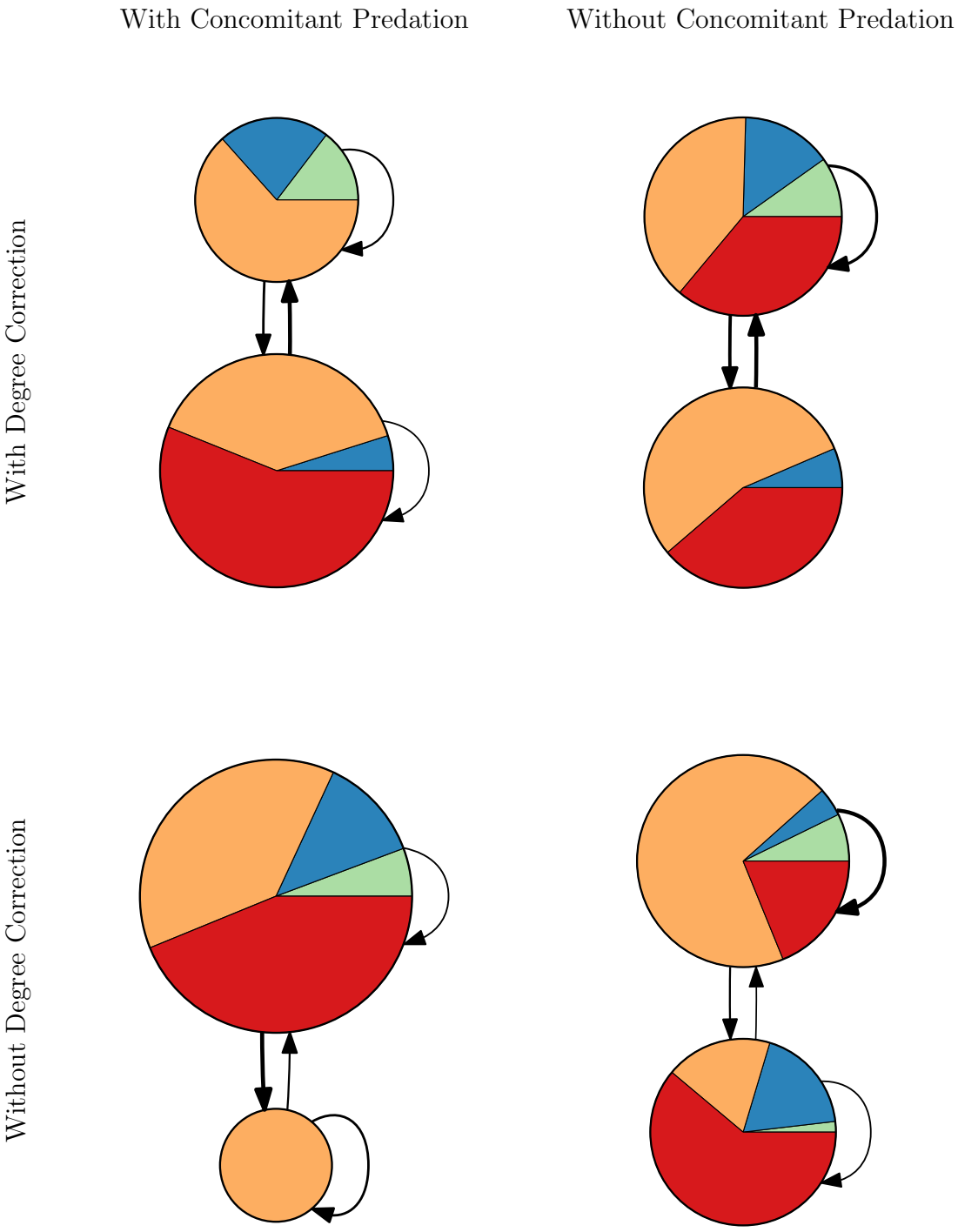

Figure 7.34: As Figure 3.3, but for the Flensburg network and $g = 2$.

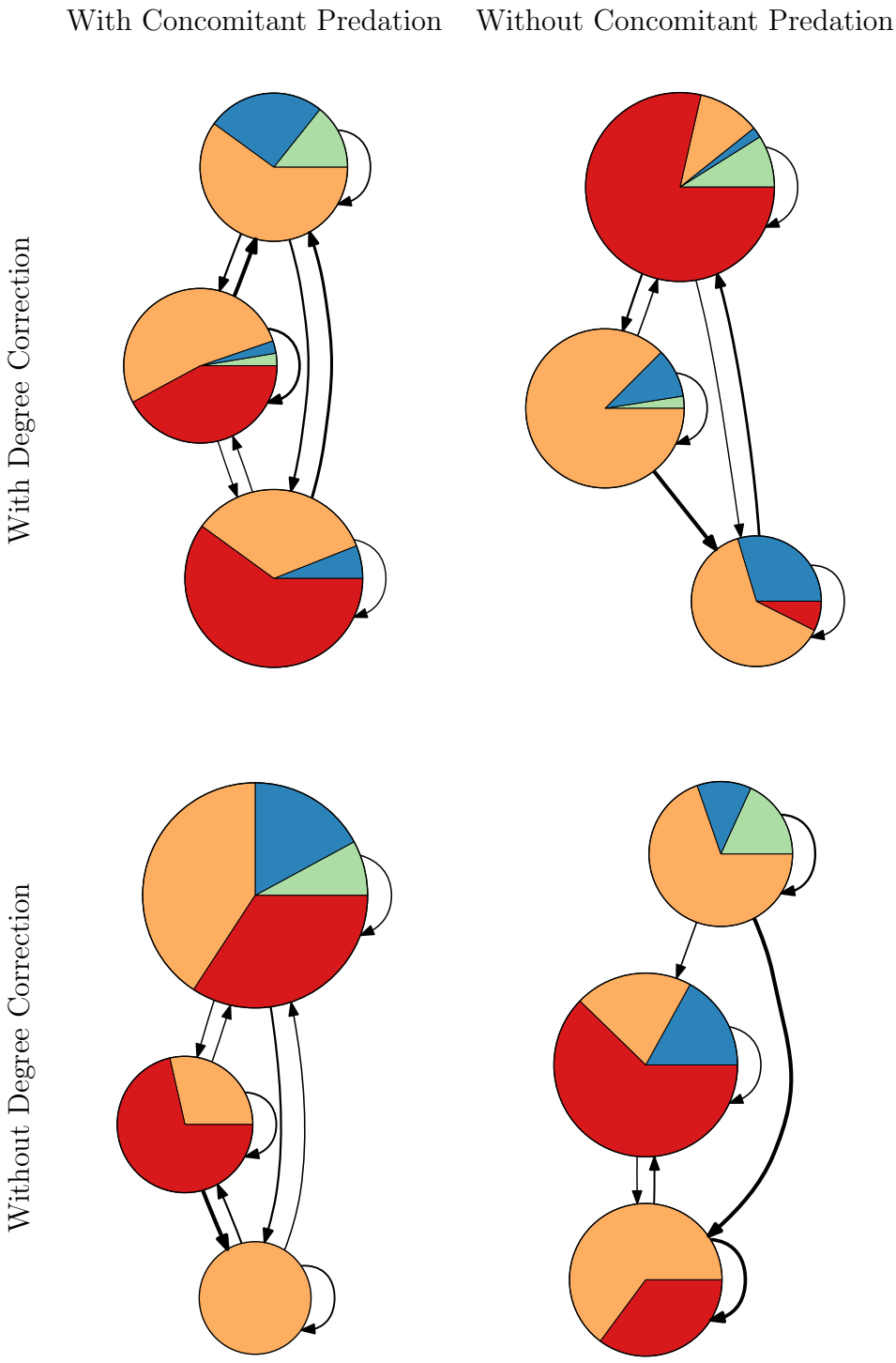

Figure 7.35: As Figure 3.3, but for the Flensburg network and $g = 3$.

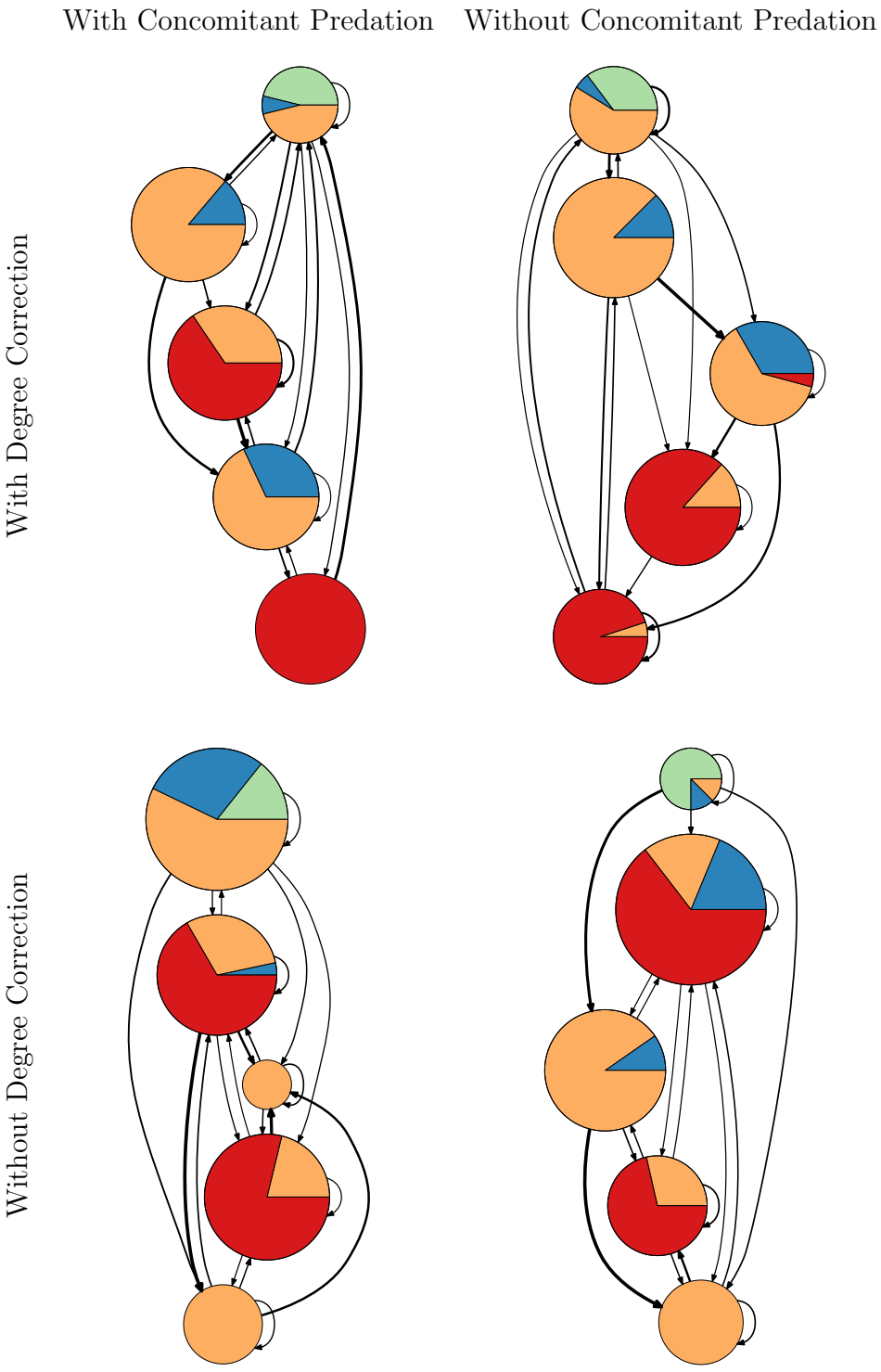

Figure 7.36: As Figure 3.3, but for the Flensburg network and $g = 5$.

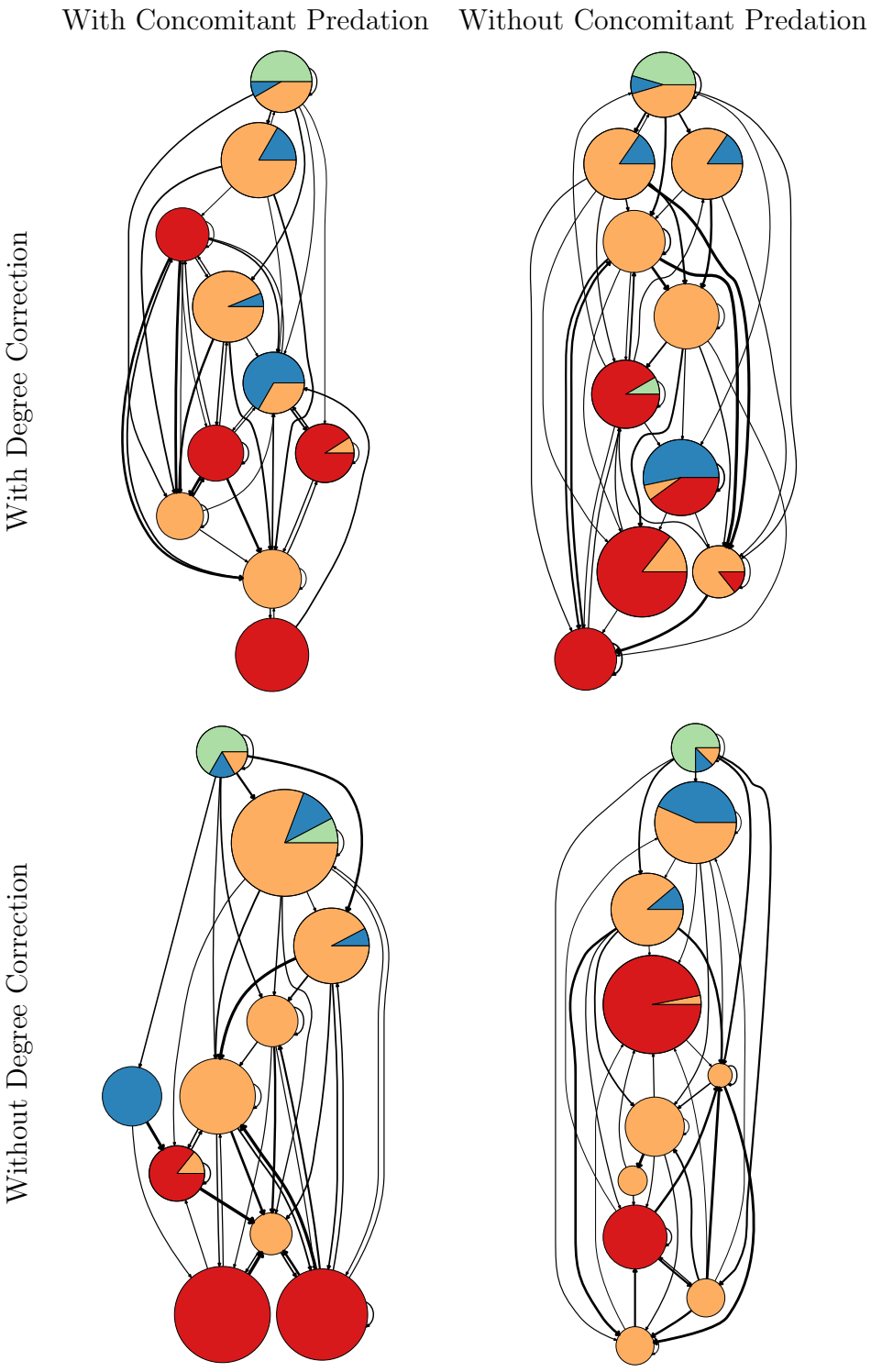

Figure 7.37: As Figure 3.3, but for the Flensburg network and $g = 10$.

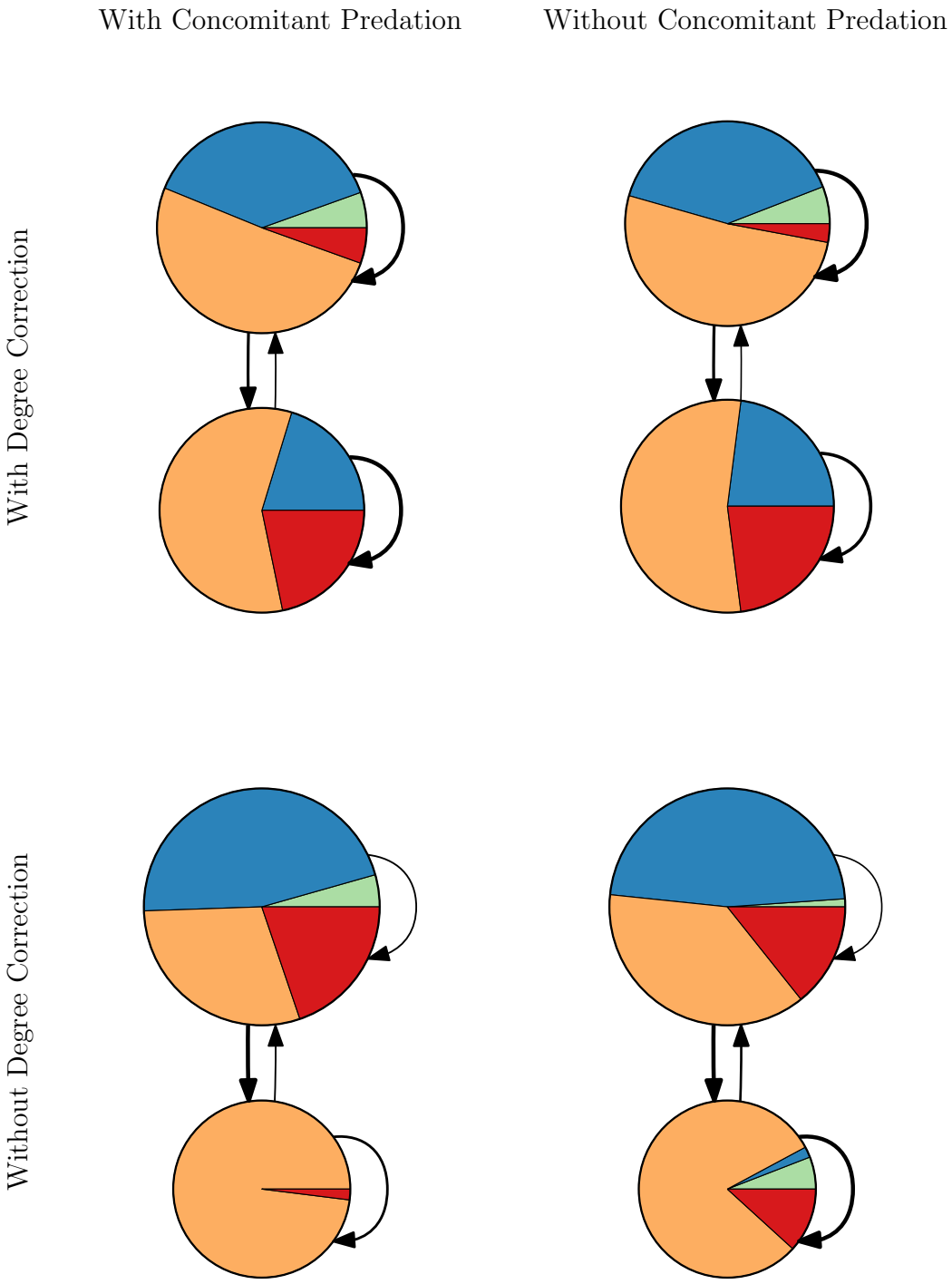

Figure 7.38: As Figure 3.3, but for the Otago network and $g = 2$.

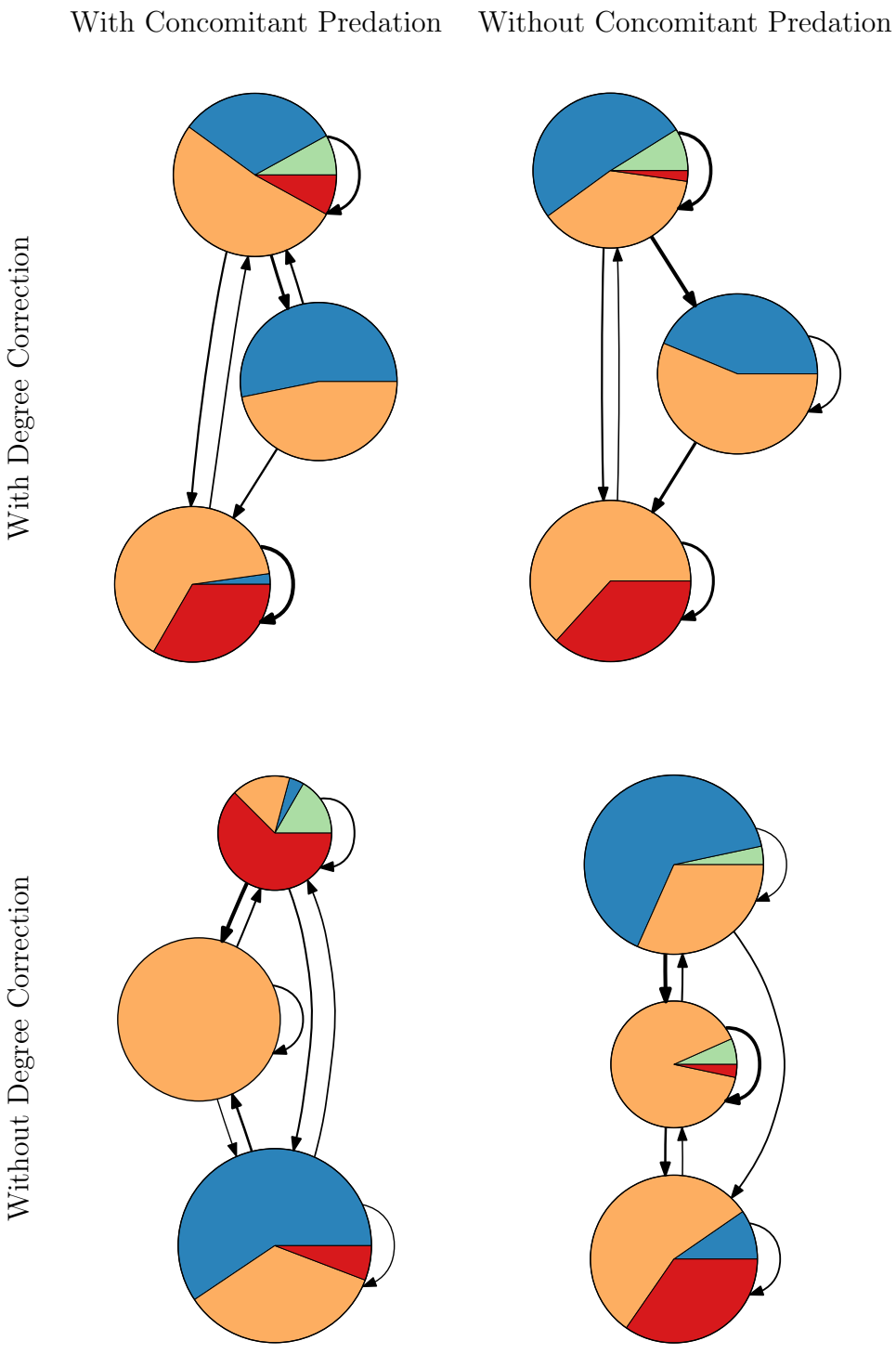

Figure 7.39: As Figure 3.3, but for the Otago network and $g = 3$.

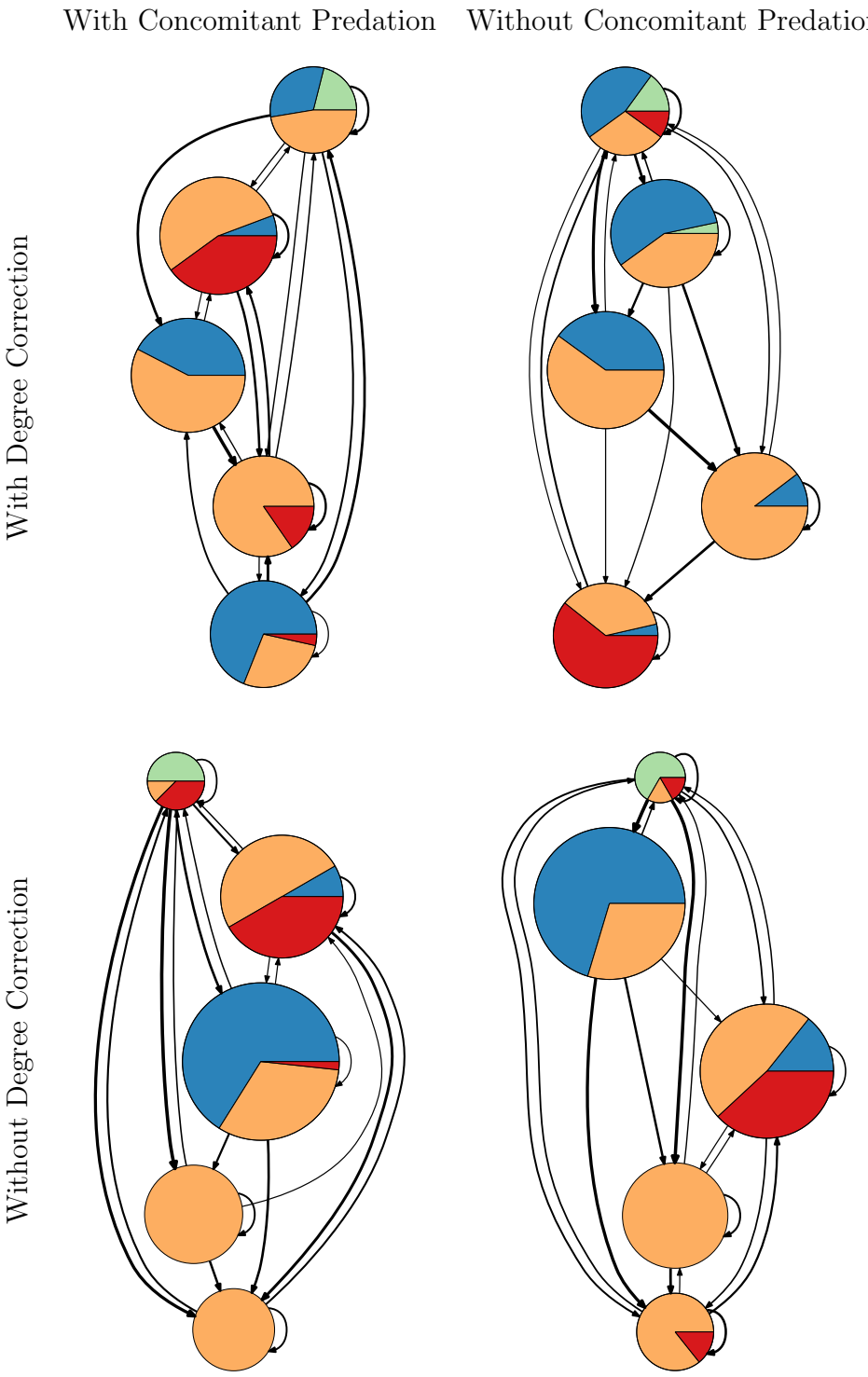

Figure 7.40: As Figure 3.3, but for the Otago network and $g = 5$.

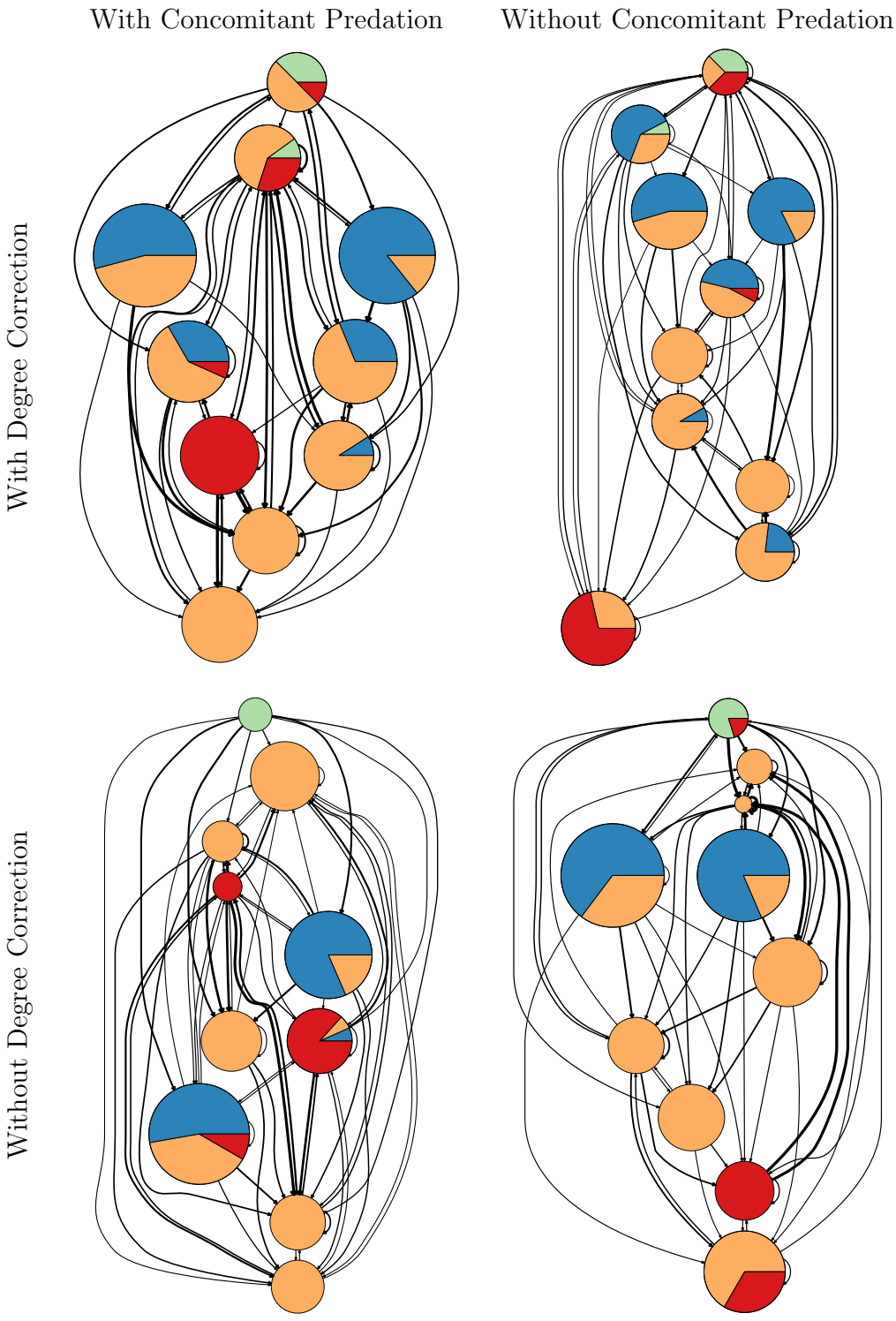

Figure 7.41: As Figure 3.3, but for the Otago network and $g = 10$.

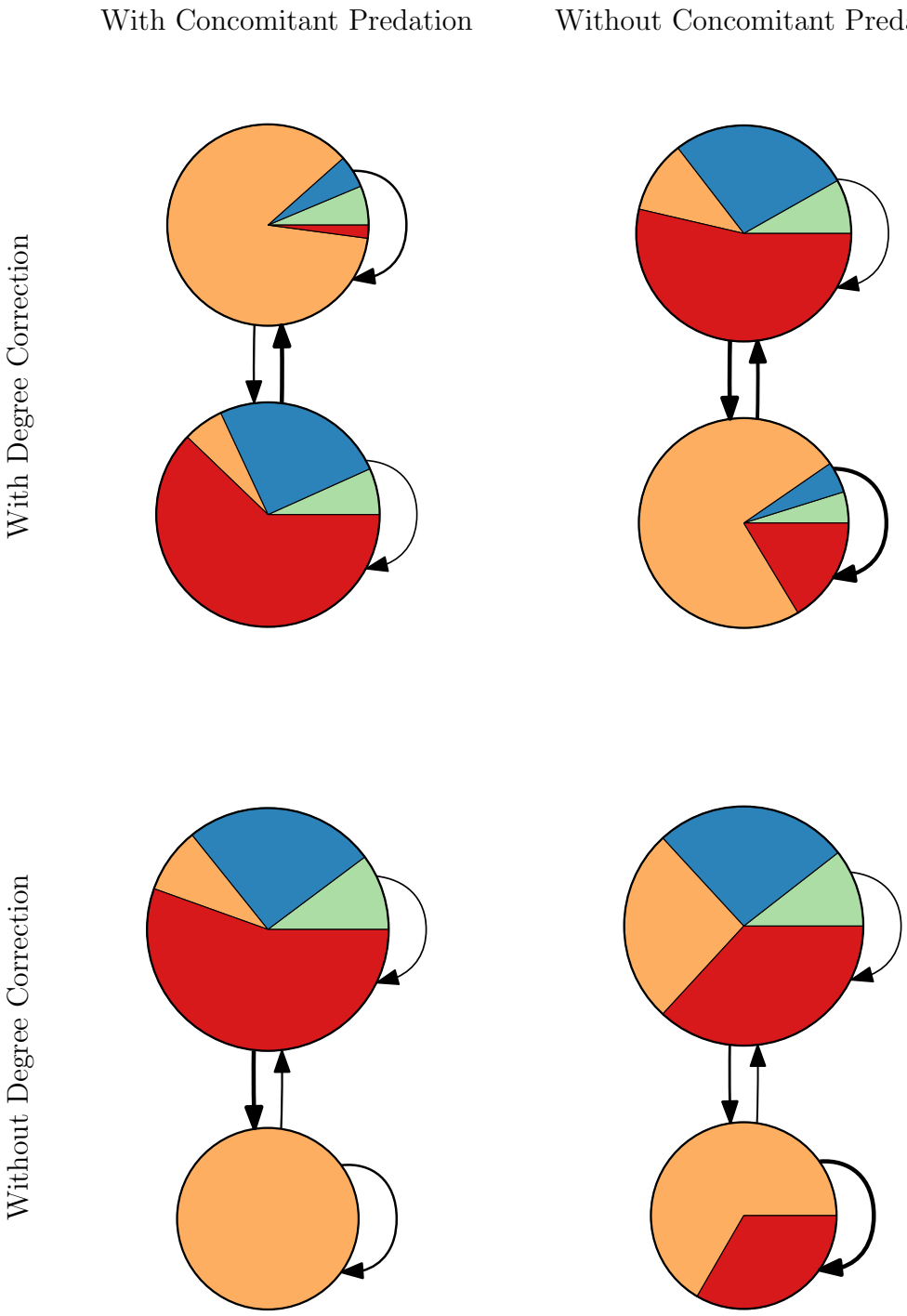

Figure 7.42: As Figure 3.3, but for the PuntaBanda network and $g = 2$.

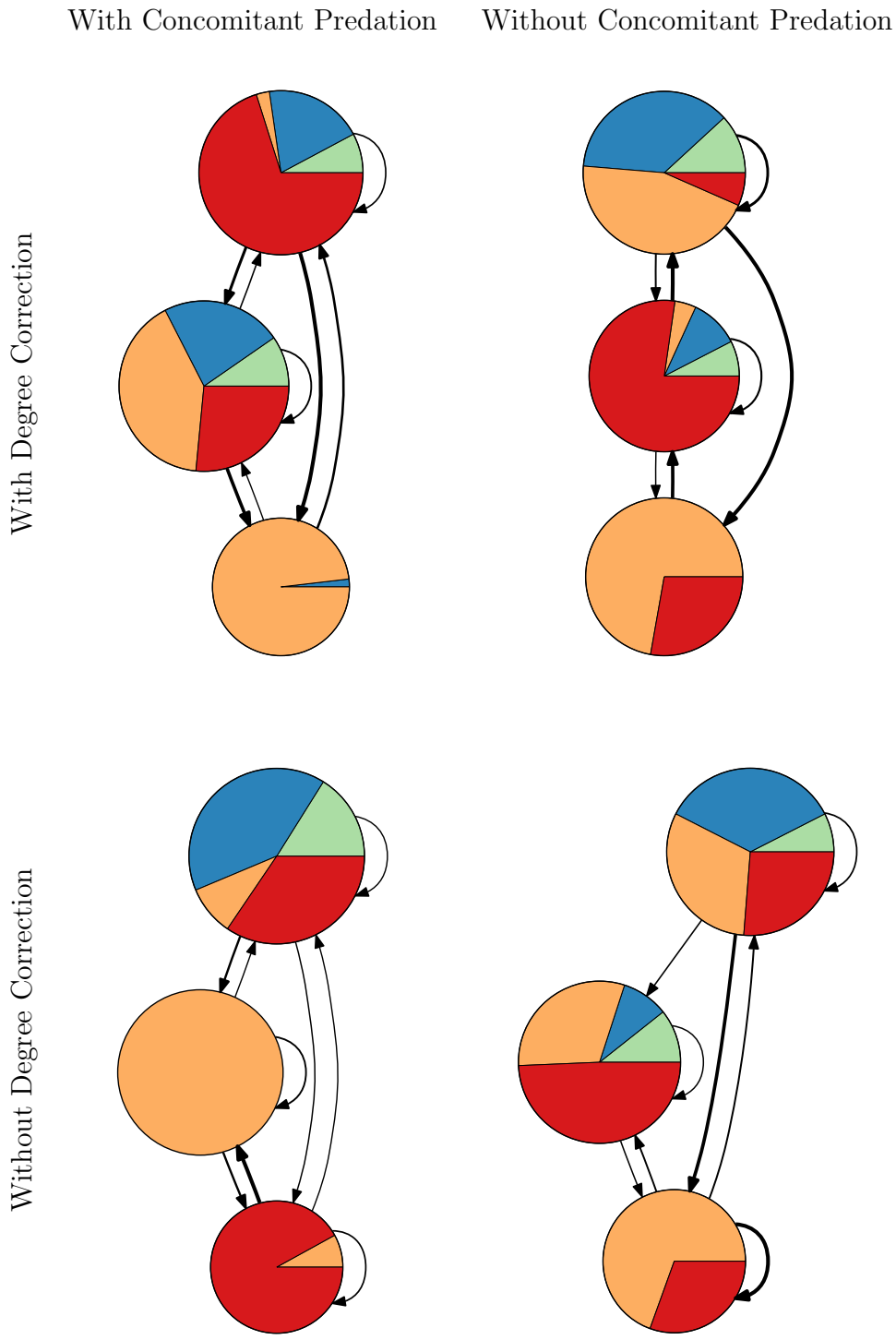

Figure 7.43: As Figure 3.3, but for the PuntaBanda network and $g = 3$.

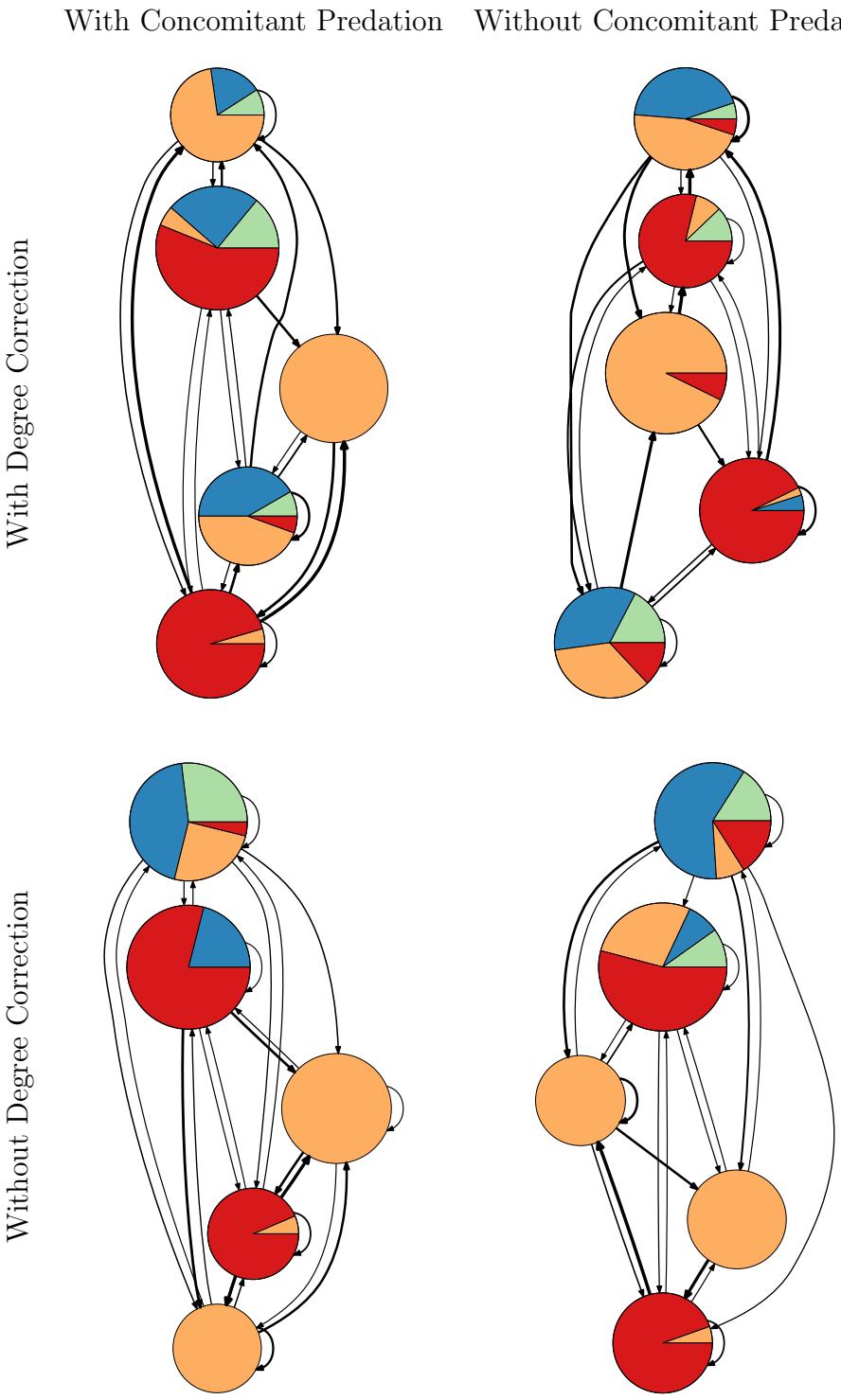

Figure 7.44: As Figure 3.3, but for the PuntaBanda network and $g = 5$.

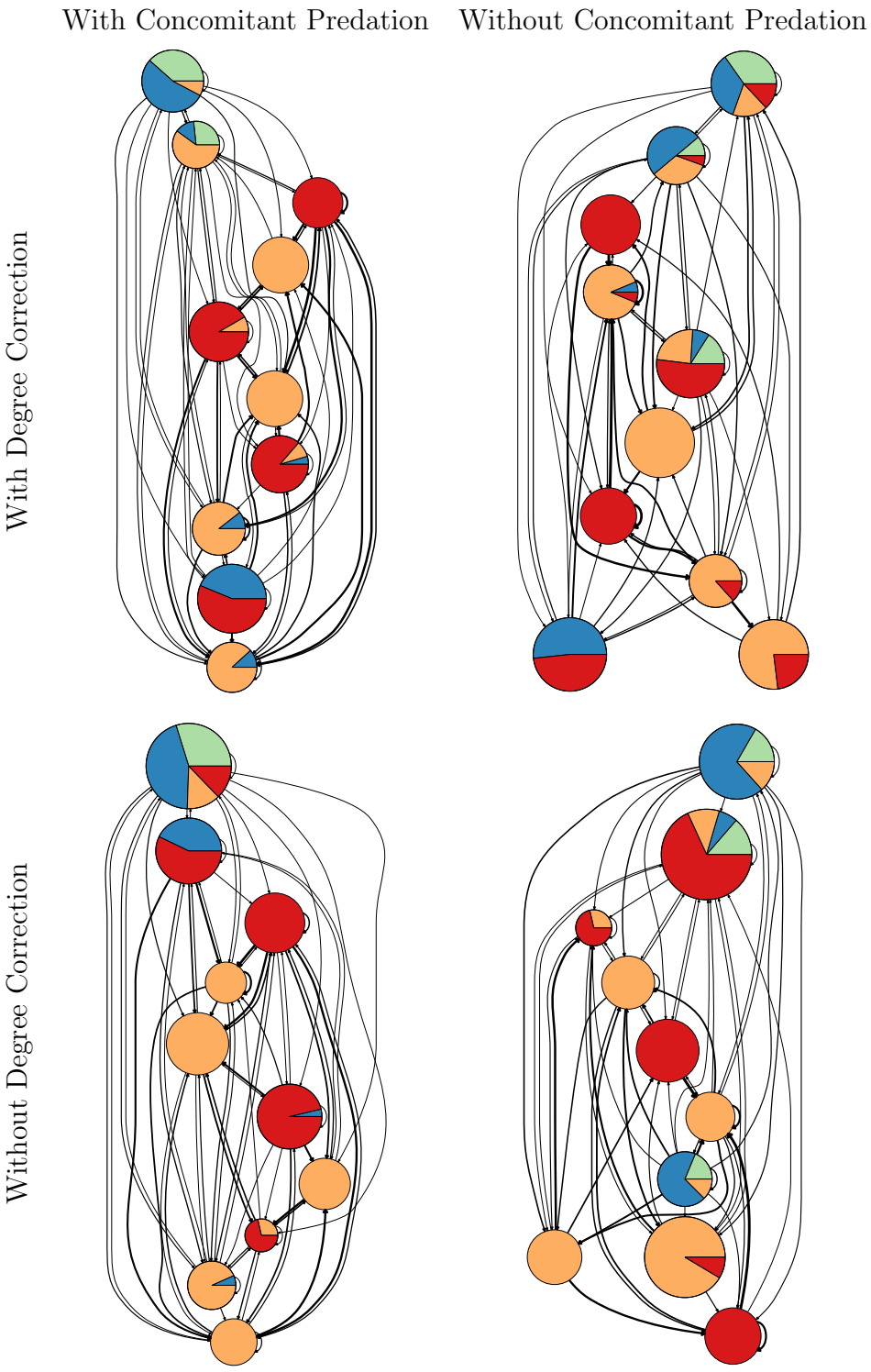

Figure 7.45: As Figure 3.3, but for the PuntaBanda network and $g = 10$.

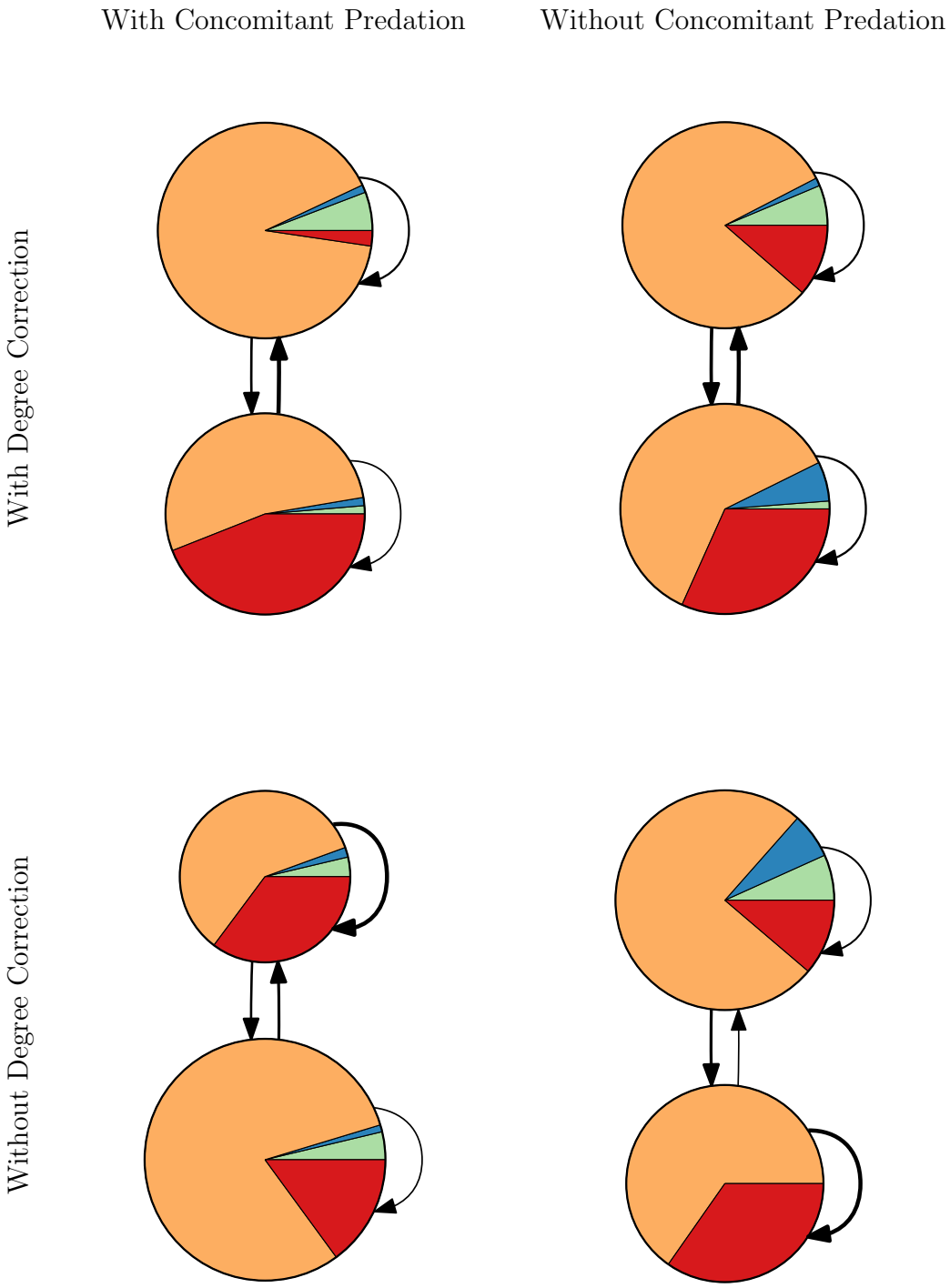

Figure 7.46: As Figure 3.3, but for the Sylt network and $g = 2$.

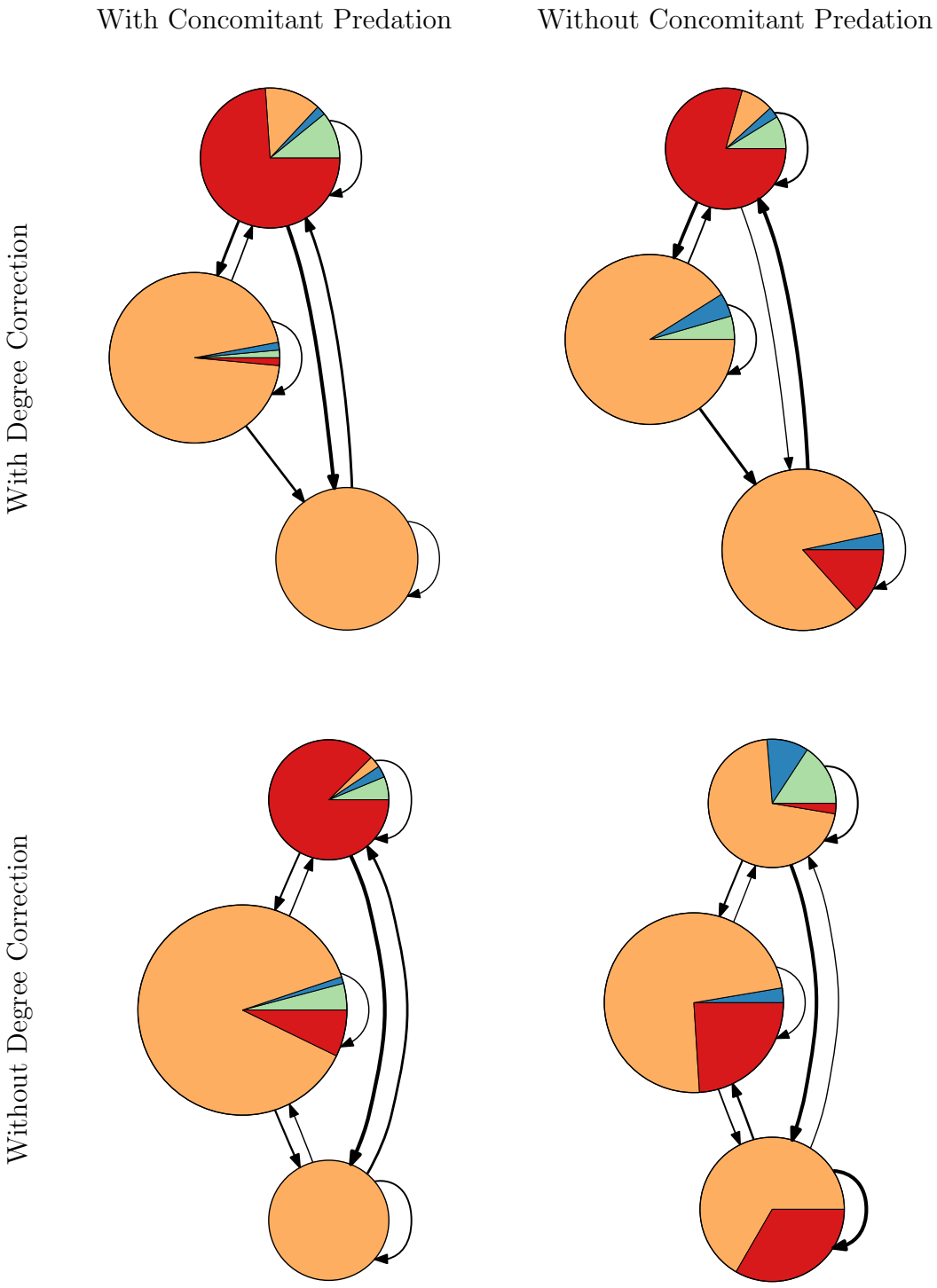

Figure 7.47: As Figure 3.3, but for the Sylt network and $g = 3$.

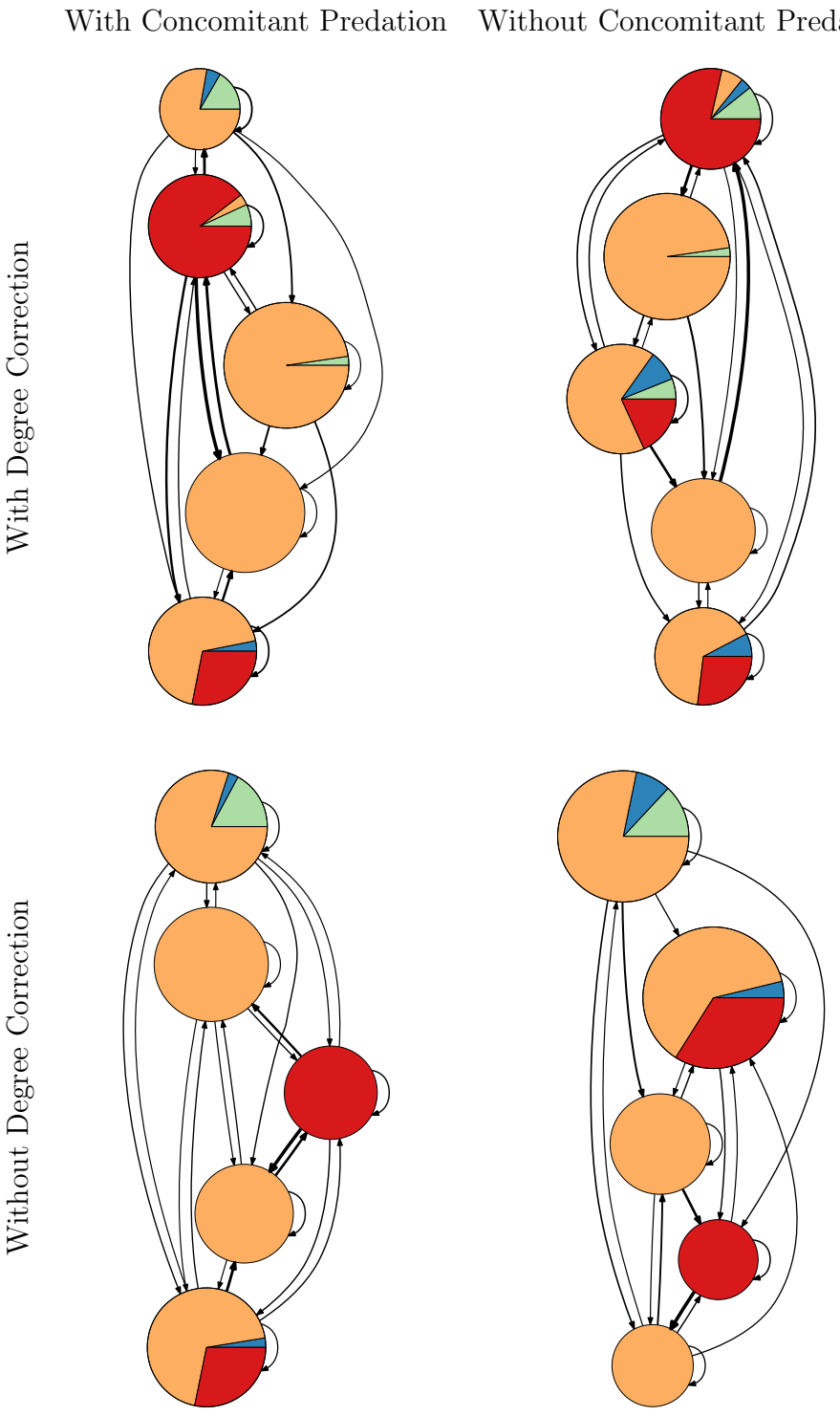

Figure 7.48: As Figure 3.3, but for the Sylt network and $g = 5$.

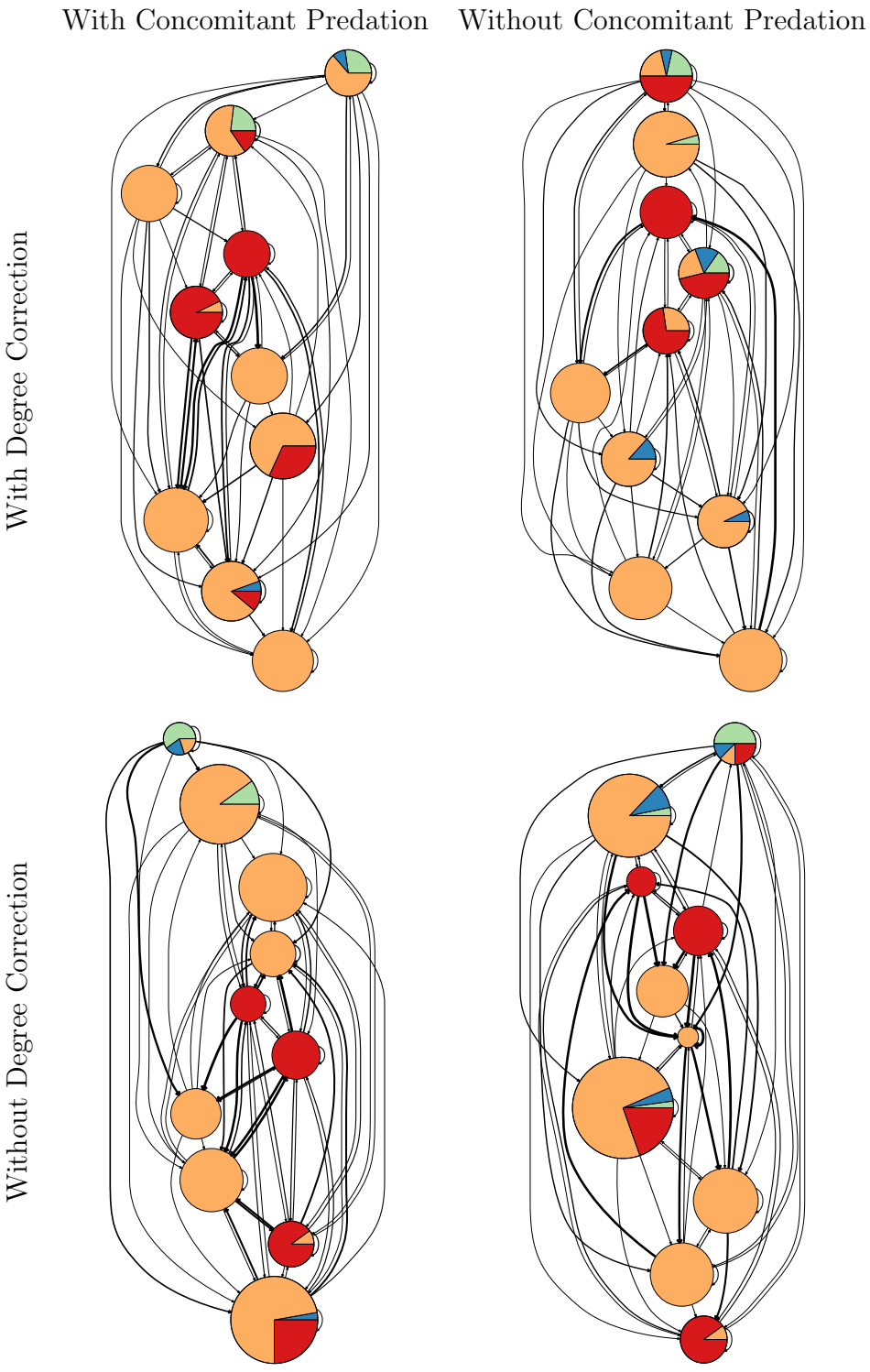

Figure 7.49: As Figure 3.3, but for the Sylt network and $g = 10$.

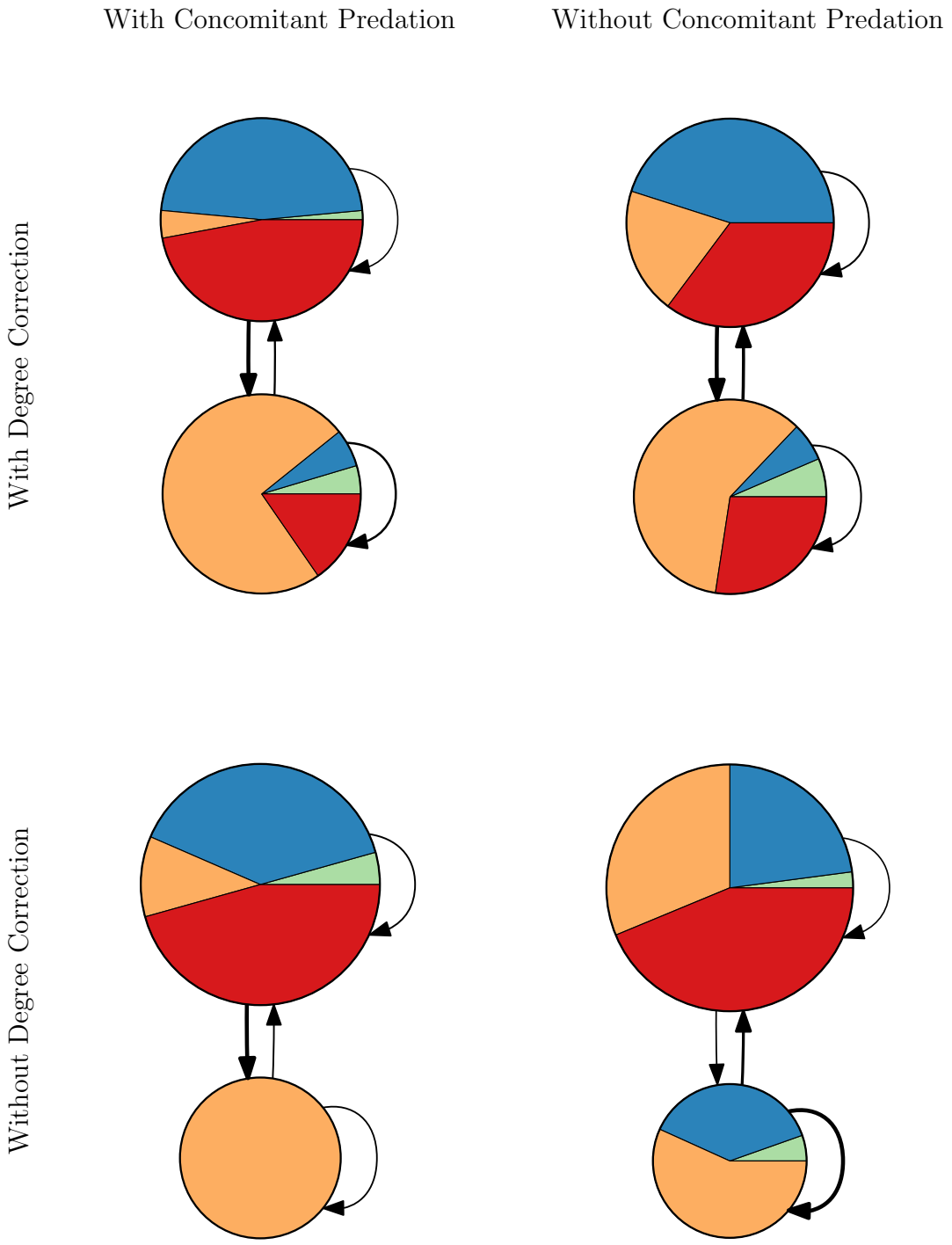

Figure 7.50: As Figure 3.3, but for the Ythan network and $g = 2$.

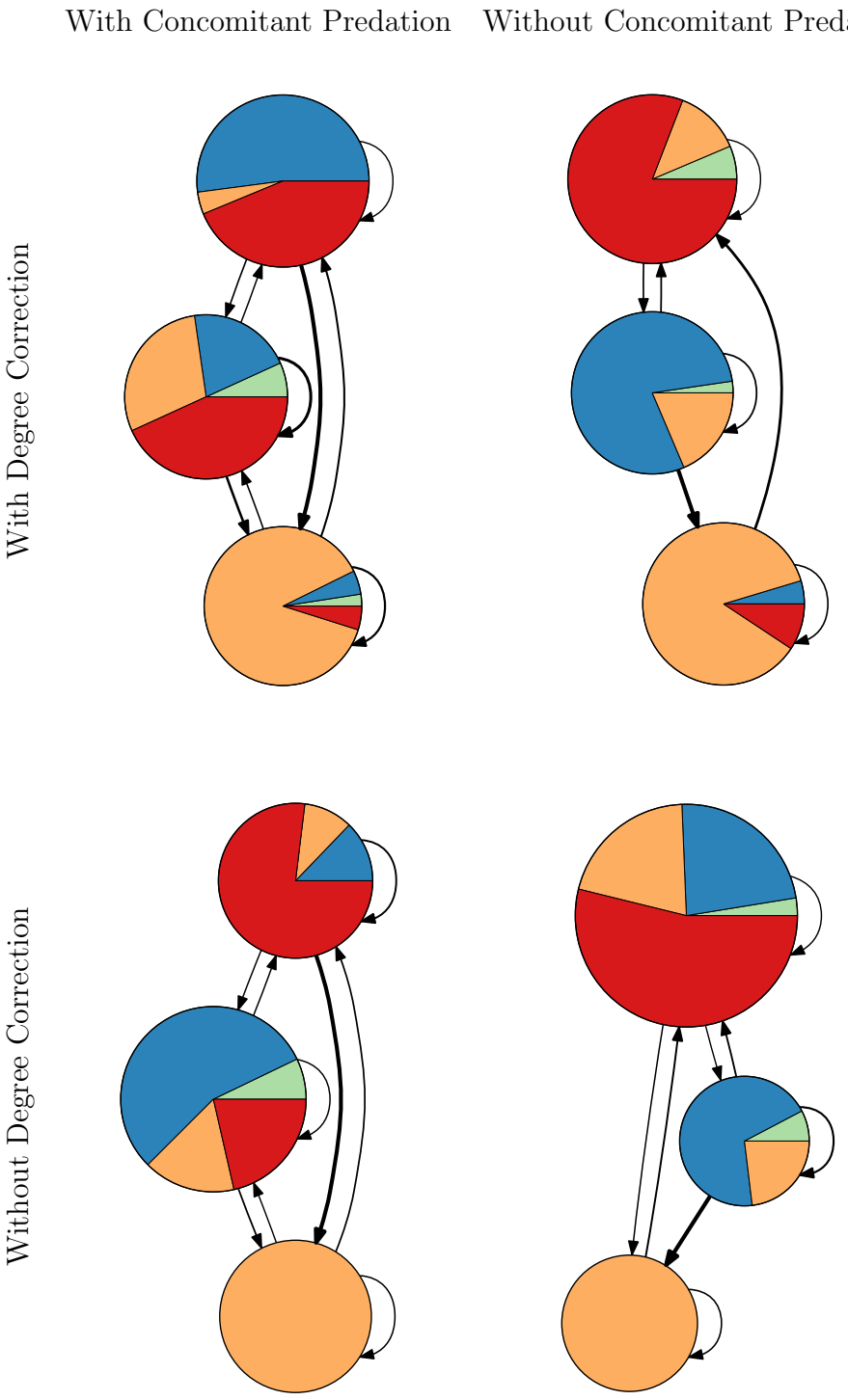

Figure 7.51: As Figure 3.3, but for the Ythan network and $g = 3$.

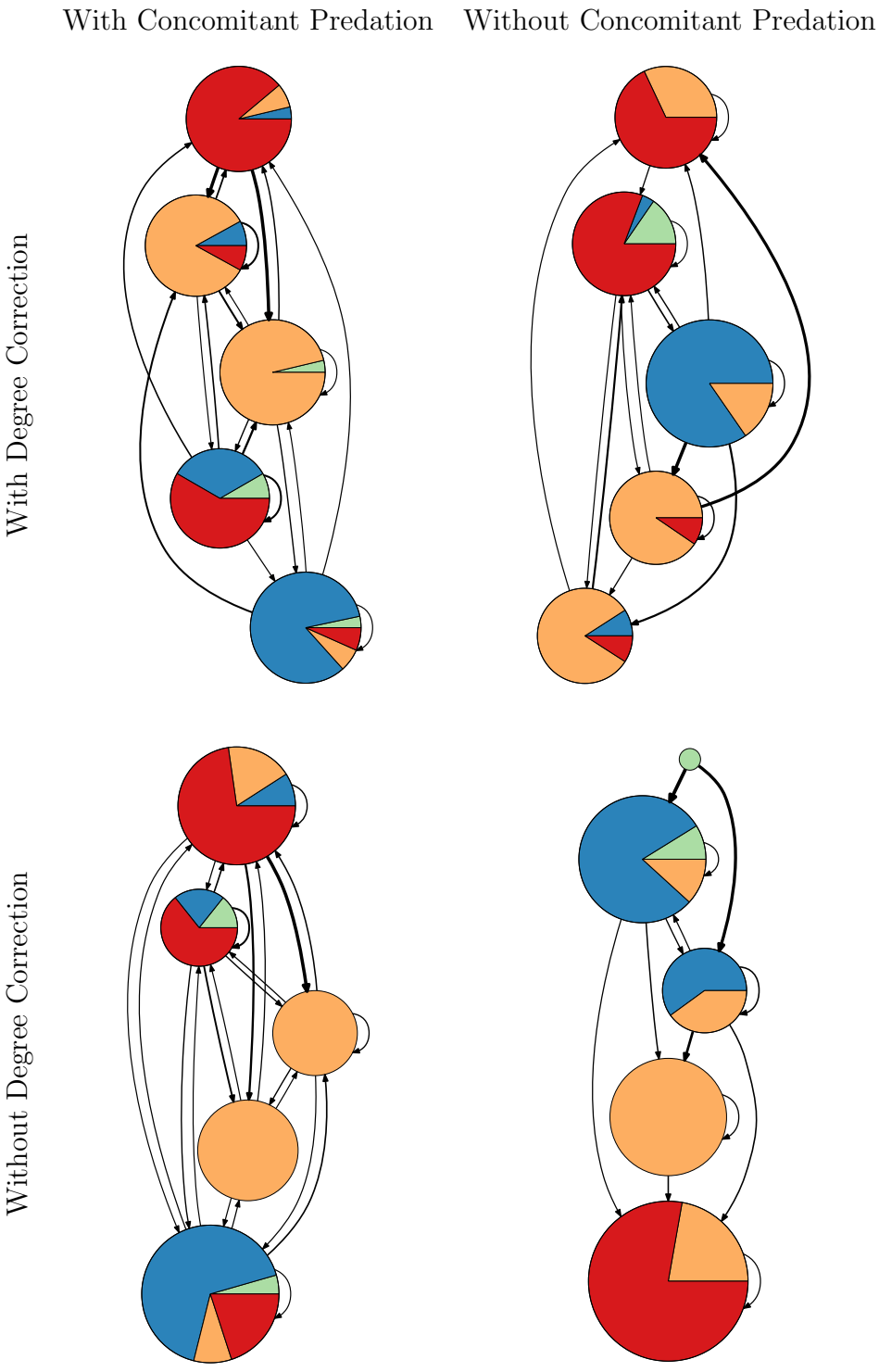

Figure 7.52: As Figure 3.3, but for the Ythan network and $g = 5$.

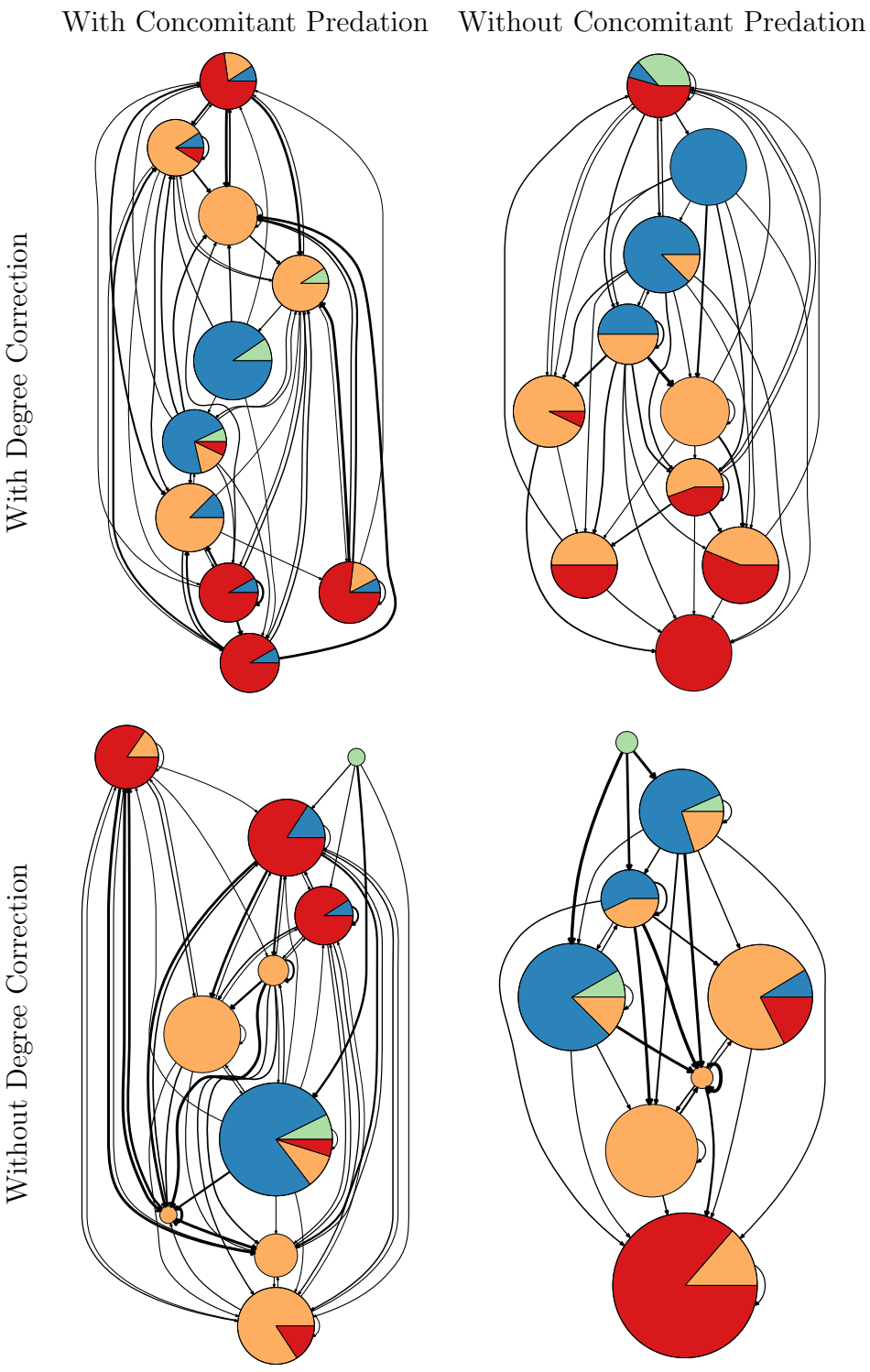

Figure 7.53: As Figure 3.3, but for the Ythan network and $g = 10$. Note that the group model without degree correction found that 8 groups outperforms 10 for the Ythan web without concomitant predation. This has the effect of making the nodes for this network disproportionately large and thus uncomparable to the condensed graphs with 10 nodes.

CHAPTER 8

APPENDIX 3: SUPPLEMENTAL INFORMATION FOR *ECOLOGICAL NETWORK INFERENCE FROM LONG-TERM PRESENCE-ABSENCE DATA*

Web	Model	Precision	Recall	p-value
Tatoosh Trophic	DBN-2	0.24	0.02	0.35
Tatoosh Trophic	DBN-3	0.15	0.01	0.78
Tatoosh Trophic	DBN-4	0.18	0.01	0.67
Tatoosh Trophic	DBN-5	0.20	0.02	0.59
Tatoosh Trophic	Lasso-1st	0.19	0.19	1.00
Tatoosh Trophic	Lasso-2nd	0.21	0.41	0.96
Tatoosh Trophic	Pearson	0.18	0.16	0.81
Tatoosh Nontrophic	DBN-2	0.31	0.03	0.02
Tatoosh Nontrophic	DBN-3	0.27	0.03	0.06
Tatoosh Nontrophic	DBN-4	0.27	0.03	0.06
Tatoosh Nontrophic	DBN-5	0.27	0.03	0.06
Tatoosh Nontrophic	Lasso-1st	0.22	0.30	0.10
Tatoosh Nontrophic	Lasso-2nd	0.20	0.53	0.12
Tatoosh Nontrophic	Pearson	0.25	0.28	0.00
France	DBN-2	0.33	0.06	0.06
France	DBN-3	0.20	0.07	0.51
France	DBN-4	0.22	0.08	0.35
France	DBN-5	0.23	0.09	0.27
France	Lasso-1st	0.24	0.57	0.02
France	Lasso-2nd	0.23	0.68	0.01
France	Pearson	0.19	0.90	0.18

Table 8.1: Expanded table of precision and recall with p-values. Columns represent the empirical network used, the model used to construct the inferred network, and the FDR-corrected p-values, rounded to two decimal places.

Model	Subsample	Median Probability	Log Likelihood
Coin	1	0.71	-1351.65
Coin	2	0.71	-1438.84
Coin	3	0.72	-1359.96
Coin	4	0.71	-1408.46
Coin	5	0.72	-1324.23
Coin	6	0.71	-1349.12
Coin	7	0.71	-1347.53
Coin	8	0.71	-1363.43
Coin	9	0.71	-1366.17
Coin	10	0.71	-1355.30
Coin	11	0.71	-1432.69
Coin	12	0.71	-1399.75
Coin	13	0.71	-1337.17
Coin	14	0.72	-1387.19
Disconnected	1	0.81	-581.15
Disconnected	2	0.79	-652.67
Disconnected	3	0.81	-630.82
Disconnected	4	0.80	-662.79
Disconnected	5	0.82	-588.37
Disconnected	6	0.81	-604.36
Disconnected	7	0.81	-598.80
Disconnected	8	0.81	-591.53
Disconnected	9	0.82	-594.54
Disconnected	10	0.81	-598.15
Disconnected	11	0.81	-624.40
Disconnected	12	0.80	-613.72
Disconnected	13	0.82	-606.19
Disconnected	14	0.81	-623.79

Table 8.2: Out-of-sample prediction accuracy for different subsamples of the Tatoosh dataset. Columns represent the model, subsample number, the median probability of correctly predicting a single out-of-sample species at one time point, and the log likelihood of correctly predicting all species at all time points, given the model trained on on that data subsample. The log likelihood values are shown visually in main text Fig. 5.2.

Model	Subsample	Median Probability	Log Likelihood
DBN-2	1	0.83	-571.70
DBN-2	2	0.82	-646.60
DBN-2	3	0.82	-617.19
DBN-2	4	0.82	-669.52
DBN-2	5	0.84	-582.59
DBN-2	6	0.84	-590.07
DBN-2	7	0.84	-579.43
DBN-2	8	0.83	-584.08
DBN-2	9	0.83	-591.26
DBN-2	10	0.84	-576.44
DBN-2	11	0.83	-594.09
DBN-2	12	0.82	-603.52
DBN-2	13	0.84	-579.63
DBN-2	14	0.83	-624.11
DBN-3	1	0.83	-571.70
DBN-3	2	0.82	-646.60
DBN-3	3	0.82	-617.19
DBN-3	4	0.82	-669.52
DBN-3	5	0.84	-582.59
DBN-3	6	0.84	-590.07
DBN-3	7	0.84	-579.43
DBN-3	8	0.83	-584.08
DBN-3	9	0.83	-591.26
DBN-3	10	0.84	-576.44
DBN-3	11	0.83	-594.09
DBN-3	12	0.82	-603.52
DBN-3	13	0.84	-579.63
DBN-3	14	0.83	-624.11

Table 8.2, continued

Model	Subsample	Median Probability	Log Likelihood
DBN-4	1	0.83	-571.70
DBN-4	2	0.82	-646.60
DBN-4	3	0.82	-617.19
DBN-4	4	0.82	-669.52
DBN-4	5	0.84	-582.59
DBN-4	6	0.84	-590.07
DBN-4	7	0.84	-579.43
DBN-4	8	0.83	-584.08
DBN-4	9	0.83	-591.26
DBN-4	10	0.84	-576.44
DBN-4	11	0.83	-594.09
DBN-4	12	0.82	-603.52
DBN-4	13	0.84	-579.63
DBN-4	14	0.83	-624.11
DBN-5	1	0.83	-571.70
DBN-5	2	0.82	-646.60
DBN-5	3	0.82	-617.19
DBN-5	4	0.82	-669.52
DBN-5	5	0.84	-582.59
DBN-5	6	0.84	-590.07
DBN-5	7	0.84	-579.43
DBN-5	8	0.83	-584.08
DBN-5	9	0.83	-591.26
DBN-5	10	0.84	-576.44
DBN-5	11	0.83	-594.09
DBN-5	12	0.82	-603.52
DBN-5	13	0.84	-579.63
DBN-5	14	0.83	-624.11

Table 8.2, continued

Model	Subsample	Median Probability	Log Likelihood
Lasso-1st	1	0.71	-580.27
Lasso-1st	2	0.71	-609.93
Lasso-1st	3	0.73	-580.36
Lasso-1st	4	0.71	-607.62
Lasso-1st	5	0.72	-571.27
Lasso-1st	6	0.71	-605.50
Lasso-1st	7	0.73	-581.86
Lasso-1st	8	0.71	-562.52
Lasso-1st	9	0.71	-580.16
Lasso-1st	10	0.70	-622.67
Lasso-1st	11	0.70	-666.94
Lasso-1st	12	0.71	-629.32
Lasso-1st	13	0.72	-598.10
Lasso-1st	14	0.74	-612.47
Lasso-2nd	1	0.71	-580.27
Lasso-2nd	2	0.71	-609.93
Lasso-2nd	3	0.73	-580.36
Lasso-2nd	4	0.71	-607.62
Lasso-2nd	5	0.72	-571.27
Lasso-2nd	6	0.71	-605.50
Lasso-2nd	7	0.73	-581.86
Lasso-2nd	8	0.71	-562.52
Lasso-2nd	9	0.71	-580.16
Lasso-2nd	10	0.70	-622.67
Lasso-2nd	11	0.70	-666.94
Lasso-2nd	12	0.71	-629.32
Lasso-2nd	13	0.72	-598.10
Lasso-2nd	14	0.74	-612.47

Table 8.2, continued

Predator	Prey	Ref
Ameiurus melas	Cyprinus carpio	[149]
Ameiurus melas	Gymnocephalus cernua	[186]
Ameiurus melas	Lepomis gibbosus	[149]
Anguilla anguilla	Abramis brama	[216, 143]
Anguilla anguilla	Barbatula barbatula	[22, 105]
Anguilla anguilla	Cottus gobio	[105, 103]
Anguilla anguilla	Gasterosteus aculeatus	[105, 143, 218]
Anguilla anguilla	Gobio gobio	[22, 105]
Anguilla anguilla	Gymnocephalus cernua	[127, 185, 62, 216, 143]
Anguilla anguilla	Leuciscus leuciscus	[22, 216]
Anguilla anguilla	Lota lota	[216]
Anguilla anguilla	Phoxinus phoxinus	[90, 245, 218, 105]
Anguilla anguilla	Perca fluviatilis	[90, 62, 216, 143]
Anguilla anguilla	Rutilus rutilus	[22, 218, 62, 216, 143]
Anguilla anguilla	Sander lucioperca	[62, 143]
Anguilla anguilla	Salmo salar	[146]
Anguilla anguilla	Squalius cephalus	[216]
Blicca bjoerkna	Gymnocephalus cernua	[185]
Cottus gobio	Perca fluviatilis	[229]
Cottus gobio	Salmo trutta	[93]
Esox lucius	Abramis brama	[255, 104]
Esox lucius	Alburnus alburnus	[255, 157, 129, 156]
Esox lucius	Ameiurus melas	[224, 186]
Esox lucius	Anguilla anguilla	[91, 2, 157, 156, 105, 104]
Esox lucius	Barbatula barbatula	[157, 156, 135, 105, 104, 103]
Esox lucius	Cottus gobio	[157, 156, 135, 105, 104]
Esox lucius	Cyprinus carpio	[164, 77, 61]
Esox lucius	Gasterosteus aculeatus	[7, 91, 255, 2, 156, 135, 105, 104, 103]
Esox lucius	Gobio gobio	[157, 156, 14, 77, 105, 104, 61, 103, 56]
Esox lucius	Gymnocephalus cernua	[76, 129, 185, 135, 186]
Esox lucius	Lepomis gibbosus	[163, 77, 153]
Esox lucius	Leuciscus leuciscus	[157, 156]
Esox lucius	Lota lota	[255, 60, 126]
Esox lucius	Perca fluviatilis	[7, 91, 255, 2, 129, 60, 126, 135]
Esox lucius	Phoxinus phoxinus	[7, 91, 255, 2, 157, 156, 126, 105, 104, 103]
Esox lucius	Pungitius pungitius	[147, 126, 135]

Table 8.3: Adjacency list of interactions in the French piscivory food web, with supporting references.

Predator	Prey	Ref
Esox lucius	Rutilus rutilus	[255, 2, 157, 129, 156, 105, 104, 103]
Esox lucius	Sander lucioperca	[129]
Esox lucius	Salmo trutta	[7, 91, 255, 2, 157, 156, 126, 122, 111, 61]
Esox lucius	Salmo salar	[2, 157, 146, 135]
Esox lucius	Scardinius erythrophthalmus	[76, 153]
Esox lucius	Squalius cephalus	[156, 14, 56]
Esox lucius	Thymallus thymallus	[157]
Esox lucius	Tinca tinca	[14, 153, 61]
Gasterosteus aculeatus	Pungitius pungitius	[116]
Gymnocephalus cernua	Gymnocephalus cernua	[185]
Lota lota	Gymnocephalus cernua	[186, 185]
Lota lota	Esox lucius	[92]
Lota lota	Pungitius pungitius	[126, 92]
Lota lota	Salmo trutta	[25, 126]
Lota lota	Salmo salar	[146]
Lota lota	Sander lucioperca	[45]
Perca fluviatilis	Alburnus alburnus	[57, 198]
Perca fluviatilis	Gasterosteus aculeatus	[6, 104]
Perca fluviatilis	Gymnocephalus cernua	[185, 63]
Perca fluviatilis	Phoxinus phoxinus	[6]
Perca fluviatilis	Rutilus rutilus	[120, 57, 79, 203, 198, 179, 65, 104]
Perca fluviatilis	Sander lucioperca	[200, 65, 63]
Perca fluviatilis	Scardinius erythrophthalmus	[204, 198]
Perca fluviatilis	Squalius cephalus	[198]
Pungitius pungitius	Gasterosteus aculeatus	[116]
Salmo salar	Gasterosteus aculeatus	[131, 119]
Salmo trutta	Barbatula barbatula	[89]
Salmo trutta	Cottus gobio	[168, 21]
Salmo trutta	Gasterosteus aculeatus	[91, 127, 145, 140, 121]
Salmo trutta	Gobio gobio	[202, 105, 89]
Salmo trutta	Gasterosteus aculeatus	[105, 140, 199, 118]
Salmo trutta	Gymnocephalus cernua	[117]
Salmo trutta	Perca fluviatilis	[91, 117, 184]
Salmo trutta	Phoxinus phoxinus	[245, 91, 202, 151, 178, 140, 139]

Table 8.3, continued

Predator	Prey	Ref
Salmo trutta	Pungitius pungitius	[126, 184, 121, 117]
Salmo trutta	Rutilus rutilus	[117]
Salmo trutta	Salmo salar	[115, 146, 98, 199, 118]
Sander lucioperca	Alburnus alburnus	[232, 130, 233, 249]
Sander lucioperca	Abramis brama	[232, 130, 233, 211, 37, 251]
Sander lucioperca	Cottus gobio	[227]
Sander lucioperca	Esox lucius	[130]
Sander lucioperca	Gymnocephalus cernua	[185, 27, 232, 95, 64, 137, 130, 233, 249, 211, 37]
Sander lucioperca	Lepomis gibbosus	[232]
Sander lucioperca	Perca fluviatilis	[27, 95, 64, 137, 130, 249, 102, 227, 211, 37, 251]
Sander lucioperca	Rutilus rutilus	[232, 95, 64, 137, 130, 249, 102, 227, 211, 37, 251]
Sander lucioperca	Salmo trutta	[122]
Squalius cephalus	Alburnoides bipunctatus	[181]
Squalius cephalus	Barbus meridionalis	[181]
Squalius cephalus	Cyprinus carpio	[250]

Table 8.3, continued

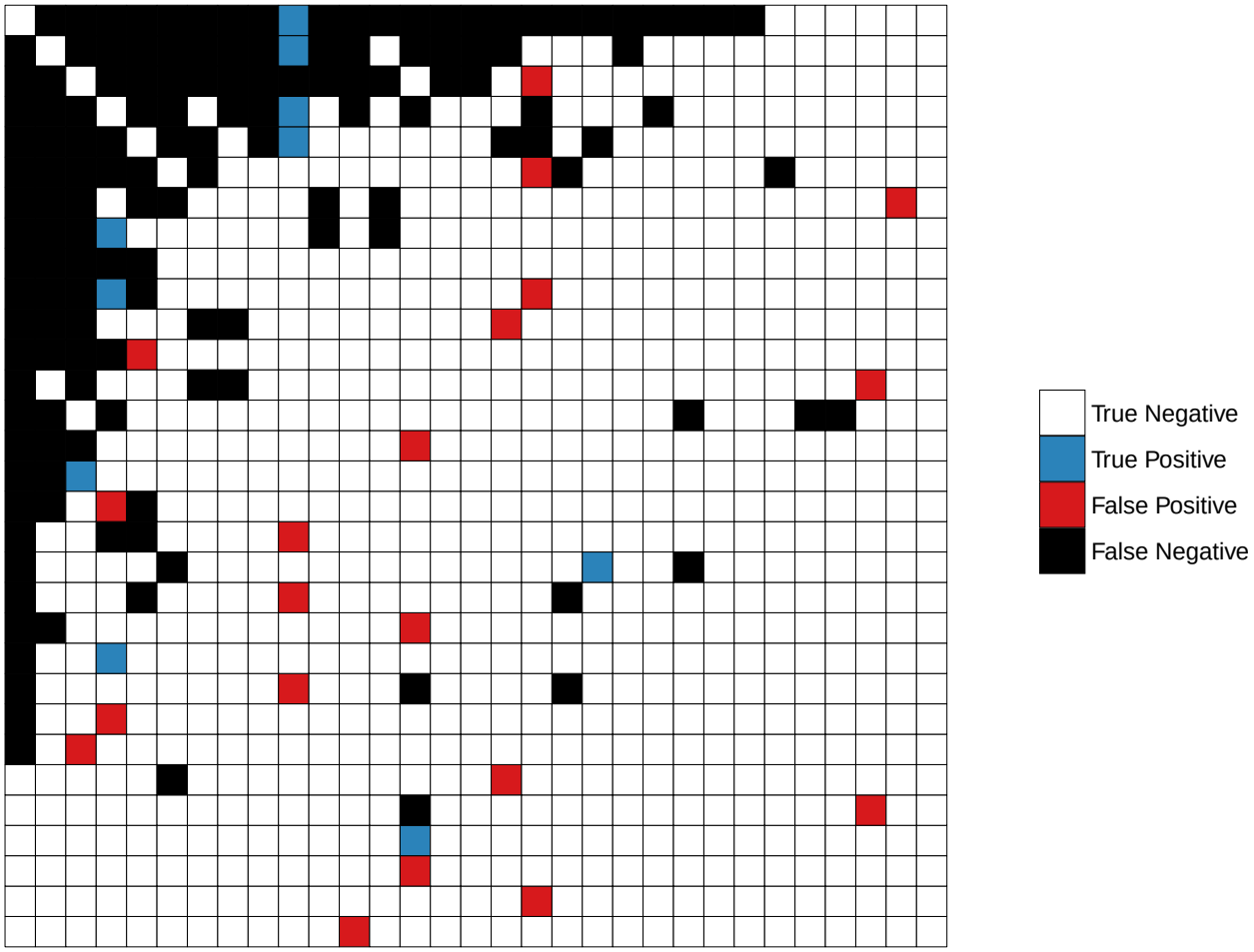

Figure 8.1: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the DBN-2 model.

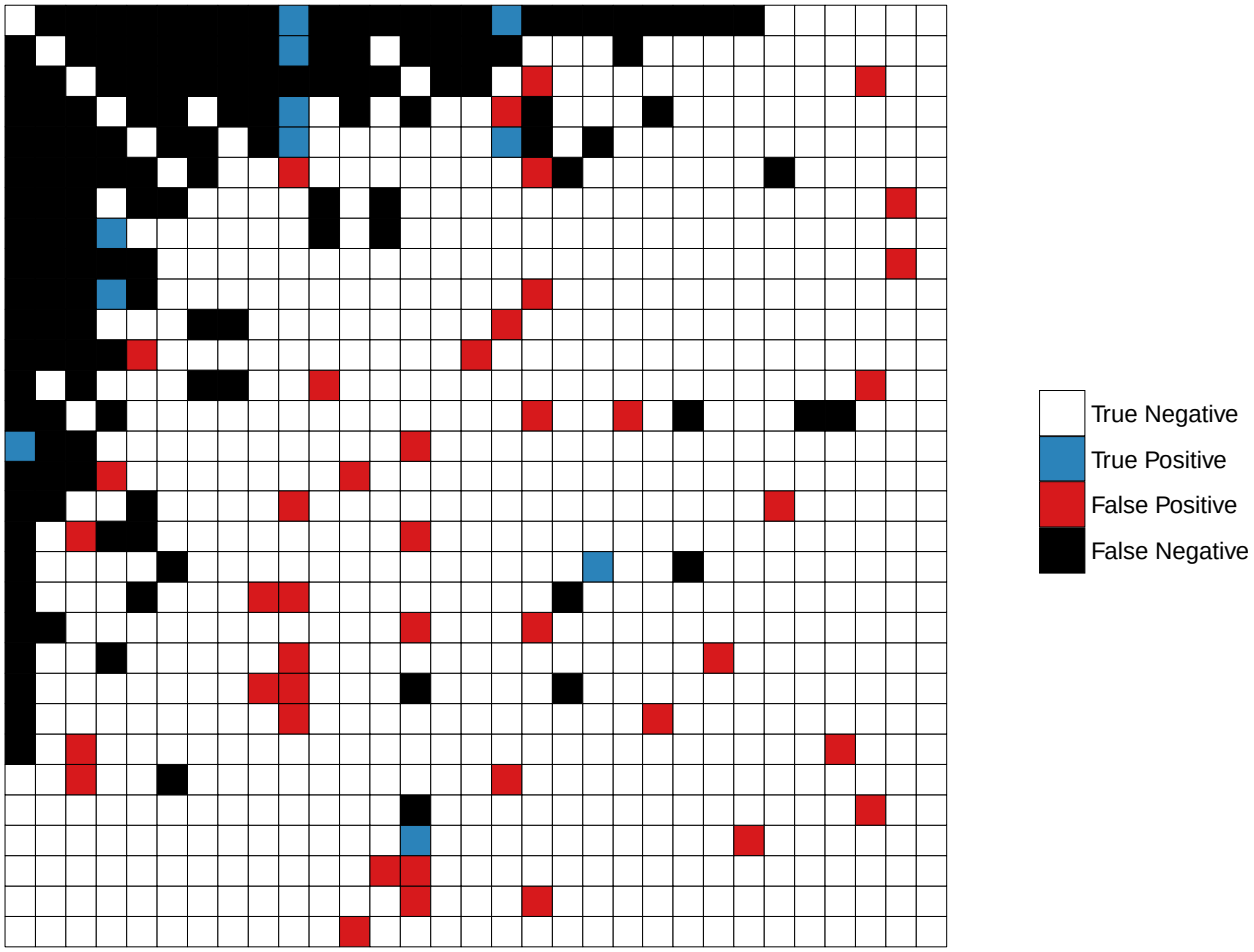

Figure 8.2: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the DBN-3 model.

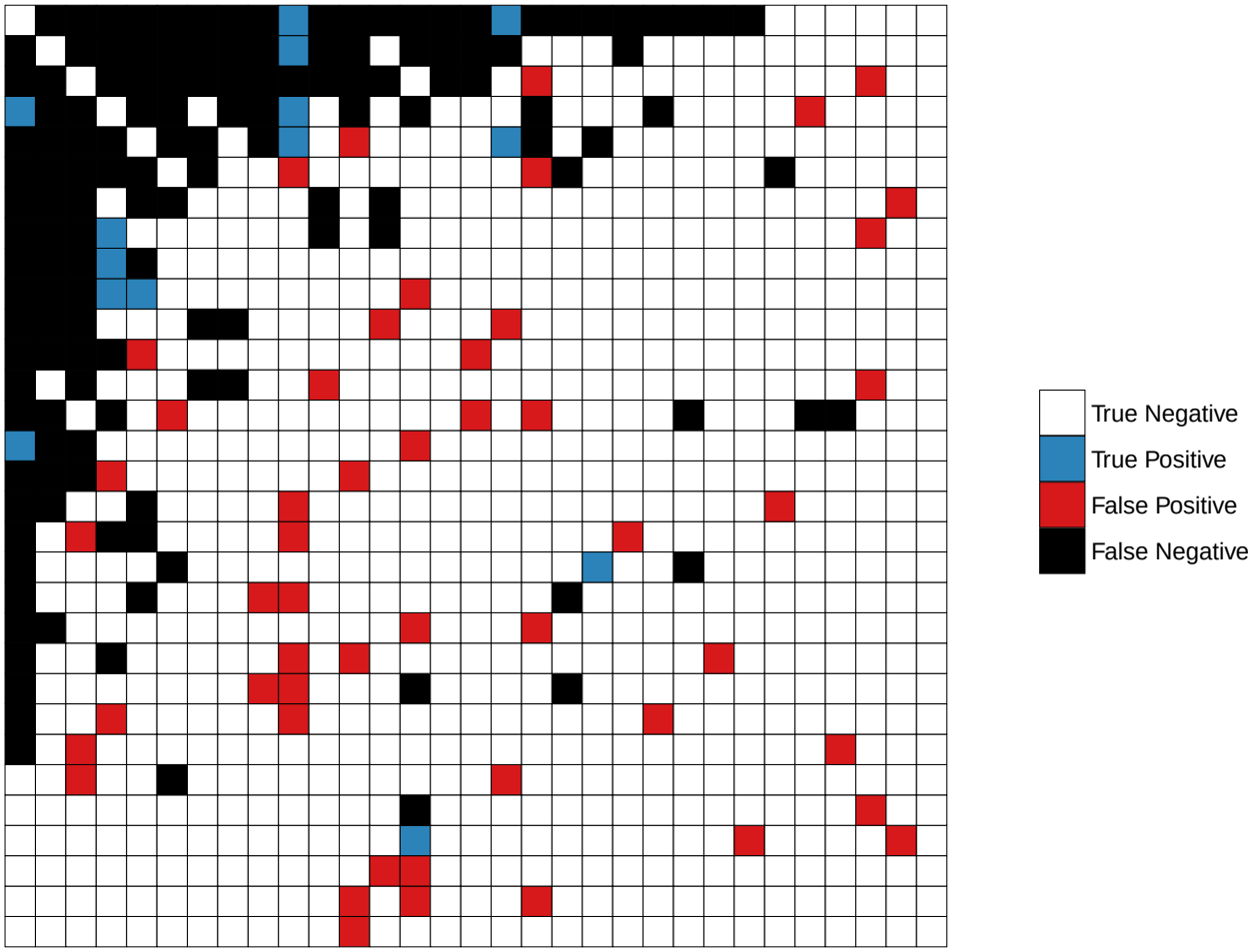

Figure 8.3: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the DBN-4 model.

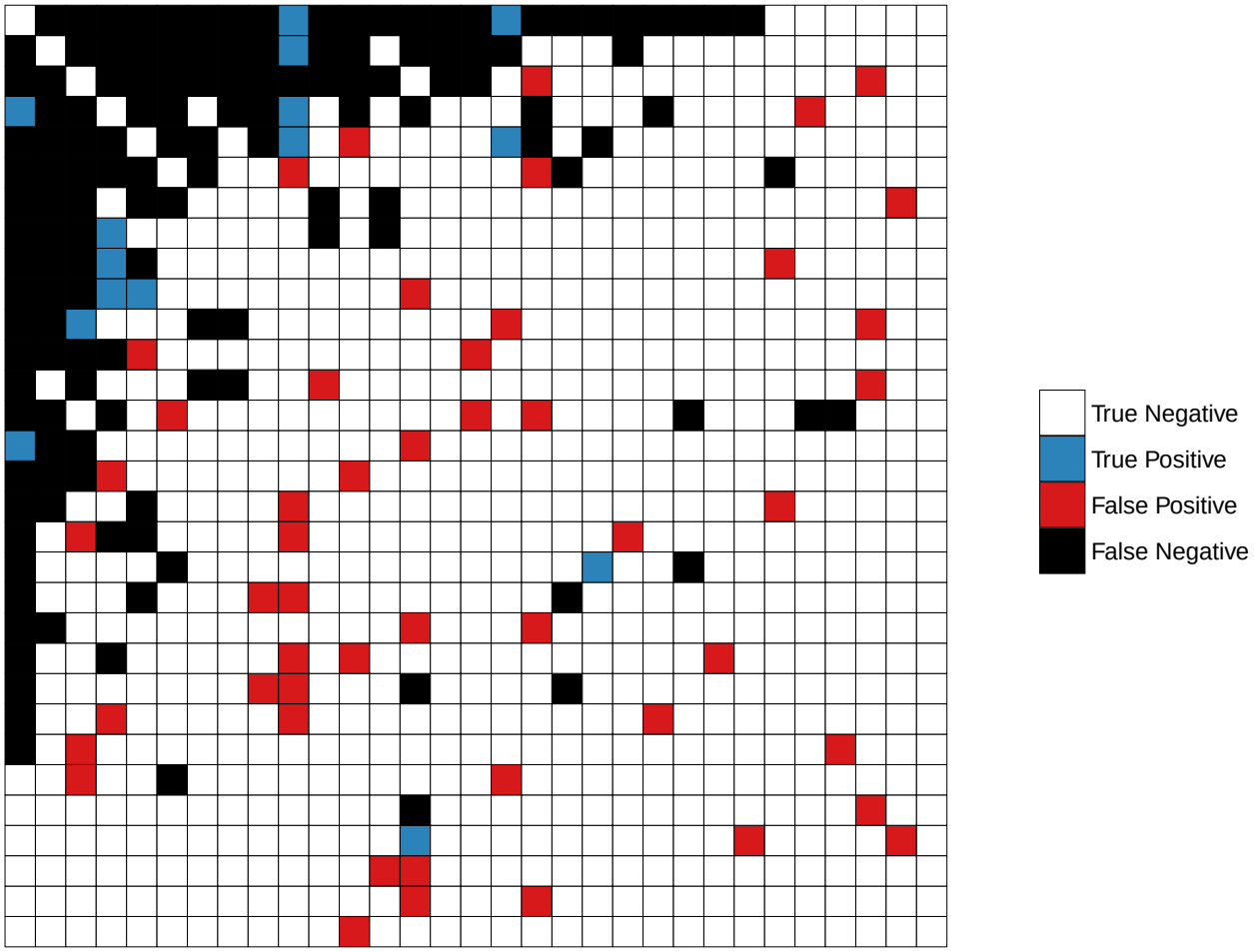

Figure 8.4: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the DBN-5 model.

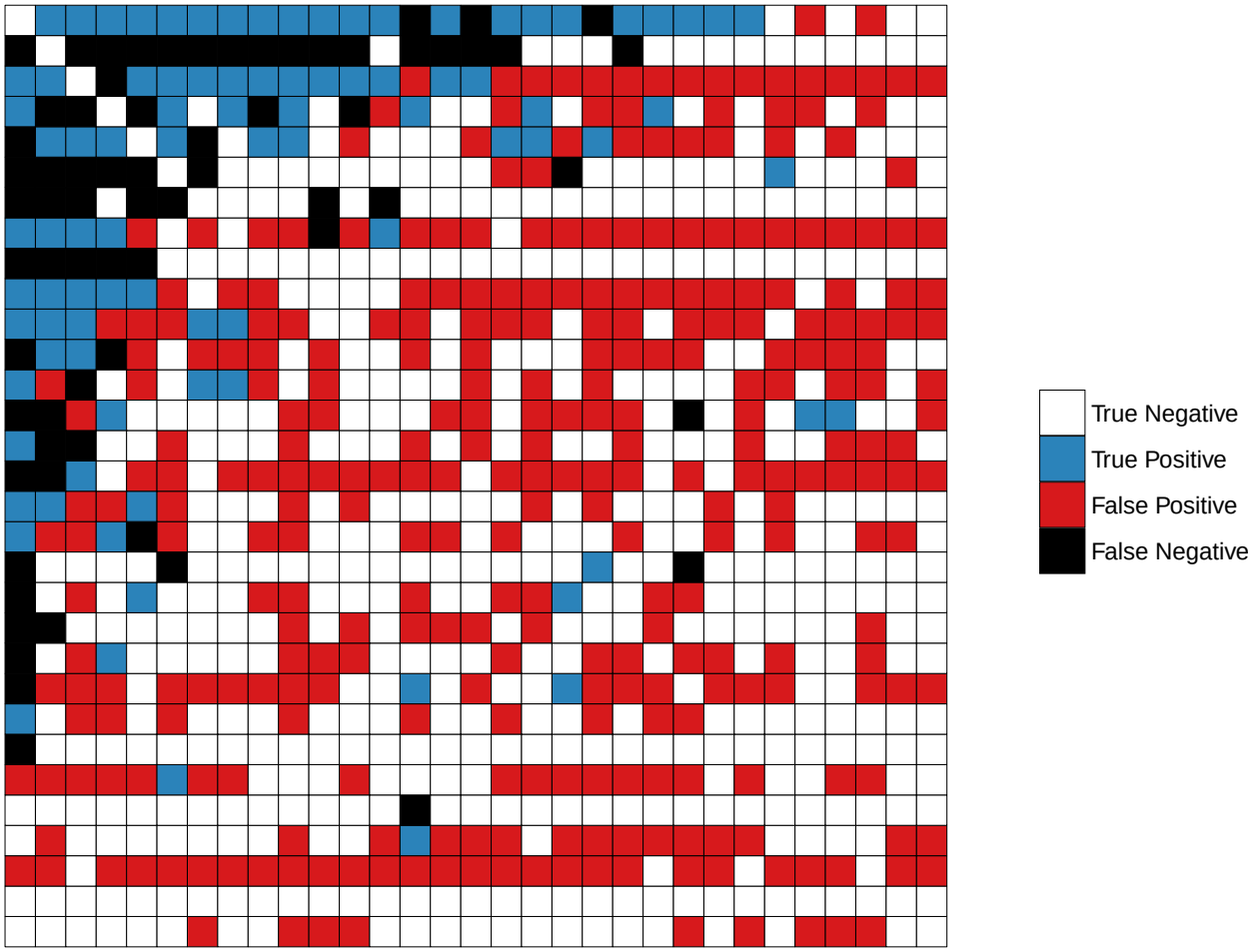

Figure 8.5: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the Lasso-1st model.

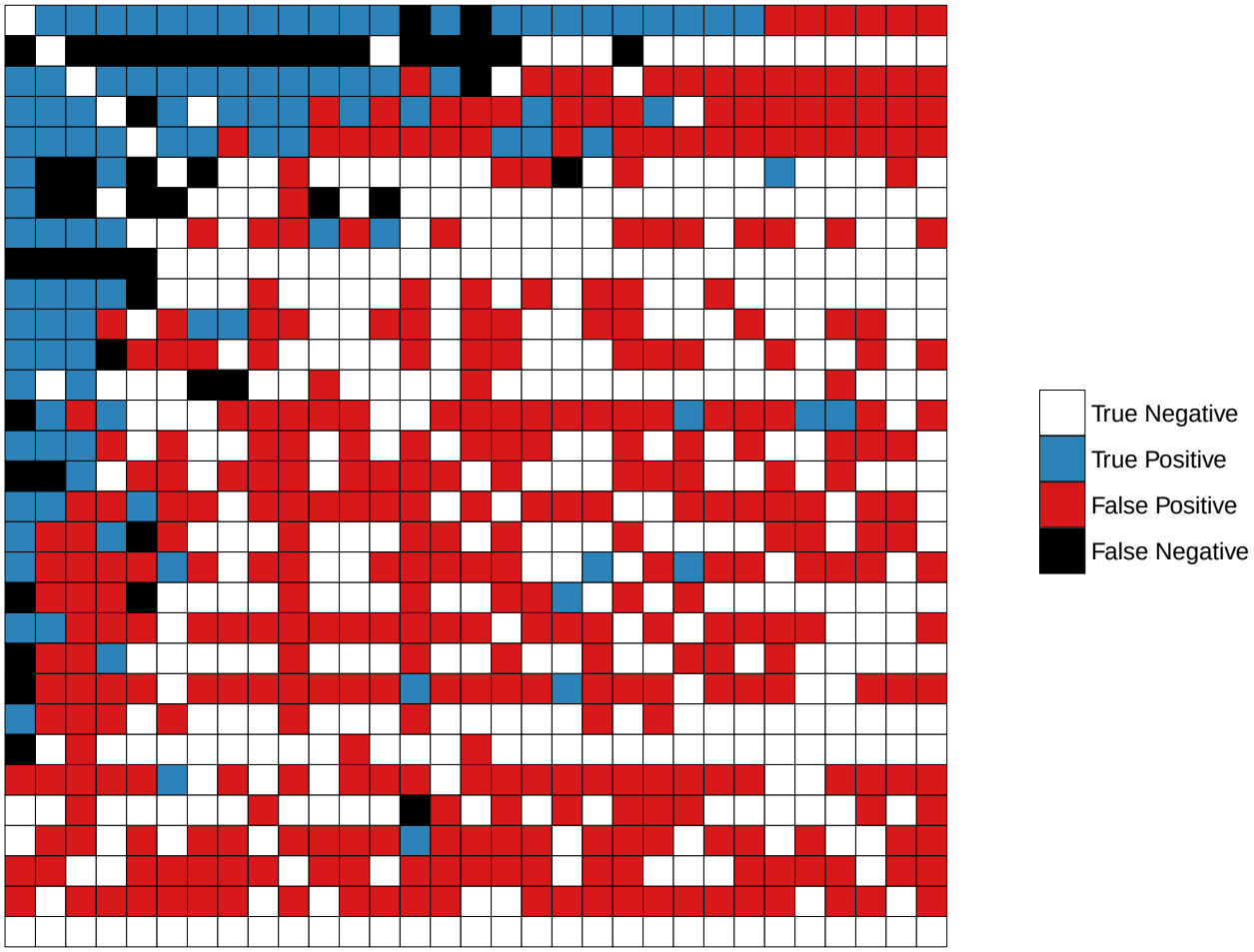

Figure 8.6: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the Lasso-2nd model.

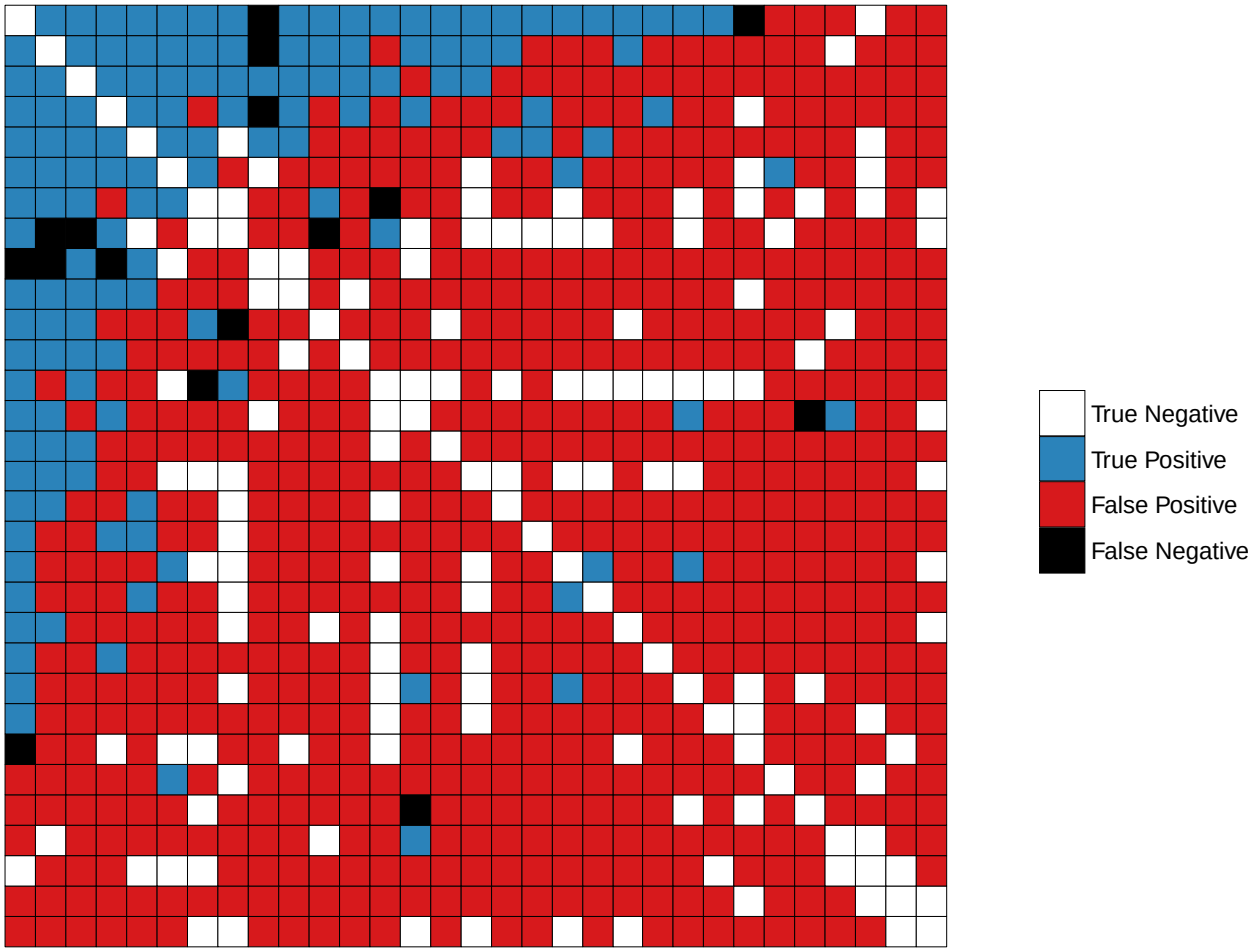

Figure 8.7: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the France piscivory network and the Pearson model.

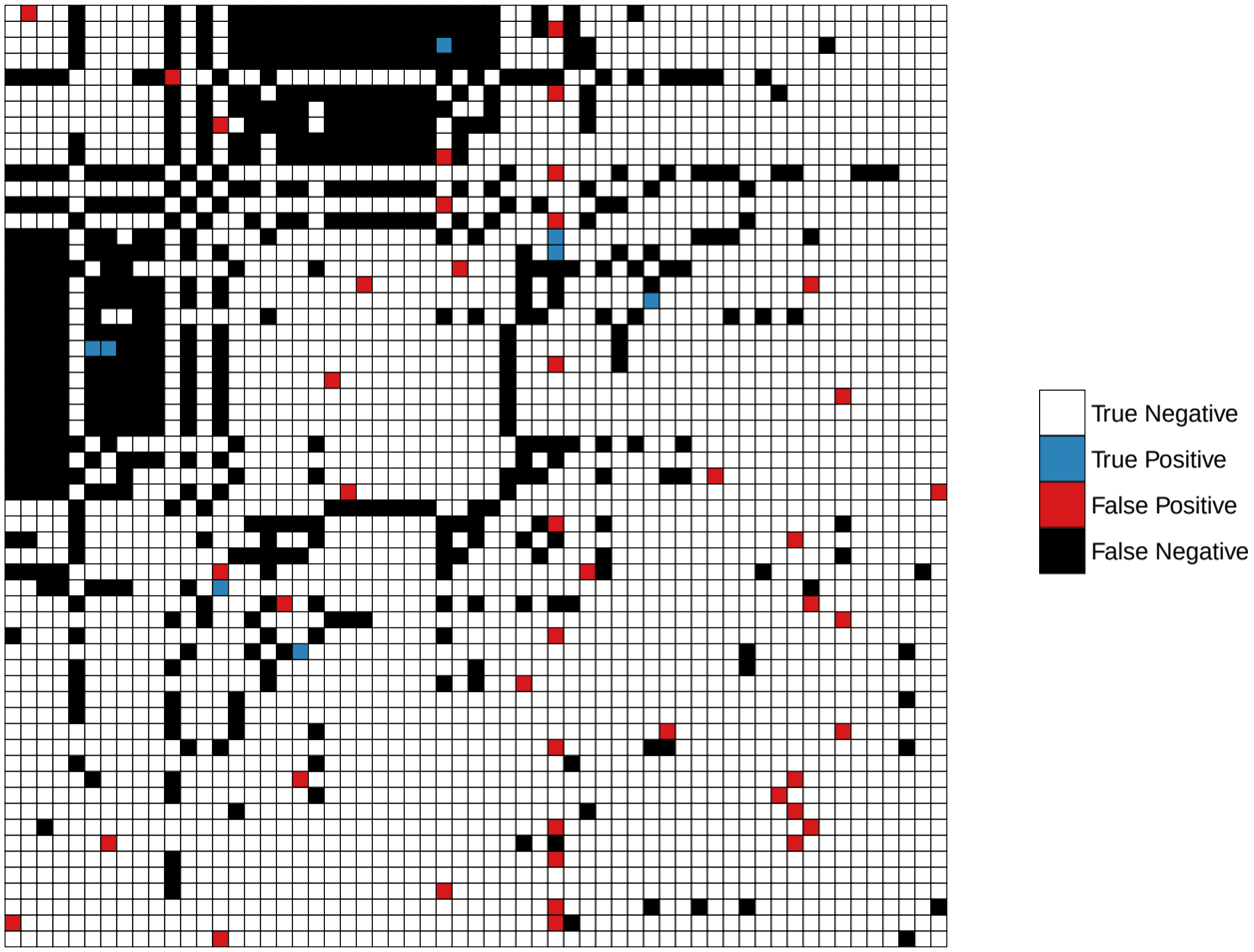

Figure 8.8: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh trophic network and the DBN-3 model.

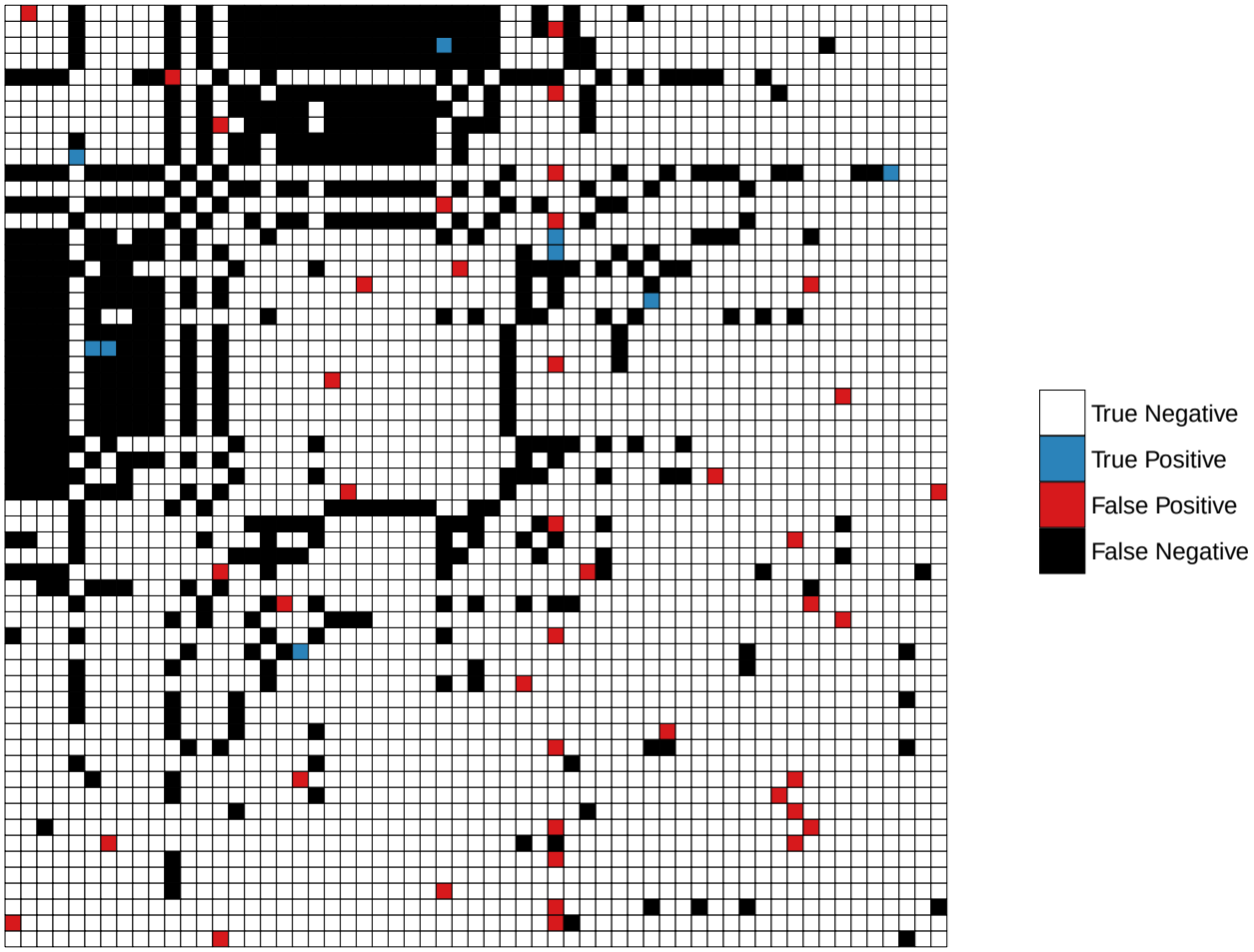

Figure 8.9: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh trophic network and the DBN-4 model.

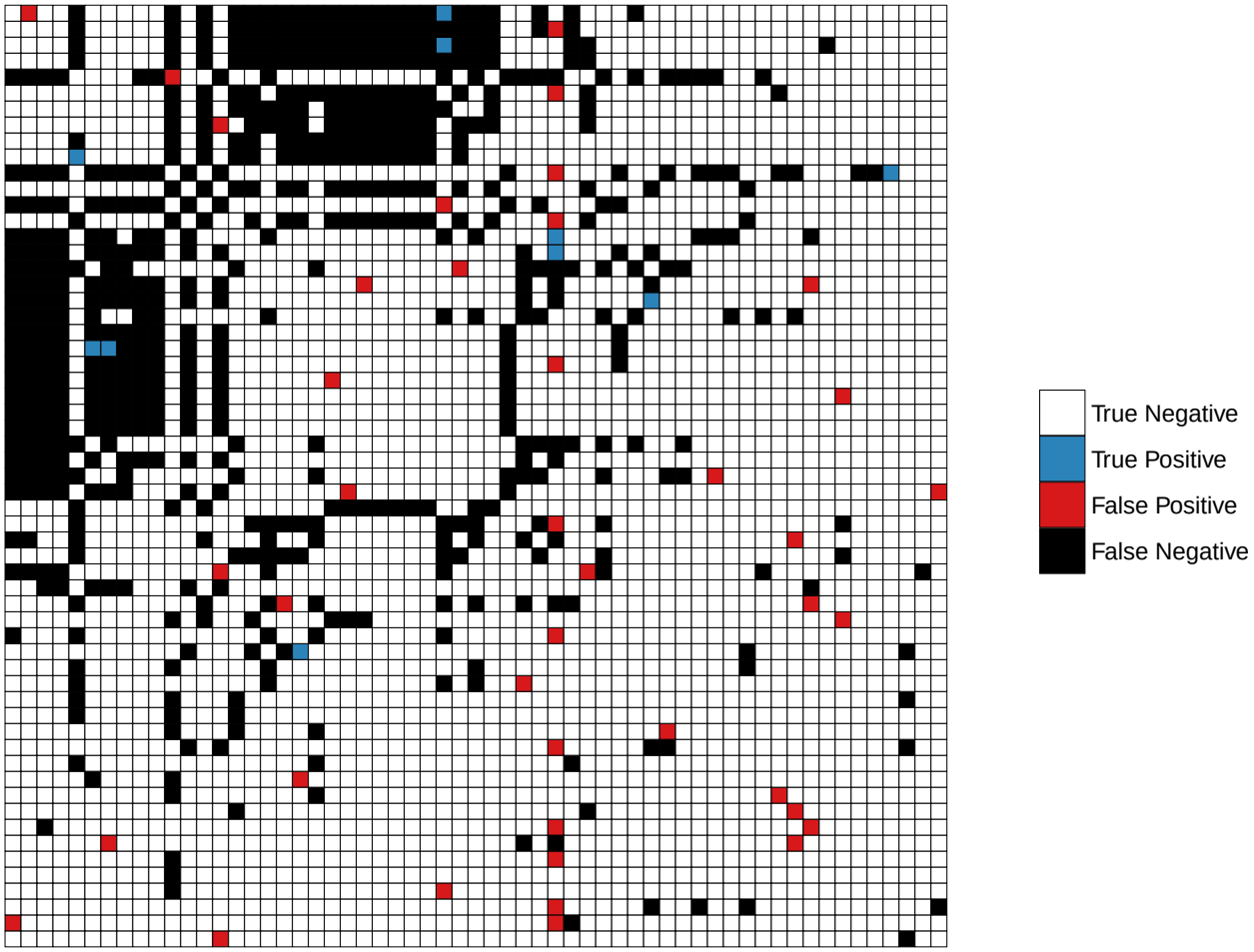

Figure 8.10: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh trophic network and the DBN-5 model.

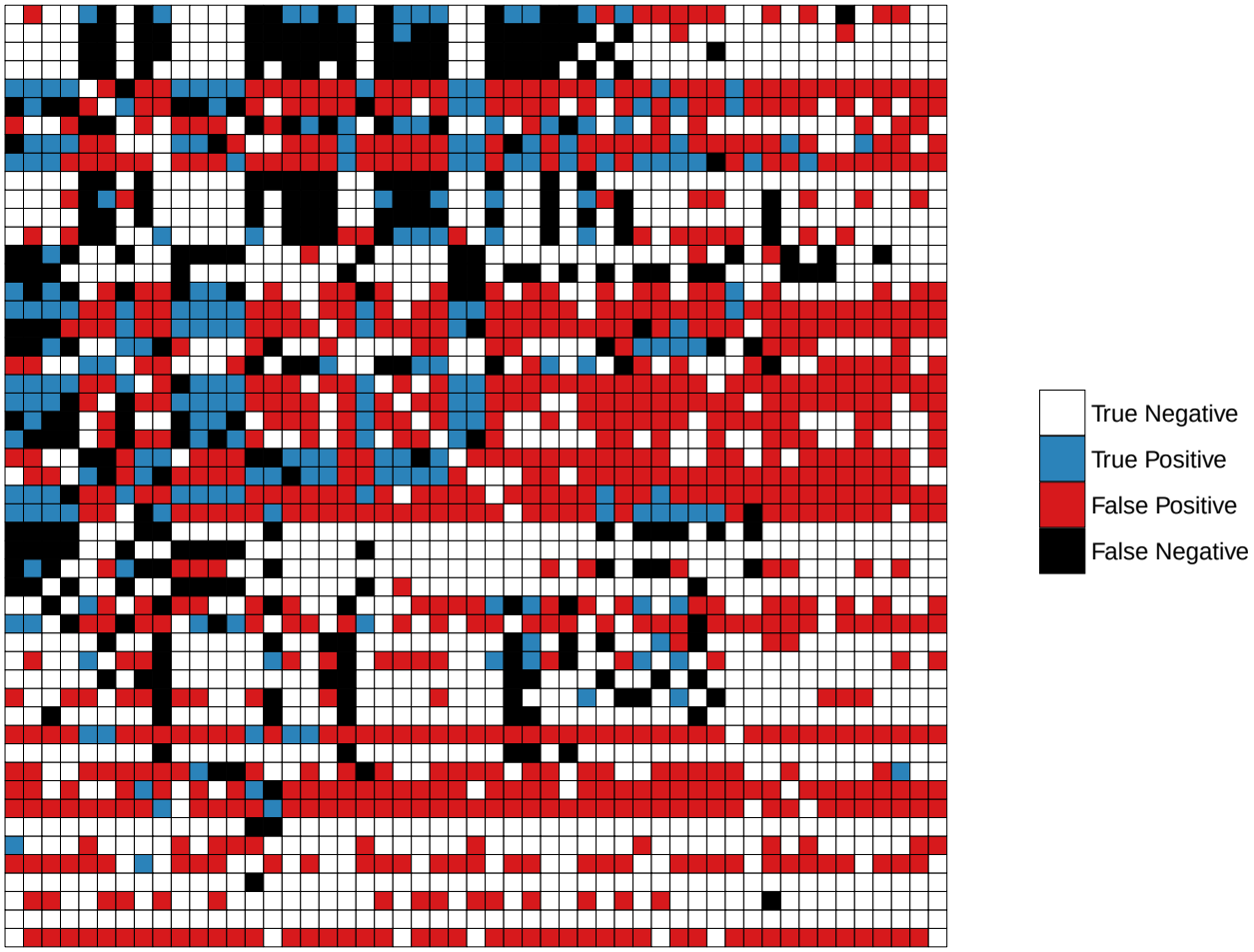

Figure 8.11: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh trophic network and the Lasso-2nd model.

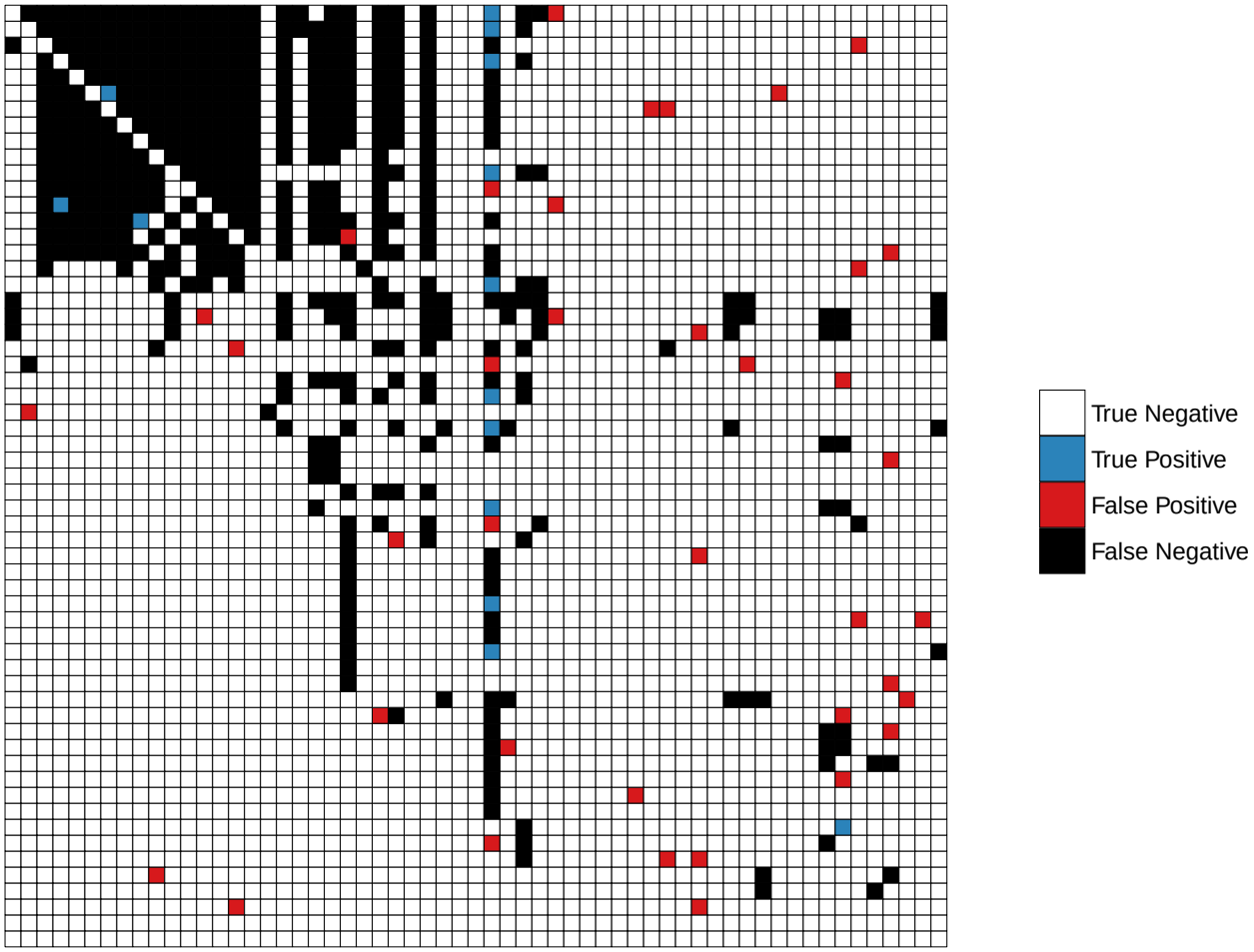

Figure 8.12: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh nontrophic network and the DBN-3 model.

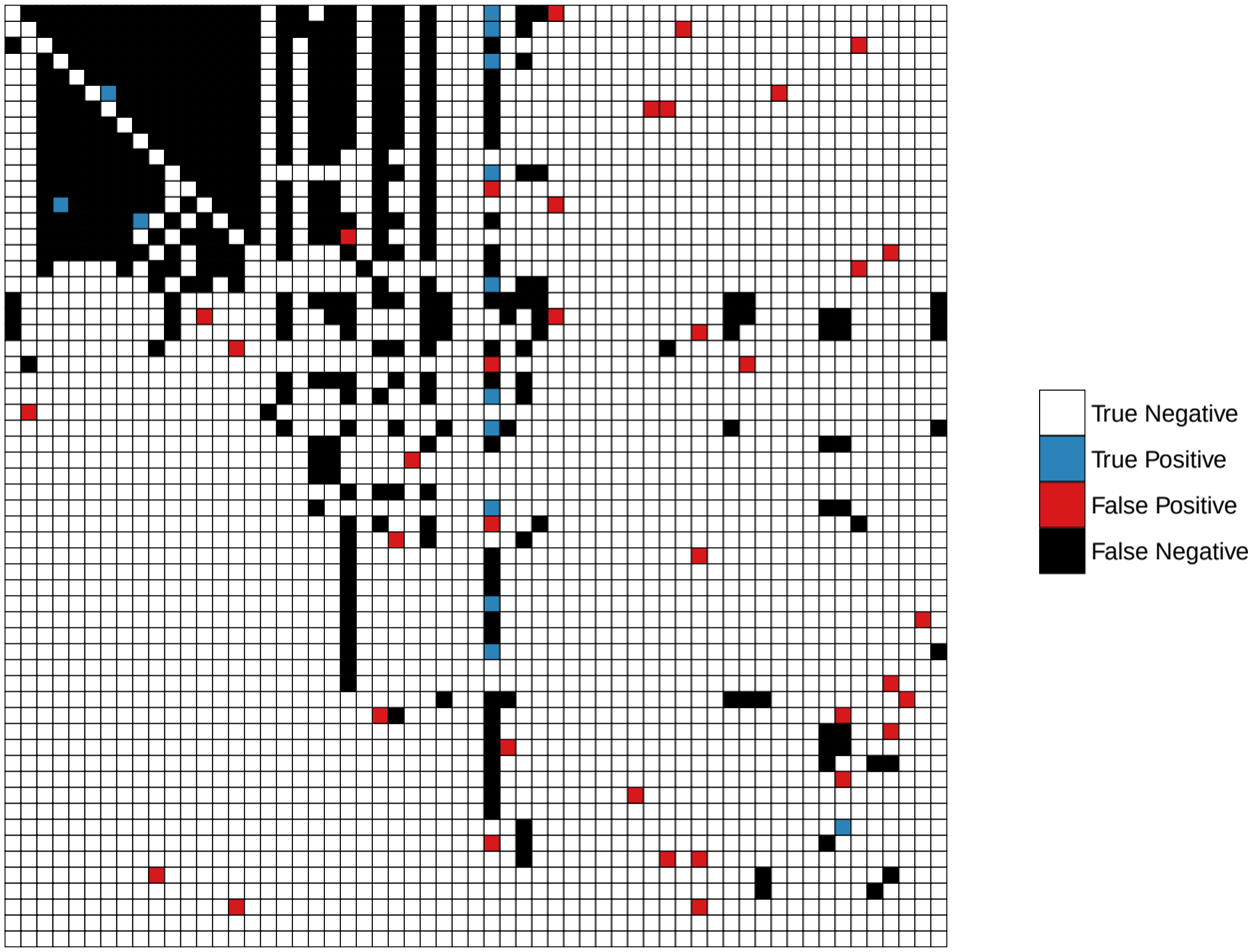

Figure 8.13: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh nontrophic network and the DBN-4 model.

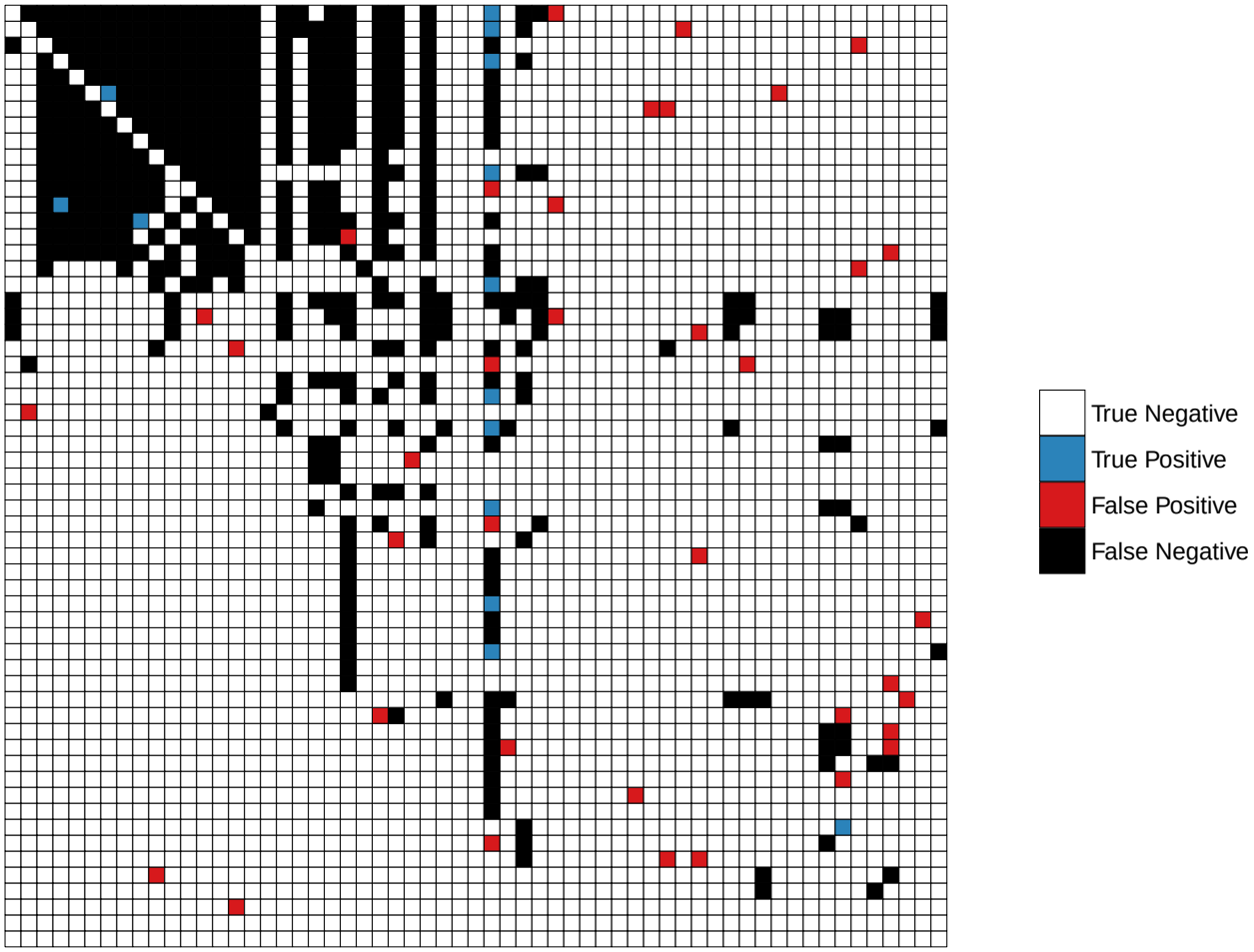

Figure 8.14: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh nontrophic network and the DBN-5 model.

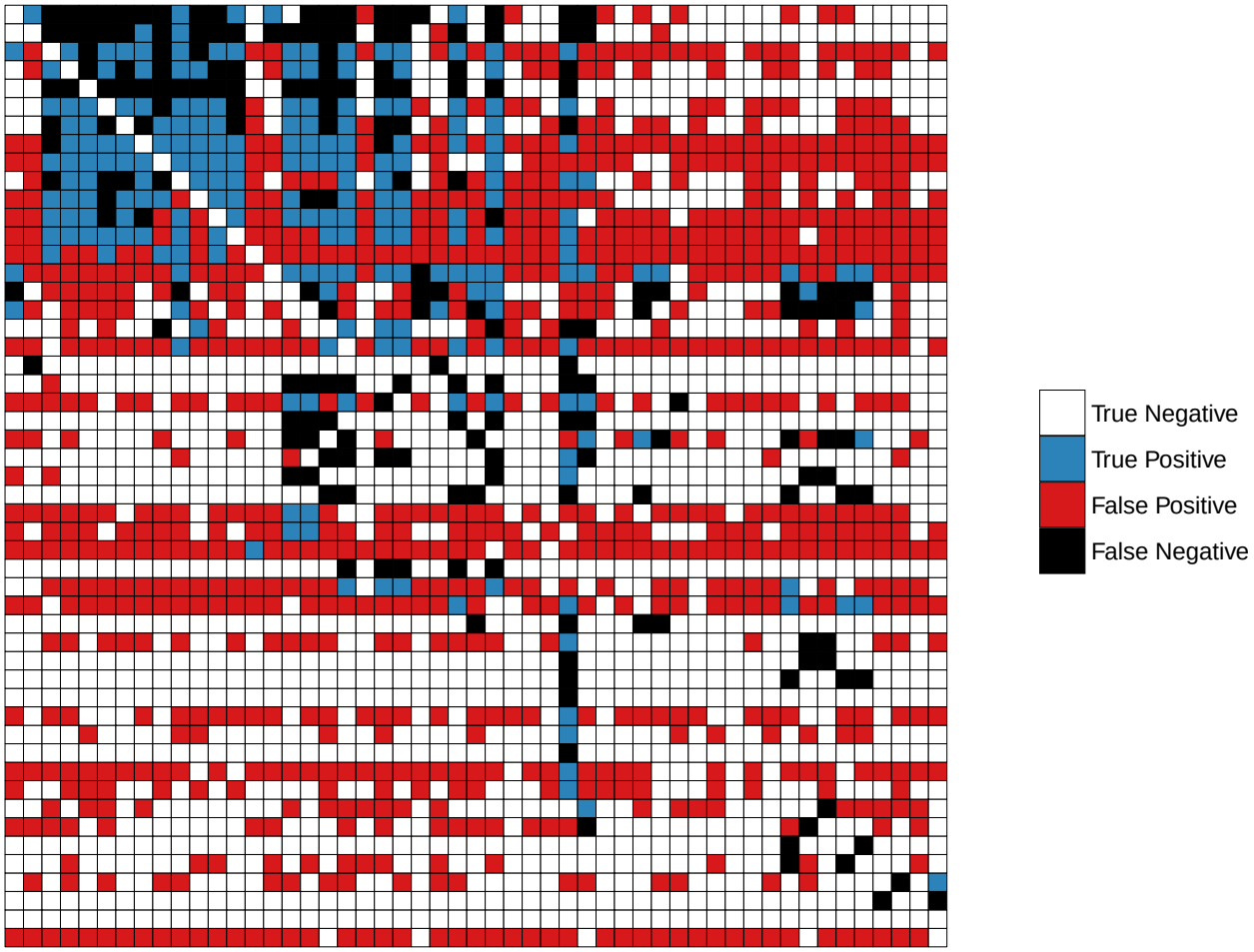

Figure 8.15: Structural comparison between empirical and model adjacency matrices as in Fig. 5.3, but for the Tatoosh nontrophic network and the Lasso-2nd model.

REFERENCES

[1] Encyclopedia of Life.

[2] C E Adams. Shift in pike Esox lucius predation pressure following the introduction of a new fish species to Loch Lomond. *Journal of Fish Biology*, 38:663–667, 1991.

[3] Lynn S Adler and Judith L Bronstein. Attracting Antagonists : Does Floral Nectar Increase Leaf Herbivory ? *Ecology*, 85(6):1519–1526, 2014.

[4] James K. Agee. *Fire Ecology of Pacific Northwest Forests*. Island press, 1996.

[5] Markos A Alexandrou, Bradley J Cardinale, John D Hall, F Delwiche, Keith J Fritschie, Anita Narwani, Patrick A Venail, Bastian Bentlage, M Sabrina Pankey, and Todd H Oakley. Evolutionary relatedness does not predict competition and co-occurrence in natural or experimental communities of green algae. *Proceedings of the Royal Society of London B*, pages 1–10, 2015.

[6] KR Allen. The Food and Migration of the Perch (Perca fluviatilis) in Windermere. *Journal of Animal Ecology*, 4(2):264–273, 1935.

[7] KR Allen. A Note on the Food of Pike (Esox lucius) in Windermere. *Journal of Animal Ecology*, 8(1):72–75, 1939.

[8] Stefano Allesina, David Alonso, and Mercedes Pascual. A General Model for Food Web Structure. *Science*, 320(5876):658–661, 2008.

[9] Stefano Allesina, Antonio Bodini, and Mercedes Pascual. Functional links and robustness in food webs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1524):1701–1709, 2009.

[10] Stefano Allesina and Mercedes Pascual. Network structure, predator - Prey modules, and stability in large food webs. *Theoretical Ecology*, 1(1):55–64, 2008.

[11] Stefano Allesina and Mercedes Pascual. Food web models: a plea for groups. *Ecology Letters*, 12(7):652–62, jul 2009.

[12] Stefano Allesina and Mercedes Pascual. Googling food webs: can an eigenvector measure species' importance for coextinctions? *PLoS computational biology*, 5(9):e1000494, oct 2009.

[13] Stefano Allesina and Si Tang. Stability criteria for complex ecosystems. *Nature*, 483(7388):205–8, mar 2012.

[14] Ahmet Alp, V. Yeen, M. Apaydin Yaci, Rahmi Uysal, E. Biçen, and A. Yaci. Diet composition and prey selection of the pike, Esox lucius, in Çivril Lake, Turkey. *Journal of Applied Ichthyology*, 24:670–677, 2008.

[15] Gautam Altekar, Sandhya Dwarkadas, John P Huelsenbeck, and Fredrik Ronquist. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics (Oxford, England)*, 20(3):407–15, feb 2004.

[16] David M. Althoff. Does parasitoid attack strategy influence host specificity? A test with New World braconids. *Ecological Entomology*, 28(4):500–502, 2003.

[17] Per-arne Amundsen, Kevin D Lafferty, Rune Knudsen, Raul Primicerio, Anders Klemetsen, and Armand M Kuris. Food web topology and parasites in the pelagic zone of a subarctic lake. *Journal of Animal Ecology*, 78:563–572, 2009.

[18] Roy M Anderson and Robert M May. Infectious Diseases and Population Cycles of Forest Insects. *Science*, 210(4470):658–661, 1980.

[19] Roger Arditi, Jerzy Michalski, and Alexandre H. Hirzel. Rheagogies: Modelling non-trophic effects in food webs. *Ecological Complexity*, 2(3):249–258, sep 2005.

[20] Robert A Armstrong and Richard McGehee. Competitive Exclusion. *The American Naturalist*, 115(2):151–170, 1980.

[21] JN Ball. On the food of the brown trout of Llyn Tegid. *Proceedings of the Zoological Society of London*, 137(4):599–622, 1961.

[22] N A.-E. Barak and C F Mason. Population density, growth and diet of eels, Anguilla anguilla L., in two rivers in eastern England. *Aquaculture and Fisheries Management*, 23:59–70, 1992.

[23] Albert Barberán, Scott T Bates, Emilio O Casamayor, and Noah Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, 6(10):343–351, 2011.

[24] Douglas Bates, Martin Mächler, Benjamin M Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[25] AM Beeton. Food Habits of the Burbot (Lota lota lacustris) in the White River, a Michigan Trout Stream. *Copeia*, 1:58–60, 1956.

[26] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, 1995.

[27] Eva Bergman. Changes in Abundance of Two Percids, Perca fluviatilis and Gymnocephalus ceruus, along a Productivity Gradient: Relations Feeding Strategies and Competitive Abilities. *Canadian Journal of Fisheries and Aquatic Sciences*, 48:536–545, 1991.

[28] Eric L Berlow, Jennifer A Dunne, Neo D Martinez, Philip B Stark, Richard J Williams, Ulrich Brose, and Simon A Levin. Simple Prediction of Interaction Strengths in Complex Food Webs. *Proceedings of the National Academy of Sciences*, 106(1):187–191, 2009.

[29] David Berry and Stefanie Widder. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5:1–14, 2014.

[30] Mani Bhushan, Raghunathan Rengaswamy, and Literature Survey. Design of sensor network based on the signed directed graph of the process for efficient fault diagnosis. *Industrial & engineering chemistry research*, 39(4):999–1019, 2000.

[31] Alice Boit, Neo D Martinez, Richard J Williams, and Ursula Gaedke. Mechanistic theory and modelling of complex food-web dynamics in Lake Constance. *Ecology Letters*, 15:594–602, jun 2012.

[32] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[33] Judith L Bronstein, William G Wilson, and William F Morris. Ecological dynamics of mutualist/antagonist communities. *The American Naturalist*, 162(4 Suppl):S24–39, oct 2003.

[34] Ulrich Brose, Lara Cushing, Eric L Berlow, Tomas Jonsson, Carolin Banašek-Richter, Louis-Félix Bersier, Julia L. Blanchard, Thomas Brey, Stephen R. Carpenter, Marie-France Cattin Blandenier, Joel E. Cohen, Hassan Ali Dawah, Tony Dell, François Edwards, Sarah Harper-Smith, Ute Jacob, Roland a. Knapp, Mark E Ledger, Jane Memmott, Katja Mintenbeck, John K. Pinnegar, Björn C. Rall, Tom Rayner, Liliane Ruess, Werner Ulrich, Philip Warren, Rich J. Williams, Guy Woodward, Peter Yodzis, and Neo D Martinez. Body Sizes of Consumers and Their Resources. *Ecology*, 86:2545–2545, 2005.

[35] Ulrich Brose, Richard J. Williams, and Neo D Martinez. Allometric scaling enhances stability in complex food webs. *Ecology Letters*, 9(11):1228–1236, 2006.

[36] Christopher Brown. hash: Full feature implementation of hash/associated arrays/dictionaries, 2013.

[37] R. N. B. Campbell. Food of an introduced population of pikeperch, Stizostedion lucioperca L., in lake Egirdir, Turkey. *Aquaculture and Fisheries Management*, 23(1):71–85, 1992.

[38] Angelo Canty and Brian Ripley. boot: Bootstrap R (S-Plus) Functions., 2016.

[39] Stephen R. Carpenter, James F. Kitchell, and James R. Hodgson. Cascading trophic interactions and lake productivity. *BioScience*, 35(10):634–639, 1985.

[40] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, Christian Von Mering, Ch Zu, and Limnological Station. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959, 2010.

[41] Hsuan-Wien Chen, Kwang-Tsao Shao, Chester Wai-Jen Liu, Wen-Hsieh Lin, and Wei-Chung Liu. The reduction of food web robustness by parasitism: fact and artefact. *International journal for parasitology*, 41(6):627–34, may 2011.

[42] Robert R. Christian and Joseph J. Luczkovich. Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. *Ecological Modelling*, 117:99–124, 1999.

[43] Alyssa R. Cirtwill and Daniel B Stouffer. Concomitant predation on parasites is highly variable but constrains the ways in which parasites contribute to food web structure. *Journal of Animal Ecology*, 84(3):734–744, 2015.

[44] Adam Thomas Clark, Hao Ye, Forest Isbell, Ethan Deyle, Jane Cowles, G David Tilman, and George Sugihara. Spatial convergent cross mapping to detect causal relationship from short time series. *Ecology*, 96(5):1174–1181, 2015.

[45] HP Clemens. The Food of the Burbot Lota Lota Maculosa (LeSueur) in Lake Erie. *Transactions of the American Fisheries Society*, 80(1):56–66, 1951.

[46] Sarah Cobey and Edward B Baskerville. Limits to causal inference with state-space reconstruction for infectious disease. *PLoS ONE*, 11(12):e0169050, 2016.

[47] Joel E. Cohen and C.M. Newman. A stochastic theory of community food webs I . Models and aggregated data. *Proceedings of the Royal Society of London B*, 224(1237):421–448, 1985.

[48] Lise Comte, Bernard Hugueny, and Gaël Grenouillet. Climate interacts with anthropogenic drivers to determine extirpation dynamics. *Ecography*, 39(10):1008–1016, 2015.

[49] Lise Comte, Bernard Hugueny, and Gaël Grenouillet. Data from: Climate interactis with anthropogenic drivers to determine extirpation dynamics. *Dryad Digital Repository*, 2016.

[50] Edward F Connor and Daniel Simberloff. Interspecific Competition and Species Co-Occurrence Patterns on Islands: Null Models and the Evaluation of Evidence. *Oikos*, 41(3):455–465, 1983.

[51] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, 2 edition, 2006.

[52] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Sy:1695, 2006.

[53] Diane W Davidson, Steven C Cook, Roy R Snelling, and Tock H Chua. Explaining the Abundance of Ants in Lowland Tropical Rainforest Canopies. *Science*, 300:969–972, 2003.

[54] AC Davison and DV Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997.

[55] André M de Roos and Lennart Persson. Size-dependent life-history traits promote catastrophic collapses of top predators. *Proceedings of the National Academy of Sciences*, 99(20):12907–12912, 2002.

[56] A. de Sostoa and J. Lobon-Cervia. Observations on feeding relationships between fish predators and fish assemblages in a Mediterranean stream. *Regulated Rivers: Research & Management*, 4(2):157–163, 1989.

[57] C.L. Deelder. A contribution to knowledge of the stunted growth of perch (Perca fluviatilis L.) in Holland. *Hydrobiologia*, 3(4):357–378, 1951.

[58] Julie Sloan Denslow. Gap Partitioning among Tropical Rainforest Trees. *Biotropica*, 12(2):47–55, 1980.

[59] JM Diamond. Assembly of species communities. In ML Cody and JM Diamond, editors, *Ecology and evolution of communities*, pages 342–444. Harvard University Press, Cambridge, Massachusetts, 1975.

[60] James S. Diana. The feeding pattern and daily ration of a top carnivore, the northern pike (Esox lucius). *Canadian Journal of Zoology*, 57(11):2121–2127, 1979.

[61] J Dominguez and J C Pena. SPATIO-TEMPORAL VARIATION IN THE DIET OF NORTHERN PIKE (Esox lucius) IN A COLONISED AREA (ESLA BASIN , NW SPAIN). *Limnetica*, 19:1–20, 2000.

[62] Hendrik Dörner and J. Benndorf. Piscivory by large eels on young-of-the-year fishes: Its potential as a biomanipulation tool. *Journal of Fish Biology*, 62(2):491–494, 2003.

[63] Hendrik Dörner, Søren Berg, Lene Jacobsen, Stephan Hülsmann, Mads Brojerg, and Annekatrin Wagner. The feeding behaviour of large perch Perca fluviatilis (L.) in relation to food availability: A comparative study. *Hydrobiologia*, 506-509:427–434, 2003.

[64] Hendrik Dörner, Stephan Hülsmann, F. Hölker, Christian Skov, and Annekatrin Wagner. Size-dependent predator-prey relationships between pikeperch and their prey fish. *Ecology of Freshwater Fish*, 16(3):307–314, 2007.

[65] Hendrik Dörner and Annekatrin Wagner. Size-dependent predator-prey relationships between perch and their fish prey. *Journal of Fish Biology*, 62(5):1021–1032, 2003.

[66] David Duggins. Kelp Beds and Sea Otters : An Experimental Approach. *Ecology*, 61(3):447–453, 1980.

[67] JA Dunne, KD Lafferty, AP Dobson, RF Hechinger, AM Kuris, ND Martinez, JP McLaughlin, KN Mouritsen, R Poulin, K Reise, DB Stouffer, DW Thieltges, RJ Williams, and CD Zander. Data from: Parasites affect food web structure primarily through increased diversity and complexity. *PLoS Biology*, 2013.

[68] Jennifer A Dunne, Kevin D Lafferty, Andrew P Dobson, Ryan F Hechinger, Armand M Kuris, Neo D Martinez, John P McLaughlin, Kim N Mouritsen, Robert Poulin, Karsten Reise, Daniel B Stouffer, David W Thieltges, Richard J Williams, and Claus Dieter Zander. Parasites affect food web structure primarily through increased diversity and complexity. *PLoS Biology*, 11(6):e1001579, jan 2013.

[69] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12917–22, oct 2002.

[70] Jennifer A Dunne, Richard J. Williams, and Neo D Martinez. Network structure and biodiversity loss in food webs: Robustness increases with connectance. *Ecology Letters*, 5(4):558–567, 2002.

[71] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Network structure and robustness of marine food webs. *Marine Ecology Progress Series*, 273:291–302, 2004.

[72] Greg Dwyer, Jonathan Dushoff, and Susan Harrell Yee. The combined effects of pathogens and predators on insect outbreaks. *Nature*, 430:341–345, 2004.

[73] L A Dyer, M S Singer, J T Till, J O Stireman, G L Gentry, R J Marquis, R E Ricklefs, H F Greeney, D L Wagner, H C Morais, I R Diniz, T A Kursar, and P D Coley. Host specificity of Lepidoptera in tropical and temperate forests. *Nature*, 448(August):696–700, 2007.

[74] David J Earl and Michael W Deem. Parallel tempering: theory, applications, and new perspectives. *Physical chemistry chemical physics*, 7(23):3910–6, dec 2005.

[75] Anna Eklöf, Matthew R Helmus, M Moore, and Stefano Allesina. Relevance of evolutionary history for food web structure. *Proceedings of the Royal Society of London B*, 279(1733):1588–96, apr 2012.

[76] P Eklöv and SF Hamrin. Predatory Efficiency and Prey Selection: Interactions between Pike Esox lucius , Perch Perca fluviatilis and Rudd Scardinus erythrophthalmus. *Oikos*, 56(2):149–156, 1989.

[77] B. Elvira, G.G. Nicola, and A. Almodovar. Pike and red swamp crayfish: a new case on predator prey relationship between aliens in central Spain. *Journal of Fish Biology*, 48:437–446, 1996.

[78] James A. Estes. Killer Whale Predation on Sea Otters Linking Oceanic and Nearshore Ecosystems. *Science*, 282(5388):473–476, 1998.

[79] S. Estlander, L. Nurminen, M. Olin, Mika Vinni, S. Immonen, Martti Rask, J. Ruuhijärvi, J. Horppila, and Hannu Lehtonen. Diet shifts and food selection of perch Perca fluviatilis and roach Rutilus rutilus in humic lakes of varying water colour. *Journal of Fish Biology*, 77(1):241–256, 2010.

[80] Ernesto Estrada. Characterization of topological keystone species. Local, global and "meso-scale" centralities in food webs. *Ecological Complexity*, 4(1-2):48–57, 2007.

[81] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews*, 10:538–550, 2012.

[82] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*, 8(7):e1002606, 2012.

[83] Colin Fontaine, Paulo R Guimarães, Sonia Kéfi, Nicolas Loeuille, Jane Memmott, Wim H van der Putten, Frank J F van Veen, and Elisa Thébault. The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, 14(11):1170–81, nov 2011.

[84] Colin Fontaine, Elisa Thébault, and Isabelle Dajoz. Are insect pollinators more generalist than insect herbivores? *Proceedings of the Royal Society of London B*, 276(1669):3027–33, aug 2009.

[85] Linton C. Freeman. Centrality in social networks: conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[86] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[87] Keith J Fritschie, Bradley J Cardinale, Markos A Alexandrou, and Todd H Oakley. Evolutionary history and the strength of species interactions: testing the phylogenetic limiting similarity hypothesis. *Ecology*, 95(5):1407–1417, 2014.

[88] R Froese and D Pauly, editors. *FishBase 2000: concepts, design and data sources. ICLARM, Los Ban Os.* Laguna, Philippines, 2000.

[89] Winifred E Frost. River Liffey survey - II. The food consumed by the brown trout (Salmo trutta Linn.) in acid and alkaline waters. *Proceedings of the Royal Irish Academy. Section B: Biological, Geological, and Chemical Science*, 45:139–206, 1938.

[90] Winifred E Frost. Observations on the Food of Eels (Anguilla anguilla) from the Windermere Catchment Area. *Journal of Animal Ecology*, 15(1):43–53, 1946.

[91] Winifred E Frost. The Food of Pike , Esox lucius L ., in Windermere. *Journal of Animal Ecology*, 23(2):339–360, 1954.

[92] Colin P. Gallagher and Terry A. Dick. Winter feeding ecology and the importance of cannibalism in juvenile and adult burbot (Lota lota) from the Mackenzie Delta, Canada. *Hydrobiologia*, 757(1):73–88, 2015.

[93] P Gaudin and L Calliere. Experimental study of the influence of presence and predation by sculpin, Cottus gobio L., on the drift of emergent brown trout, Salmo trutta L. *Archiv fur Hydrobiologie*, 147(3):257–271, 2000.

[94] C J Geyer. Markov chain Monte Carlo maximum likelihood. In Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, 1991.

[95] Kai Ginter, Külli Kangur, Andu Kangur, Peeter Kangur, and Marina Haldna. Diet patterns and ontogenetic diet shift of Pikeperch, Sander lucioperca (L.) fry in lakes Peipsi and Võrtsjärv (Estonia). *Hydrobiologia*, 660(1):79–91, 2011.

[96] Nicholas J Gotelli and Declan J. McCabe. Species co-occurrence: A meta-analysis of J. M. Diamond's assembly rules model. *Ecology*, 83(8):2091–2096, 2002.

[97] Alexandra Goudard and Michel Loreau. Nontrophic interactions, biodiversity, and ecosystem functioning: an interaction web model. *The American Naturalist*, 171(1):91–106, jan 2008.

[98] J. Grey, S. J. Thackeray, R. I. Jones, and A. Shine. Ferox trout (Salmo trutta) as 'Russian dolls': Complementary gut content and stable isotope analyses of the Loch Ness foodweb. *Freshwater Biology*, 47(7):1235–1243, 2002.

[99] Nabil Guelzim, Samuele Bottani, Paul Bourgine, and François Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature genetics*, 31(1):60–63, 2002.

[100] Jan O. Haerter, Namiko Mitarai, and Kim Sneppen. Food Web Assembly Rules for Generalized Lotka-Volterra Equations. *PLoS Computational Biology*, 12(2):1–17, 2016.

[101] S J Hall and D Raffaelli. Food-web patterns: lessons from a species-rich web. *Journal of Animal Ecology*, 60:823–842, 1991.

[102] Sture Hansson, Fredrik Arrhenius, and Sture Nellbring. Diet and growth of pikeperch (Stizostedion lucioperca L.) in a Baltic Sea area. *Fisheries Research*, 31:163–167, 1997.

[103] PHT Hartley. The food of coarse fish, being the Interim Report on the Coarse Fish Investigation. *Scientific Publications of the Freshwater Biological Association of the British Empire*, pages 3–33, 1940.

[104] PHT Hartley. The Natural History of some British Freshwater Fishes. *Proceedings of the Zoological Society of London*, 117:129–206, 1947.

[105] PHT Hartley. Food and Feeding Relationships in a Community of Fresh-Water Fishes. *Journal of Animal Ecology*, 17(1):1–14, 1948.

[106] Ryan F Hechinger, Kevin D Lafferty, John P McLaughlin, Brian L Fredensborg, Todd C Huspeni, Julio Lorda, Parwant K Sandhu, Jenny C Shaw, Mark E Torchin, Kathleen L Whitney, and KM Kuris. Food webs including parasites, biomass, body sizes, and life stages for three California/Baja California estuaries. *Ecology*, 92(3):791–791, 2011.

[107] David Heckerman. A Tutorial on Learning with Bayesian Networks. In Dawn E Holmes and Lakhmi C Jain, editors, *Innovations in Bayesian Networks: Theory and Applications*, chapter 3, pages 33–82. Springer, Berlin, 2008.

[108] David Heckerman, Dan Geiger, and David M Chlckering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *AAAI Technical Report*, WS-94-03, 1995.

[109] Stein Joar Hegland, Anders Nielsen, Amparo Lázaro, Anne Line Bjerknes, and Ørjan Totland. How does climate warming affect plant-pollinator interactions? *Ecology Letters*, 12(2):184–195, 2009.

[110] David W Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, 2nd edition, 2005.

[111] RL Hunt. Food of Northern Pike in a Wisconsin Trout Stream. *Transactions of the American Fisheries Society*, 94(1):95–97, 1965.

[112] Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[113] Mark Huxham, S Beaney, and Dave Raffaelli. Do parasites reduce the chances of triangulation in a real food web? *Oikos*, 76(2):284–300, 1996.

[114] Mark Huxham, Dave Raffaelli, and Alan Pike. Parasites and Food Web Patterns. *Journal of Animal Ecology*, 64(2):168–176, 1995.

[115] N A Hvidsten and P I Møkkelgjerd. Predation on salmon smolts, Salmo salar L., in the estuary of the River Surna, Norway. *Journal of Fish Biology*, 30(3):273–280, 1987.

[116] HBN Hynes. The Food of Fresh-Water Sticklebacks (Gasterosteus aculeatus and Pygosteus pungitius), with a Review of Methods Used in Studies of the Food of Fishes. *Journal of Animal Ecology*, 19(1):36–58, 1950.

[117] P. Hyvärinen and A. Huusko. Diet of brown trout in relation to variation in abundance and size of pelagic fish prey. *Journal of Fish Biology*, 68(1):87–98, 2006.

[118] C Idyll. Food of Rainbow, Cutthroat and Brown Trout in the Cowichan River System, B.C. *Journal of the Fisheries Research Board of Canada*, 5c(5):448–458, 1942.

[119] Jan Arge Jacobsen and Lars Petter Hansen. Feeding habits of wild and escaped farmed Atlantic salmon, Salmo salar L., in the Northeast Atlantic. *ICES Journal of Marine Science*, 58(4):916–933, 2001.

[120] Jean-Louis Jamet. Feeding activity of roach (Rutilus rutilus (L.)), perch (Perca fluviatilis, L.) and ruffe (Gymnocephalus cernuus (L.)) in eutrophic lake Aydat (France). *Aquatic Sciences*, 56(4):376–387, 1994.

[121] H. Jensen, M. Kiljunen, and Per-Arne Amundsen. Dietary ontogeny and niche shift to piscivory in lacustrine brown trout Salmo trutta revealed by stomach content and stable isotope analyses. *Journal of Fish Biology*, 80(7):2448–2462, 2012.

[122] Niels Jepsen, Susanne Pedersen, and Eva Thorstad. Behavioural interactions between prey (trout smolts) and predators (pike and pikeperch) in an impounded river. *Regulated Rivers: Research & Management*, 16(2):189–198, 2000.

[123] Ferenc Jordán, Wei-chung Liu, and A J Davis. Topological Keystone Species : Measures of Positional Importance in Food Webs in of positional Topological keystone species : measures importance food webs. *Oikos*, 112(May):535–546, 2014.

[124] Ferenc Jordán, Thomas a. Okey, Barbara Bauer, and Simone Libralato. Identifying important species: Linking structure and function in ecological networks. *Ecological Modelling*, 216(1):75–80, aug 2008.

[125] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[126] K. Kahilainen and Hannu Lehtonen. Piscivory and prey selection of four predator species in a whitefish dominated subarctic lake. *Journal of Fish Biology*, 63(3):659–672, 2003.

[127] S Kålås. The ecology of ruffe, Gymnocephalus cernuus (Pisces: percidae) Introduced to Mildevatn, western Norway. *Environmental Biology of Fishes*, 42(3):219–232, 1995.

[128] Jason M Kamilar and Justin A Ledogar. Species Co-Occurrence Patterns and Dietary Resource Competition in Primates Species. *American Journal of Physical Anthropology*, 144:131–139, 2011.

[129] Andu Kangur and Peeter Kangur. Diet composition and size-related changes in the feeding of pikeperch, Stizostedion lucioperca (Percidae) and pike, Esox lucius (Esocidae) in the Lake Peipsi (Estonia). *Italian Journal of Zoology*, 65:255–259, 1998.

[130] Peeter Kangur, Andu Kangur, and Külli Kangur. Dietary importance of various prey fishes for pikeperch Sander lucioperca (L.) in large shallow lake Võrtsjärv (Estonia). *Proceedings of the Estonian Academy of Sciences: Biology, Ecology*, 56(2):154–167, 2007.

[131] Lars Karlsson, Erkki Ikonen, Andis Mitans, and Sture Hansson. The Diet of Salmon (Salmo salar) in the Baltic Sea and Connections with the M74 Syndrome. *Ambio*, 28(1):37–42, 1999.

[132] Brian Karrer and M E J Newman. Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(1):016107, aug 2011.

[133] Sonia Kéfi, Eric L Berlow, Evie a Wieters, Sergio A. Navarrete, Owen L Petchey, Spencer a Wood, Alice Boit, Lucas N Joppa, Kevin D Lafferty, Richard J Williams, Neo D Martinez, Bruce A Menge, Carol a Blanchette, Alison C Iles, and Ulrich Brose. More than a meal... integrating non-feeding interactions into food webs. *Ecology Letters*, 15:291–300, feb 2012.

[134] Sonia Kéfi, Vincent Miele, Evie A. Wieters, Sergio A. Navarrete, and Eric L Berlow. How Structured Is the Entangled Bank? The Surprisingly Simple Organization of Multiplex Ecological Networks Leads to Increased Persistence and Resilience. *PLOS Biology*, 14(8):e1002527, 2016.

[135] J. Kekäläinen, T Niva, and H. Huuskonen. Pike predation on hatchery-reared Atlantic salmon smolts in a northern Baltic river. *Ecology of Freshwater Fish*, 17(1):100–109, 2008.

[136] Bruce E Kendall, Cheryl J Briggs, William W Murdoch, Peter Turchin, Stephen P Ellner, Ed McCauley, Roger M Nisbet, and Simon N Wood. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80(6):1789–1805, 1999.

[137] T. Keskinen and T. J. Marjomäki. Diet and prey size spectrum of pikeperch in lakes in central Finland. *Journal of Fish Biology*, 65(4):1147–1153, 2004.

[138] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology*, 11(5):e1004226, 2015.

[139] J H L'Abée-Lund, P. Aass, and H. Sægrov. Long-term variation in piscivory in a brown trout population : effect of changes in available prey organisms. *Ecology of Freshwater Fish*, 11:260–269, 2002.

[140] J H L'Abée-Lund, A Langeland, and H Saegrov. Piscivory by brown trout Salmo trutta L. and Arctic charr Salvelinus alpinus (L.) in Norwegian lakes. *Journal of Fish Biology*, 41:91–101, 1992.

[141] Kevin D Lafferty, Stefano Allesina, Matias Arim, Cherie J Briggs, Giulio De Leo, Andrew P Dobson, Jennifer A Dunne, Pieter T J Johnson, Armand M Kuris, David J Marcogliese, Neo D Martinez, Jane Memmott, Pablo a Marquet, John P McLaughlin, Erin a Mordecai, Mercedes Pascual, Robert Poulin, and David W Thieltges. Parasites in food webs : the ultimate missing links. *Ecology Letters*, 11(6):533–46, jun 2008.

[142] Kevin D Lafferty, Andrew P Dobson, and Armand M Kuris. Parasites dominate food web links. *Proceedings of the National Academy of Sciences of the United States of America*, 103(30):11211–6, jul 2006.

[143] Eddy H R R Lammens and Jentsje T. Visser. Variability of mouth width in European eel, Anguilla anguilla, in relation to varying feeding conditions in three Dutch lakes. *Environmental Biology of Fishes*, 26(1):63–75, 1989.

[144] Armin Landmann and Norbert Winding. Niche segregation in high-altitude Himalayan chats (Aves, Turdidae): does morphology match ecology? *Oecologia*, 95:506–519, 1993.

[145] A Langeland, J H L'Abée-Lund, Bror Jonsson, and N Jonsson. Resource Partitioning and Niche Shift in Arctic Charr Salvelinus alpinus and Brown Trout Salmo trutta. *Journal of Animal Ecology*, 60(3):895–912, 1991.

[146] P-O. Larsson. Predation on migrating smolt as a regulating factor in Baltic salmon, Salmo salar L., populations. *Journal of Fish Biology*, 26:391–397, 1985.

[147] GH Lawler. The Food of the Pike Esox lucius, in Heming Lake, Manitoba. *Journal of the Fisheries Research Board of Canada*, 22(6):1357–1377, 1965.

[148] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed Networks in Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370, 2010.

[149] P. M. Leunda, J. Oscoz, B. Elvira, A. Agorreta, S. Perea, and R. Miranda. Feeding habits of the exotic black bullhead Ameiurus melas (Rafinesque) in the Iberian Peninsula: First evidence of direct predation on native fish species. *Journal of Fish Biology*, 73(1):96–114, 2008.

[150] Ping Li, C Zhang, E J Perkins, P Gong, and Y Deng. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics*, 8(Suppl 7):S13, 2007.

[151] Leif Lien. Biology of the Minnow Phoxinus phoxinus and Its Interactions with Brown Trout Salmo in Øvre Heimdalsvatn , Norway. *Holarctic Ecology*, 4(3):191–200, 1981.

[152] Carmen Maria Livi, Ferenc Jordán, Paola Lecca, and Thomas A. Okey. Identifying key species in ecosystems with stochastic sensitivity analysis. *Ecological Modelling*, 222(14):2542–2551, 2011.

[153] M Lorenzoni, M Corboli, AJM Dörr, G Giovinazzo, S Selvi, and M Mearelli. Diets of Micropterus Salmoides Lac. and Esox Lucius L. in Lake Trasimeno (Umbria, Italy) and Their Diet Overlap. *Bulletin Français de la Pêche et de la Pisciculture*, 365/366:537–547, 2002.

[154] John E. Losey and Robert F. Denno. Positive predator-predator interactions: Enhanced predation rates and synergistic suppression of aphid populations. *Ecology*, 79(6):2143–2152, 1998.

[155] Joseph J. Luczkovich, Stephen P. Borgatti, Jeffrey C. Johnson, and Martin G. Everett. Defining and Measuring Trophic Role Similarity in Food Webs Using Regular Equivalence. *Journal of Theoretical Biology*, 220(3):303–321, feb 2003.

[156] R. H. K. Mann. Observations on the age, growth, reproduction and food of the pike Esox lucius (L.) in the two rivers in southern England. *Journal of Fish Biology*, 8:179–197, 1976.

[157] R. H. K. Mann. The Annual Food Consumption and Prey Preferences of Pike (Esox lucius) in the River Frome, Dorset. *Journal of Animal Ecology*, 51(1):81–95, 1982.

[158] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[159] Daniel Marbach, JC Costello, Robert Küffner, and NM Vega. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

[160] David J Marcogliese. Food webs and biodiversity: are parasites the missing link? *Journal of Parasitology*, 89(6):S106–S113, 2003.

[161] David J Marcogliese and David K Cone. Food webs: a plea for parasites. *Trends in Ecology & Evolution*, 12(8):320–324, 1997.

[162] Neo D Martinez, Bradford A Hawkins, Hassan Ali Dawah, and Brian P Feifarek. Effects of sampling effort on characterization of food-web structure. *Ecology*, 80(3):1044–1055, 1999.

[163] R Alastair Mathers and Peter H Johansen. The effects of feeding ecology on mercury accumulation in walleye (Stizostedion Vitreum) and pike (Esox lucius) in Lake Simcoe. *Canadian Journal of Zoology*, 63:2006–2012, 1985.

[164] WL Mauck and DW Coble. The Vulnerability of Some Fishes to Northern Pike (Esox lucius) Predation. *Journal of the Fisheries Research Board of Canada*, 28(7):957–969, 1971.

[165] Robert M May. Qualitative Stability in Model Ecosystems. *Ecology*, 54(3):638–641, 1973.

[166] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt, New York, 2013.

[167] Andrew McAfee and Erik Brynjolfsson. Big Data. *Harvard Business Review*, 90(10):61–67, 2012.

[168] Jean C McCormack. The Food of Young Trout (Salmo trutta) in Two Different Becks. *Journal of Animal Ecology*, 31(2):305–316, 1962.

[169] Carlos J. Melián, Jordi Bascompte, Pedro Jordano, and Vlastimil Krivan. Diversity in a complex ecological network with two interaction types. *Oikos*, 118(1):122–130, jan 2009.

[170] Jane Memmott, Nickolas M Waser, and Mary V Price. Tolerance of pollination networks to species extinctions. *Proceedings. Biological sciences / The Royal Society*, 271(1557):2605–11, dec 2004.

[171] Bruce a Menge. Organization of the New England rocky intertidal community: role of predation, competition, and environmental heterogeneity. *Ecological Monographs*, 46:355–393, 1976.

[172] Bruce A Menge. Top-down and bottom-up community regulation in marine rocky intertidal habitats. *Journal of Experimental Marine Biology and Ecology*, 250(1-2):257–289, 2000.

[173] Matthew Joseph Michalska-smith. The Effect of Intra- and Interspecific Competition on Coexistence in Multispecies Communities. *The American Naturalist*, 188(1):E1–E12, 2016.

[174] Isobel Milns, Colin M Beale, and V Anne Smith. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91(7):1892–1899, 2010.

[175] a Mougi and M Kondoh. Diversity of interaction types and ecological community stability. *Science (New York, N.Y.)*, 337(6092):349–51, jul 2012.

[176] K. N. Mouritsen, R. Poulin, J. P. McLaughlin, and D. W. Thieltges. Food web including metazoan parasites for an intertidal ecosystem in New Zealand. *Ecology*, 92(10):2006–2006, 2011.

[177] Kim N Mouritsen, Robert Poulin, John P McLaughlin, and David W. Thieltges. Food web including metazoan parasites for an intertidal ecosystem in New Zealand. *Ecology*, 92:2006, 2011.

[178] J. Museth, R. Borgstrøm, T. Hame, and L. Å Holen. Predation by brown trout: A major mortality factor for sexually mature European minnows. *Journal of Fish Biology*, 62(3):692–705, 2003.

[179] Jií Musil and Zdeněk Adámek. Piscivorous fishes diet dominated by the Asian cyprinid invader, topmouth gudgeon (Pseudorasbora parva). *Biologia*, 62(4):488–490, 2007.

[180] Anita Narwani, Markos a Alexandrou, Todd H Oakley, Ian T Carroll, and Bradley J Cardinale. Experimental evidence that evolutionary relatedness does not affect the ecological mechanisms of coexistence in freshwater green algae. *Ecology Letters*, 16(11):1373–81, nov 2013.

[181] R Nastova-Gjorgjioska, V Kostov, and S Georgiev. Nutrition of chub Leuciscus cephalus (Linnaeaus, 1758) from the river Babuna. *Croatian Journal of Fisheries*, 55(1):53–65, 1997.

[182] Chris J. Needham, James R. Bradford, Andrew J. Bulpitt, and David R. Westhead. A primer on learning in Bayesian networks for computational biology. *PLoS Computational Biology*, 3(8):1409–1416, 2007.

[183] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[184] T Niva. Relations between diet, growth, visceral lipid content and yield of the stocked brown trout in three small lakes in northern Finland. *Annales Zoologici Fennici*, 36(2):103–120, 1999.

[185] Derek H Ogle. A Synopsis of the Biology and Life History of Ruffe. *Journal of Great Lakes Research*, 24(2):170–185, 1998.

[186] Derek H Ogle, James H Selgeby, JF Savino, Raymond M Newman, and Mary G Henry. Predation on Ruffe by Native Fishes of the St. Louis River Estuary, Lake Superior, 1989-1991. *North American Journal of Fisheries Management*, 16(1):115–123, 1996.

[187] Thomas A. Okey, Stuart Banks, Abraham F. Born, Rodrigo H. Bustamante, Mónica Calvopiña, Graham J. Edgar, Eduardo Espinoza, José Miguel Fariña, Lauren E. Garske, Günther K. Reck, Sandie Salazar, Scoresby Shepherd, Veronica Toral-Granda, and Petra Wallem. A trophic model of a Galápagos subtidal rocky reef for evaluating fisheries and conservation strategies. *Ecological Modelling*, 172(2-4):383–401, 2004.

[188] Toshinori Okuyama and J Nathaniel Holland. Network structural properties mediate the stability of mutualistic communities. *Ecology Letters*, 11(3):208–16, mar 2008.

[189] Jens M Olesen, Jordi Bascompte, Yoko L Dupont, Heidi Elberling, Claus Rasmussen, and Pedro Jordano. Missing and forbidden links in mutualistic networks. *Proceedings of the Royal Society of London B*, 278(1706):725–32, mar 2011.

[190] Sylvia Opitz. *Trophic Interactions in Caribbean Coral Reefs*. WorldFish, Manila, 1996.

[191] Robert T Paine. Food Web Complexity and Species Diversity. *The American Naturalist*, 100(910):65–75, 1966.

[192] Robert T Paine. A note on trophic complexity and community stability. *The American Naturalist*, 103(929):91–93, 1969.

[193] Robert T Paine. Intertidal Community Structure: Experimental Studies on the Relationship between a Dominant Competitor and Its Principal Predator. *Oecologia*, 15(2):93–120, 1974.

[194] Robert T Paine. Disaster , Catastrophe , and Local Persistence of the Sea Palm Postelia palmaeformis. *Science*, 205(4407):685–687, 1979.

[195] Robert T Paine. Food webs: Linkage, Interaction Strength and Community Infrastracture. *Journal of Animal Ecology*, 49(3):666–685, 1980.

[196] Robert T Paine and Simon A Levin. Intertidal Landscapes : Disturbance and the Dynamics of Pattern. *Ecological Monographs*, 51(2):145–178, 1981.

[197] Todd M Palmer, Maureen L Stanton, Truman P Young, Jacob R Goheen, Robert M Pringle, and Richard Karban. Breakdown of an ant-plant mutualism follows the loss of large herbivores from an African savanna. *Science (New York, N.Y.)*, 319(5860):192–5, jan 2008.

[198] Milena Pavlović, Momir Paunović, and Vladica Simić. Feeding of Eurasian perch (Perca fluviatilis L.) in three reservoirs in Serbia. *Water Research and Management*, 3(4):41–46, 2013.

[199] R B Pedley and J W Jones. The comparative feeding behaviour of brown trout, Salmo trutta L. and Atlantic salmon, Salmo salar L. in Llyn Dwythwch, Wales. *Journal of Fish Biology*, 12(3):239–256, 1978.

[200] Zeynep Pekcan-Hekim and Jyrki Lappalainen. Effects of clay turbidity and density of pikeperch (Sander lucioperca) larvae on predation by perch (Perca fluviatilis). *Naturwissenschaften*, 93(7):356–359, 2006.

[201] Xingxing Peng, Feng Guo, Feng Ju, and Tong Zhang. Shifts in the microbial community, nitrifiers and denitrifiers in the biofilm in a full-scale rotating biological contactor. *Environmental Science and Technology*, 48(14):8044–8052, 2014.

[202] FTK Pentelow. The Food of the Brown Trout (Salmo trutta L.). *Journal of Animal Ecology*, 1(2):101–107, 1932.

[203] Anders Persson and Lars-Anders Hansson. Diet shift in fish following competitive release. *Canadian Journal of Fisheries and Aquatic Sciences*, 56:70–78, 1999.

[204] Lennart Persson. Effects of Reduced Interspecific Competition on Resource Utilization in Perch (Perca Fluviatilis). *Ecology*, 67(2):355–364, 1986.

[205] O L Petchey, P T McPhearson, T M Casey, and P J Morin. Environmental warming alters food-web structure and ecosystem function. *Nature*, 402(November):69–72, 1999.

[206] Owen L Petchey, Andrew P Beckerman, Jens O Riede, and Philip H Warren. Size, foraging, and food web structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11):4191–4196, 2008.

[207] Owen L. Petchey, Amy L. Downing, Gary G. Mittelbach, Lennart Persson, Christopher F. Steiner, Philip H. Warren, and Guy Woodward. Species loss and the structure and functioning of multitrophic aquatic systems. *Oikos*, 104(3):467–478, 2004.

[208] Evelyn C Pielou. Species-diversity and pattern-diversity in the study of ecological succession. *Journal of Theoretical Biology*, 10(2):370–383, 1966.

[209] Stuart L Pimm. Food web design and the effect of species deletion. *Oikos*, 35(2):139–149, 1980.

[210] Michael J O Pocock, Darren M Evans, and Jane Memmott. The robustness and restoration of a network of ecological networks. *Science*, 335(6071):973–7, feb 2012.

[211] OA Popova and LA Sytina. Food and Feeding Relations of Eurasian Perch (Perca fluviatilis) and Pikeperch (Stizostedion lucioperca) in Various Waters of the USSR. *Journal of the Fisheries Research Board of Canada*, 34:1559–1570, 1977.

[212] Nicolas Poulet, L. Beaulaton, and S. Dembski. Time trends in fish populations in metropolitan France: Insights from national monitoring data. *Journal of Fish Biology*, 79(6):1436–1452, 2011.

[213] M. E. Power. Top-down and bottom-up forces in food webs: do plants have primacy?, 1992.

[214] C.J Puccia and Richard Levins. *Qualitative Modeling of Complex Systems: An Introduction to Loop Analysis and Time Averaging.* Harvard University Press, Cambridge, 1986.

[215] Peng Qiu and Sylvia K. Plevritis. Reconstructing directed signed gene regulatory network from microarray data. *IEEE Transactions on Biomedical Engineering*, 58(12):3518–3521, 2011.

[216] Robert J Radke and Reiner Eckmann. Piscivorous eels in Lake Constance: can they influence year class strength of perch? *Annales Zoologici Fennici*, 33(3):489–494, 1996.

[217] Haseeb Sajjad Randhawa and Robert Poulin. Determinants and consequences of interspeci W c body size variation in tetraphyllidean tapeworms. *Oecologia*, 161:759–769, 2009.

[218] G Rasmussen and B Therkildsen. Food, growth, and production of Anguilla anguilla L. in a small Danish stream. *Rapports et Proces-verbaux des Reunions Conseil International pour L'Exploration de la Mar*, 174:32–40, 1979.

[219] Enrico L Rezende, Jessica E Lavabre, Paulo R Guimarães, Pedro Jordano, and Jordi Bascompte. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature*, 448(7156):925–8, aug 2007.

[220] Sonia Romero-Romero, Axayacatl Molina-Ramírez, Juan Höfer, and José Luis Acuña. Body size- based trophic structure of a deep marine ecosystem. *Ecology*, 97(1):171–181, 2016.

[221] Elizabeth L Sander, J Timothy Wootton, and Stefano Allesina. What Can Interaction Webs Tell Us About Species Roles? *PLoS Computational Biology*, 11(7):1–22, 2015.

[222] Dirk Sanders and F. J. Frank Van Veen. The impact of an ant-aphid mutualism on the functional composition of the secondary parasitoid community. *Ecological Entomology*, 35(6):704–710, dec 2010.

[223] Oswald J. Schmitz, Peter a. Hambäck, and Andrew P Beckerman. Trophic Cascades in Terrestrial Systems: A Review of the Effects of Carnivore Removals on Plants. *The American Naturalist*, 155(2):141–153, 2000.

[224] Keith G Seaburg and John B Moyle. Feeding Habits, Digestive Rates, and Growth of Some Minnesota Warmwater Fishes. *Transactions of the American Fisheries Society*, 93(3):269–285, 1964.

[225] Jonathan B Shurin, Elizabeth T Borer, Eric W Seabloom, Carol A Blanchette, Bernardo Broitman, Scott D Cooper, and Benjamin S Halpern. A cross-ecosystem comparison of the strength of trophic cascades. *Ecology Letters*, 5:785–791, 2002.

[226] Jonathan B Shurin, Daniel S Gruner, and Helmut Hillebrand. All wet or dried up? Real differences between aquatic and terrestrial food webs. *Proceedings of the Royal Society of London B*, 273(1582):1–9, jan 2006.

[227] Phillip A Smith, Richard T Leah, and John W Eaton. Removal of pikeperch (Stizostedion lucioperca) from a British Canal as a management technique to reduce impact on prey fish populations. *Annales Zoologici Fennici*, 33(3):537–545, 1996.

[228] V Anne Smith, Jing Yu, Tom V. Smulders, Alexander J. Hartemink, and Erich D. Jarvis. Computational inference of neural information flow networks. *PLoS Computational Biology*, 2(11):1436–1449, 2006.

[229] WJP Smyly. The Life-history of the bullhead or miller's thumb (Cottus gobio L.). *Proceedings of the Zoological Society of London*, 128(3):431–454, 1957.

[230] Tom A. B. Snijders and Nowicki Krzysztof. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14:75–100, 1997.

[231] R V Solé and José M Montoya. Complexity and fragility in ecological networks. *Proceedings. Biological sciences / The Royal Society*, 268(1480):2039–45, 2001.

[232] András Specziár. First year ontogenetic diet patterns in two coexisting Sander species, S. lucioperca and S. volgensis in Lake Balaton. *Hydrobiologia*, 549(1):115–130, 2005.

[233] András Specziár. Size-dependent prey selection in piscivorous pikeperch Sander lucioperca and Volga pikeperch Sander volgensis shaped by bimodal prey size distribution. *Journal of Fish Biology*, 79(7):1895–1917, 2011.

[234] Phillip Staniczenko, Matthew J Smith, and Stefano Allesina. Selecting food web models using normalized maximum likelihood. *Methods in Ecology and Evolution*, 5(6):551–562, 2014.

[235] Daniel B Stouffer and Jordi Bascompte. Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9):3648–52, mar 2011.

[236] Daniel B Stouffer, Marta Sales-Pardo, M Irmak Sirer, and Jordi Bascompte. Evolutionary conservation of species' roles in food webs. *Science*, 335(6075):1489–1492, 2012.

[237] Sharon Y Strauss. Floral Characters Link Herbivores, Pollinators, and Plant Fitness. *Ecology*, 78(6):1640–1645, 1997.

[238] George Sugihara, Robert M May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, 338(6106):496–500, oct 2012.

[239] Yasuhiro Takeuchi. *Global dynamical properties of Lotka-Volterra systems*. World Scientific, London, 1996.

[240] Si Tang and Stefano Allesina. Reactivity and stability of large ecosystems. *Frontiers in Ecology and Evolution*, 2:1–8, 2014.

[241] L. R. Taylor. Aggregation, Variance and the Mean. *Nature*, 189(4766):732–735, 1961.

[242] R Core Team. R: A Language and Environment for Statistical Computing, 2016.

[243] D. W. Thieltges, K. Reise, K. N. Mouritsen, J. P. McLaughlin, and R. Poulin. Food web including metazoan parasites for a tidal basin in Germany and Denmark. *Ecology*, 92(10):2005–2005, 2011.

[244] David W. Thieltges, Karsten Reise, Kim N Mouritsen, John P Mclaughlin, and Robert Poulin. Food web including metazoan parasites for a tidal basin in Germany and Denmark. *Ecology*, 92:2005, 2011.

[245] Author J D Thomas. The Food and Growth of Brown Trout (Salmo trutta L .) and its Feeding Relationships with the Salmon Parr (Salmo salar L .) and the Eel (Anguilla anguilla (L .)) in the River Teify , West Wales. *Journal of Animal Ecology*, 31(2):175–205, 1962.

[246] Ross M Thompson, Kim N Mouritsen, and Robert Poulin. Importance of parasites and their life cycle characteristics in determining the structure of a large marine food web. *Journal of Animal Ecology*, 74:77–85, 2005.

[247] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

[248] G David Tilman, David Wedin, and Johannes Knops. Productivity and sustainability influenced by biodiversity in grassland ecosystems, 1996.

[249] H. Turesson, Anders Persson, and C. Brönmark. Prey size selection in piscivorous pikeperch (Stizostedion lucioperca) includes active prey choice. *Ecology of Freshwater Fish*, 11:223–233, 2002.

[250] B. Ünver and F. Erk'akan. Diet composition of chub, Squalius cephalus (Teleostei: Cyprinidae), in Lake Tödürge, Sivas, Turkey. *Journal of Applied Ichthyology*, 27(6):1350–1355, 2011.

[251] WLT van Densen. Piscivory and the development of bimodality in the size distribution of O+ pikeperch (Stizostedion lucioperca L.). *Journal of Applied Ichthyology*, 3:119–131, 1985.

[252] Frank J F van Veen, Rebecca J Morris, and H Charles J Godfray. Apparent competition, quantitative food webs, and the structure of phytophagous insect communities. *Annual review of entomology*, 51(107):187–208, jan 2006.

[253] Vera Vasas and Ferenc Jordán. Topological keystone species in ecological interaction networks: Considering link quality and non-trophic effects. *Ecological Modelling*, 196(3-4):365–378, 2006.

[254] O L E R Vetaas and John-arvid Grytnes. Distribution of vascular plant species richness and endemic richness along the Himalayan elevation gradient in Nepal. *Global Ecology & Biogeography*, 11:291–301, 2002.

[255] L A Vøllestad, J Skurdal, and T Qvenild. Habitat use, growth and feeding of pike (Esox lucius L.) in four Norwegian lakes. *Archiv fur Hydrobiologie*, 108:107–117, 1986.

[256] E Weiher and P A Keddy, editors. *Ecological Assembly Rules: Perspectives, Advances, Retreats*. Cambridge University Press, Cambridge, England, 1999.

[257] Hadley Wickham. Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[258] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[259] Hadley Wickham. testthat: Get Started with Testing. *The R Journal*, 3(1):5–10, 2011.

[260] Hadley Wickham. stringr: Simple, Consistent Wrappers for Common String Operations, 2017.

[261] Hadley Wickham. tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions, 2017.

[262] Hadley Wickham and Romain Francois. dplyr: A Grammar of Data Manipulation, 2016.

[263] John J Wiens, David D Ackerly, Andrew P Allen, Brian L Anacker, Lauren B Buckley, Howard V Cornell, Ellen I Damschen, T Jonathan Davies, John-Arvid Grytnes, Susan P Harrison, Bradford a Hawkins, Robert D Holt, Christy M McCain, and Patrick R Stephens. Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, 13(10):1310–24, oct 2010.

[264] Richard J Williams and Neo D Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–3, mar 2000.

[265] Richard J Williams and Neo D Martinez. Limits to trophic levels and omnivory in complex food webs: theory and data. *The American naturalist*, 163(3):458–68, mar 2004.

[266] J Timothy Wootton. Indirect Effects, Prey Susceptibility, and Habitat Selection: Impacts of Birds on Limpets and Algae. *Ecology*, 73(3):981–991, 1992.

[267] J. Timothy Wootton. Indirect Effects and Habitat Use in an Intertidal Community: Interaction Chains and Interaction Modifications. *Ecology*, 141(1):71–89, 1993.

[268] J Timothy Wootton. Predicting Direct and Indirect Effects: An Integrated Approach Using Experiments and Path Analysis. *Ecology*, 75(751):151–165, 1994.

[269] J Timothy Wootton. Estimates and Tests of per capita Interaction Strength: Diet, Abundance, and Impact of Intertidally Foraging Birds. *Ecological Monographs*, 67(1):45–64, 1997.

[270] J Timothy Wootton. Local interactions predict large-scale pattern in empirically derived cellular automata. *Nature*, 413(6858):841–4, oct 2001.

[271] J Timothy Wootton. Mechanisms of successional dynamics : Consumers and the rise and fall of species dominance. *Ecological Research*, 17:249–260, 2002.

[272] J Timothy Wootton. Experimental species removal alters ecological dynamics in a natural ecosystem. *Ecology*, 91:42–8, 2010.

[273] J Timothy Wootton. A 20-year data set of species replacement patterns in the middle-intertidal zone of Tatoosh Island, Washington, USA. *Ecology*, 97(3):810–810, 2016.

[274] J Timothy Wootton and James D Forester. Complex population dynamics in mussels arising from density-linked stochasticity. *PloS one*, 8(9):e75700, jan 2013.

[275] Jing Yu, V Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.

[276] C. D. Zander, N. Josten, K. C. Detloff, R. Poulin, J. P. McLaughlin, and D. W. Thieltges. Food web including metazoan parasites for a brackish shallow water ecosystem in Germany and Denmark. *Ecology*, 92(10):2007–2007, 2011.

[277] C. Dieter Zander, Neri Josten, Kim C. Detloff, Robert Poulin, John P McLaughlin, and David W. Thieltges. Food web including metazoan parasites for a brackish shallow water ecosystem in Germany and Denmark. *Ecology*, 92:2007–2007, 2011.