PHAGE–BACTERIA INFECTION NETWORKS: FROM NESTEDNESS TO MODULARITY AND BACK AGAIN

A Thesis
Presented to
The Academic Faculty

by

César Omar Flores García

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Physics

Georgia Institute of Technology
December 2014

PHAGE–BACTERIA INFECTION NETWORKS: FROM NESTEDNESS TO MODULARITY AND BACK AGAIN

Approved by:

Professor Joshua S. Weitz, Advisor
School of Physics
Georgia Institute of Technology

Professor James C. Gumbart
School of Physics
Georgia Institute of Technology

Doctor Sergi Valverde
Complex Systems Lab and Institute of Evolutionary Biology
Universitat Pompeu Fabra

Professor Flavio H. Fenton
School of Physics
Georgia Institute of Technology

Professor Kurt Wiesenfeld
School of Physics
Georgia Institute of Technology

Date Approved: 15 August 2014

To all my family, who have always supported me,

and the memory of my father.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Joshua S. Weitz, for his patience and support. The stimulating atmosphere at his lab made the ideas possible that eventually culminated in this thesis. Many thanks to the current and former group members who gave me feedback and comments during all my research: Dr. Richard I. Joh, Dr. Gabriel J. Mitchell, Luis F. Jover, Bradford P. Taylor, Abhiram Das, and Charles H. Wigington. I would like to thank Lauren Farr, who did an splendid job in collecting and digitizing the data in which this work is based. My thanks also goes to Dr. Michael H. Cortez, Dr. Lauren M. Childs, and Dr. Olga Symonova for offering me guidance at certain stages of my PhD studies. Very special thanks goes to Dr. Alexander Bucksh, who I considered my second PhD mentor.

I am greatly indebted with Dr. Sergi Valverde who was before a collaborator, a mentor who share his very broad knowledge of complex networks to me. I am also thankful to Dr. Justin R. Meyer, who besides being a collaborator, give me the opportunity of learning some of the basis in Experimental Microbiology when he was a student of Richard Lenski, who I am also thankful for the opportunity.

My thanks also goes to Dr. Drew Purves and all the people of the Computational Ecology and Environmental Science at Microsoft Research Cambridge for giving me the big opportunity of doing an internship there. I had an unique experience of interacting with some of the smartest scientists in my field. My internship at Microsoft helped me to open and formulate new questions besides my main PhD research topic.

Finally, I would like to thank my family and friends for their encouragement and support. Special thanks goes to my mother and brothers that have been always there, when I need them the most.

My research was supported by CONACyT fellowship for graduate studies and grants from the Burroughs Wellcome Fund and the National Science Foundation Division of Biological Oceanography awarded to Joshua S. Weitz.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Bacteriophages (viruses that infect bacteria) are the most abundant biological life-forms on Earth. However, very little is known regarding the structure of phage-bacteria infections. In a recent study we showed that phage-bacteria infection assay datasets are statistically nested in small scale communities while modularity is not statistically present [60]. We predicted that at large macroevolutionary scales, phage-bacteria infection assay datasets should be typified by a modular structure, even if there is nested structure at smaller scales. We evaluate and confirm this hypothesis using the largest study of the kind to date [62].

The study in question represents a phage-bacteria infection assay dataset in the Atlantic Ocean region between the European continental shelf and the Sargasso Sea. We present here a digitized version of this study that consist of a bipartite network with 286 bacteria and 215 phages including 1332 positive interactions, together with an exhaustive structural analysis of this network. We evaluated the modularity and nestedness of the network and its communities using a variety of algorithms including BRIM (Bipartite, Recursively Induced Modules), NTC (Nestedness Temperature Calculator) and NODF (Nestedness Metric based on Overlap and Decreasing Filling). We also developed extensions of these standard methods to identify multi-scale structure in large phage-bacteria interaction datasets. In addition, we performed an analysis of the degree of geographical diversity and specialization among all the hosts and phages.

We find that the largest-scale ocean dataset study [124] , as anticipated by Flores et al. [60], is highly modular and not significantly nested (computed in comparison to null models). More importantly is the fact that some of the communities extracted

from Moebus and Nattkemper dataset were found to be nested. We examine the role of geography in driving these modular patterns and find evidence that phage-bacteria interactions can exhibit strong similarity despite large distances between sites. We discuss how models can help determine how coevolutionary dynamics between strains, within a site and across sites, drives the emergence of nested, modular and other complex phage-bacteria interaction networks.

Finally, we releases a computational library (BiMAT)to help to help the ecology research community to perform bipartite network analysis of the same nature I did during my PhD.

CHAPTER I

INTRODUCTION

1.1 Context

1.1.1 A little bit of history in complex networks

The foundation of *complex networks* research started with the famous Euler's proof that no solution exist for the *Königsberg bridge* problem, which consisted in finding a way of crossing seven bridges once and only once across different points of the city of Königsberg. The genius of Euler's proof, performed in 1735 was to realize that the only relevant information to solve this problem was the list of land points (vertices or nodes) and the bridges interconnecting them (edges or links). Hence, he introduced the concept of a graph, thus giving birth to the very fruitful mathematical field of graph theory. This graph concept can be mathematically described as $G = \{V, E\}$, with V and E representing nodes and edges respectively. Because any set of items with interactions between each other can be represented as a graph, a broad set of important applications have been found for graph theory (*i.e.*, route problems, search problems, map coloring, and many others).

More than two hundred years later, Erdös and Rényi introduced the first *random* graph model (better known as Erdös–Rényi model) in 1959 [56], which is a very important concept for the *complex networks* research field. One way to describe it mathematically is by $G(n, p)$, where n represents the number of nodes and p the probability of having an edge between each pair of nodes. As we will see next, the importance of this concept is that many real–world networks (or graphs) can be better understood as deviation from *random* graphs.

A revival of interest in the study of networks was made possible by increasing

availability of network data, as well as algorithms (and computing power) to analyze them. Initiated by two very famous papers [184] and [12], real–world networks study has exploded in recent years, giving rise to the birth of *complex networks* research. Some examples of *complex networks* are social networks of acquaintances, the WWW, the Internet, food webs, financial networks, neural networks, metabolic networks and many others. What these networks have in common is that they are not well described by the Erdös–Rényi model. In other words, having features that deviate from *random*, they have make scientists wonder why they exist. Some of these features are high clustering [184], power-law degree distribution [12], community structure [128], motifs (sub-graphs that appear more or less than randomly expected) [121], to name some of the most well studied. Understanding what are the causes for real–world networks to deviate from *random* has been the main focus of study of the *complex networks* field.

1.1.2 Complex networks in ecology

Many ecological systems can be represented as a graph, where nodes are populations of living organisms and edges are the interactions among them. Hence, it is not surprising that the *complex networks* research field has made many important contributions to ecology during the last years. One of the most important examples in ecology are food webs, which are networks composed of species (nodes) and *who-eats-whom?* relationships (edges). Some of the contributions to food web research are studies about how degree distribution in food webs compare to other networks[53], network models to explain its network structure [38, 39, 191, 192, 5], and sampling strategies to improve the quality in the data [111].

Another important ecological relationship, in which this thesis work has been based, is plant–pollinator networks [17, 130, 18, 20]. These networks have the ecological property that both plants and pollinators benefit each other with regard to

survival. Hence, they are called mutualistic networks, which can be represented as a *bipartite* network (see below). Researchers have shown that these networks have specific features that distinguish them from *random networks*. Specifically, they are nested networks [18] (see Figure 1).

As with any other real–world network, understanding the structure of ecological networks have profound implications. For instance, it can help conservation policy makers predict which species are most likely to go extinct [84].

1.2 Phage–bacteria interactions: Who-kills-whom?

Bacteria are among the most abundant organisms on Earth, with estimates of around 10^{30} individuals. A very important predator of bacteria are bacteriophages (virus that infect and kill bacteria). As is the case for plant–pollinator networks, this relationship can be represented as a *bipartite complex network*. A *bipartite* network is a network in which nodes can be grouped in two different subsets such that edges can happen only between nodes across subsets. Mathematically, it can be defined as $G = (U, V, E)$, where U and V representing the two different subsets of nodes (in this case bacteria and phages), and E the edges across them (in this case which phage can infect which bacteria). However, a big difference exist between this type of networks and plant–pollinator networks. While the last one is a mutualistic relationship (plant and pollinators benefit one another), the phage-bacteria relationship is often an antagonistic one (phages need bacteria to survive, but bacteria do not need phages, and in fact they often do better without phages).

Despite the importance of phage–bacteria communities very little is know about the structure of the *who-kills-whom?* interaction network. To prove that these networks have features that distinguish them from *random* can give us insight about what are the ecological and biological mechanisms that lead to the structure in this type of ecological networks. To find the features that distinguish this kind of networks

from *random* networks is the main topic of this thesis. These features have a very important implication in microbiology and ecology in general. For instance it can help us to understand how the co–evolution race between these two species happens (*i.e.*, does phage evolution follow bacterial evolution, the opposite, or a mixing of both). It can give us insight about how coexistence mechanisms exist between these two species. However, to study those mechanisms is far beyond the scope of this thesis, which concentrates on the quantitative characterization of phage-bacteria infection networks.

1.2.1 Possible scenarios

In order to describe and find features in phage–bacteria networks, we started from the hypothesis that *bipartite* networks can be described in reference to four general *bipartite* network patterns, which are described in Figure 1. More details about these patterns will be given in the body of this thesis.

1.3 And then, what are the phage–bacteria network features?

It is not possible to describe the features of a network without having appropriate data sets. In order to solve this problem, the Weitz group performed a literature search for phage–bacteria cross infection networks looking at papers that date as far back as 1950 to current years (2011)[1]. In order for a cross–infection study to be considered an appropriate data set we required that at least two species of both phage and bacteria to exist in the study, and that no *NA* (no value available) interactions to be present. Further, quantitative infections were considered Boolean. Therefore all the analyzed networks were treated as Boolean (with 1/0 indicating interaction/no interaction respectively). Altogether 38+1 matrices were found and analyzed with the

[1]I want to thank Lauren Farr (a Biology undergrad student at the time), who did a splendid job in collecting the data.

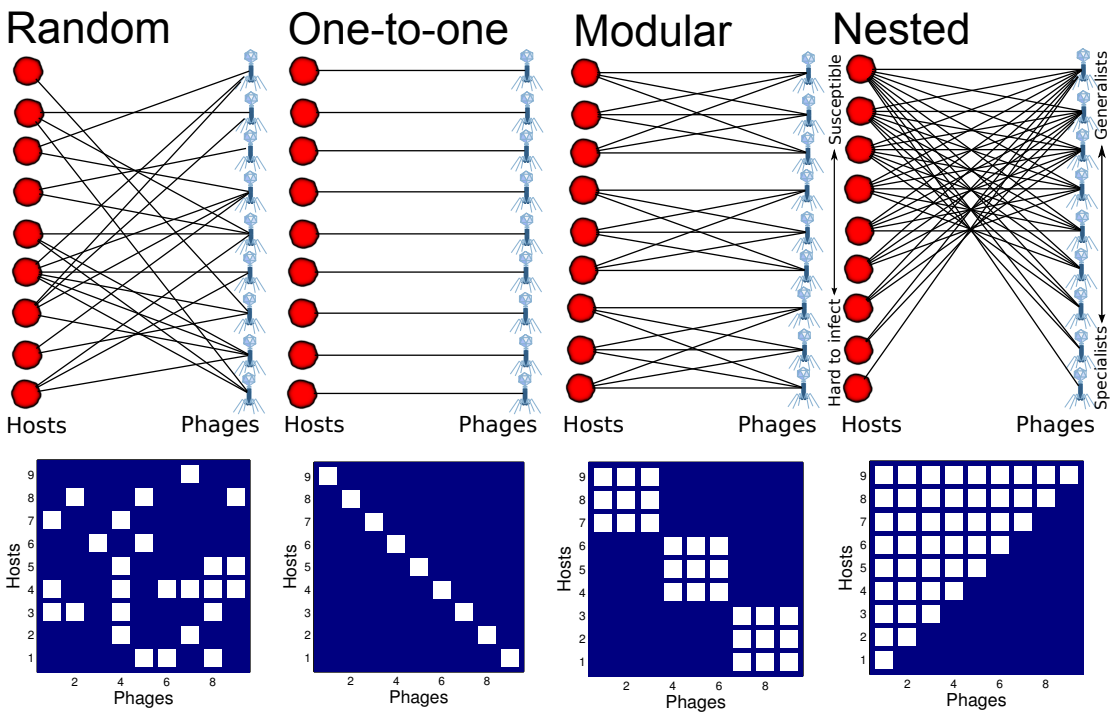

Figure 1: Potential patterns that could exist in phage–bacteria infection networks. **Random**: the pattern of who infects whom is not statistically different than what would be expected if interactions occurred by chance. **One-to-one**: an infection network with elevated specialization, such that each phage can only infect one host, and each host is only infected by one phage. **Perfectly Modular**: interactions happen only between predefined sets of bacteria and phages with no interactions across these sets. **Perfectly Nested**: Bacteria can be ranked in increasing order of infectivity from bacteria that can be infected by a single phage (hard to infect) to bacteria that can be infected by all phage species (susceptible). Similarly, phages can be ranked in terms of the number of bacteria they can infect (increasing from specialists to generalists).

last one being the largest cross–infection study to date (to my knowledge). Hence, this data–collection is by far the largest compilation of phage–bacteria cross–infection studies.

Visual inspection immediately discarded the **one–to–one** type of network (*i.e.*, most species in the analyzed networks interact with more than one specie). Hence, we focus in distinguishing **nested** and **modular** patterns from *random* ones. For accomplishing this goal, we used a series of algorithms belonging to the *complex networks* literature to evaluate nestedness and modularity values, and then compare

them to the ones evaluated in **random** networks.

1.3.1 From Nestedness ...

We first analyzed the first 38 networks, which are by nature small in size compared to the last matrix. We used the simplest null model, which is basically the generalization of the Erdös–Rényi model [56] to *bipartite* networks with number of nodes and probability of interaction the same as the number of nodes and connectance of the network that is being tested. After performing appropriate statistical tests (to be described in Chapter 2, which is adapted from Flores et al. [60]), we found that these networks are nested (27 of 38), which is a very similar result to what is found in plant–pollinator networks [17, 130, 18, 20]. The fact that mutualistic and antagonistic networks can be explained as a deviation of the same pattern is an intriguing discovery because the their underlying mechanisms would seem to be totally different.

1.3.2 ... To Modularity ...

One important characteristic of the previous 38 studies is that they have a small number of tested strains, and that in most cases they belong to the same species (*i.e.*, *E. coli* vs. λ–phage strains). Hence, the genetic distance between them is short. Our next step in the study of this kind of networks was to understand what will happen if the network increased by including strains of different species and geographical locations. To answer this question we used the largest data set to date, which is the K. Moebus and H. Nattkemper [124] study (details on Chapter 3, which is adapted from Flores et al. [62]). We find that in fact the network become **modular** and the **nestedness** is lost. Hence, we demonstrated that structure depends on size of the network.

1.3.3 . . . And back again

Finally, we used the same study to look at local parts of the network. Specifically, we look at the identified modules of the community structure algorithms, and found similar results that we found for the case of the 38 matrices. That is, we found that the structure at this smaller scale is also nested. This result is very significant because it tell us that the structure of these networks will depend on the scale at which we look at them.

1.4 *BiMAT: a software for performing bipartite network analysis*

Data analysis, in general, involves (i) getting data, (ii) asking insightful questions, and (iii) visualizing the answers. My final contribution to the ecology research community was the release of `BiMAT`. While nothing can be done about (i) without a direct collaboration with the community, `BiMAT`'s main goal is to help researchers to attack (ii) and (iii) for the case of *bipartite* ecological network data.

This library comes after three years of thinking about what are the best algorithms and relevant questions of *bipartite* ecological networks that can be answered. In a sense, we have to take a lot of decisions about which kind of analysis give us insightful results and which do not. Hence, the library allows the users to perform exhaustive initial analysis of their data without having them invest time in deciding what are the appropriate analyses and algorithms that can be used (decisions that take a lot of time when the users are inexperienced). This library will be introduced in Chapter 4, which is adapted from Flores et al. [61].

CHAPTER II

STATISTICAL STRUCTURE OF HOST–PHAGE INTERACTIONS

Adapted from Cesar O. Flores, Justin R. Meyer, Sergi Valverde, Lauren Farr, and Joshua S. Weitz. Statistical structure of host–phage interactions. PNAS (2011) [60].

Interactions between bacteria and the viruses that infect them (i.e., phages) have profound effects on biological processes, but despite their importance, little is known on the general structure of infection and resistance between most phages and bacteria. For example, are bacteria–phage communities characterized by complex patterns of overlapping exploitation networks, do they conform to a more ordered general pattern across all communities, or are they idiosyncratic and hard to predict from one ecosystem to the next? To answer these questions, we collect and present a detailed metaanalysis of 38 laboratory–verified studies of host–phage interactions representing almost 12,000 distinct experimental infection assays across a broad spectrum of taxa, habitat, and mode of selection. In so doing, we present evidence that currently available host–phage infection networks are statistically different from random networks and that they possess a characteristic nested structure. This nested structure is typified by the finding that hard to infect bacteria are infected by generalist phages (and not specialist phages) and that easy to infect bacteria are infected by generalist and specialist phages. Moreover, we find that currently available host–phage infection networks do not typically possess a modular structure. We explore possible underlying mechanisms and significance of the observed nested host–phage interaction structure. In addition, given that most of the available host–phage infection networks examined here are composed of taxa separated by short phylogenetic distances, we propose that

the lack of modularity is a scale–dependent effect, and then, we describe experimental studies to test whether modular patterns exist at macroevolutionary scales.

2.1 Introduction

Bacteria and their viruses (phages) make up two of the most abundant and genetically diverse groups of organisms [55, 175, 67]. The extent of this diversity has become increasingly apparent with the advent of community genomics. Microbial DNA isolated from oceans, lakes, soils, and human guts has revealed tremendous taxonomic diversity in a broad range of environmental habitats and conditions [180, 88, 10, 195, 43, 177, 73, 176]. The ongoing discovery of new taxonomic diversity has, thus far, outpaced gains in understanding the function of specific microbes and their most basic ecology of who interacts with whom. One of the starkest examples of this disparity is the lack of an efficient (bioinformatic or otherwise) approach for determining which viruses can infect which bacteria. Although it is well–known that individual phages do not infect all bacteria, we have little understanding of what the precise host range for any given phage is or whether there are universal patterns or principles governing the set of viruses able to infect a given bacterium and the set of bacteria that a given virus can infect. This deficit is unfortunate given that phage–bacterial interactions are important for both human health and ecosystem function [106, 145, 57, 58, 76].

Phages have multifaceted effects on their hosts: they can lyse host cells, thereby releasing new virons, transfer genes between hosts, and form lysogens that can modify host function [185, 2]. In some cases, phages can transfer genes for pathogenicity between pathogenic and labile strains (e.g., for both *Vibrio cholerae* and *Shigella*), facilitating the spread of bacterial infections [23, 155, 149]. Phages also alter ecosystem functions by the high levels of bacterial mortality that they cause. Bacteria lysed by phage will release their contents, which consequently are scavenged by other bacteria

rather than being incorporated into bactivorous eukaryotes [68, 74]. This weakened connection early in the food chain can have effects that ripple throughout the ecosystem. Information on a general pattern of infection by phages on hosts could improve predictions of microbial population dynamics, ecosystem functioning, and microbial community assembly [22, 170].

What is our expectation for the general pattern of host–phage infection networks? Host–phage infection networks have, in the past, been measured by performing pair–wise infections of hosts by phages isolated from natural ecological communities, evolution experiments, or strain collections. The results of such pair–wise infections can be represented as a network or a matrix, where the rows indicate host isolates, the columns indicate phage isolates, and the cells within the matrix describe whether each combination results in a successful infection. We consider different classes of host–phage interaction networks as alternative hypotheses for an expected pattern (Figure 2). First, phages may infect a unique host or a limited number of closely related hosts, leading to nearly diagonal matrices (Figure 2A) or block–like matrices that exhibit high degrees of modularity (Figure 2B). These patterns should occur if host–viral interactions are the result of coevolutionary processes that lead to specialization. Second, diversification of hosts and phages may result in nested matrices in which the most specialist phages infect those hosts that are most susceptible to infection rather than infecting those hosts that are most resistant to infection (Figure 2C). The nested pattern is the predicted outcome of a prominent theory of gene–for–gene coevolution, where phages evolve so as to broaden host ranges and bacteria evolve so as to increase the number of phages to which they are resistant [139, 4]. We should note that these two patterns and hypotheses for the forms of coevolution are not mutually exclusive and in fact, could be scale–dependent. Nested patterns could form within modules if, for instance, microevolutionary changes result in nestedness; however, genetic differences between species or genera that accumulate over

macroevolutionary time may limit the exchange of viruses between these phylogenetic groups and create an overall modular structure. Finally, we consider a null model to be that matrices of host–phage infection are statistically indistinguishable from random matrices (Figure 2D).

Contrary to this null expectation, we show that currently available host–phage interaction matrices are, as a whole, statistically distinguishable from random matrices and possess a characteristic nested structure. We reach this conclusion by performing a metaanalysis on the patterns of host–phage infection matrices collected by a comprehensive search of the literature and supplementing these matrices with an experimental analysis of host–phage infection. The data that we assemble consist of 38 matrices of host–phage infection assays representing the cumulative study of 1,009 bacterial isolates, 502 phage isolates, and almost 12,000 separate attempts to infect a bacteria host with a phage strain [1, 14, 24, 29, 31, 32, 33, 41, 42, 47, 49, 54, 70, 77, 83, 86, 97, 101, 102, 104, 114, 118, 120, 122, 131, 133, 139, 140, 147, 152, 161, 163, 164, 168, 182, 189, 197] (See Appendix A, Tables 7 and 8 have more information on the examined studies). This work is an attempt to subject host–phage infection assays to a unified analysis. In doing so, we find a general pattern of host–phage interactions. We discuss biophysical, ecological, and evolutionary mechanisms that could lead to this nested (and not modular) pattern as well as future studies to explore how such a pattern may change as a function of phylogenetic scale.

2.2 Results

2.2.1 Compiling a Large–Scale Host–Phage Interaction Dataset

We compiled a set of 37 studies with direct laboratory evidence of host–phage interactions using an extensive literature search supplemented by an experimental study of an evolved *Escherichia coli* and phage λ–system (Appendix A, Tables 7 and 8 have complete details of all studies) [1, 14, 24, 29, 31, 32, 33, 41, 42, 47, 49, 54, 70, 77, 83, 86,

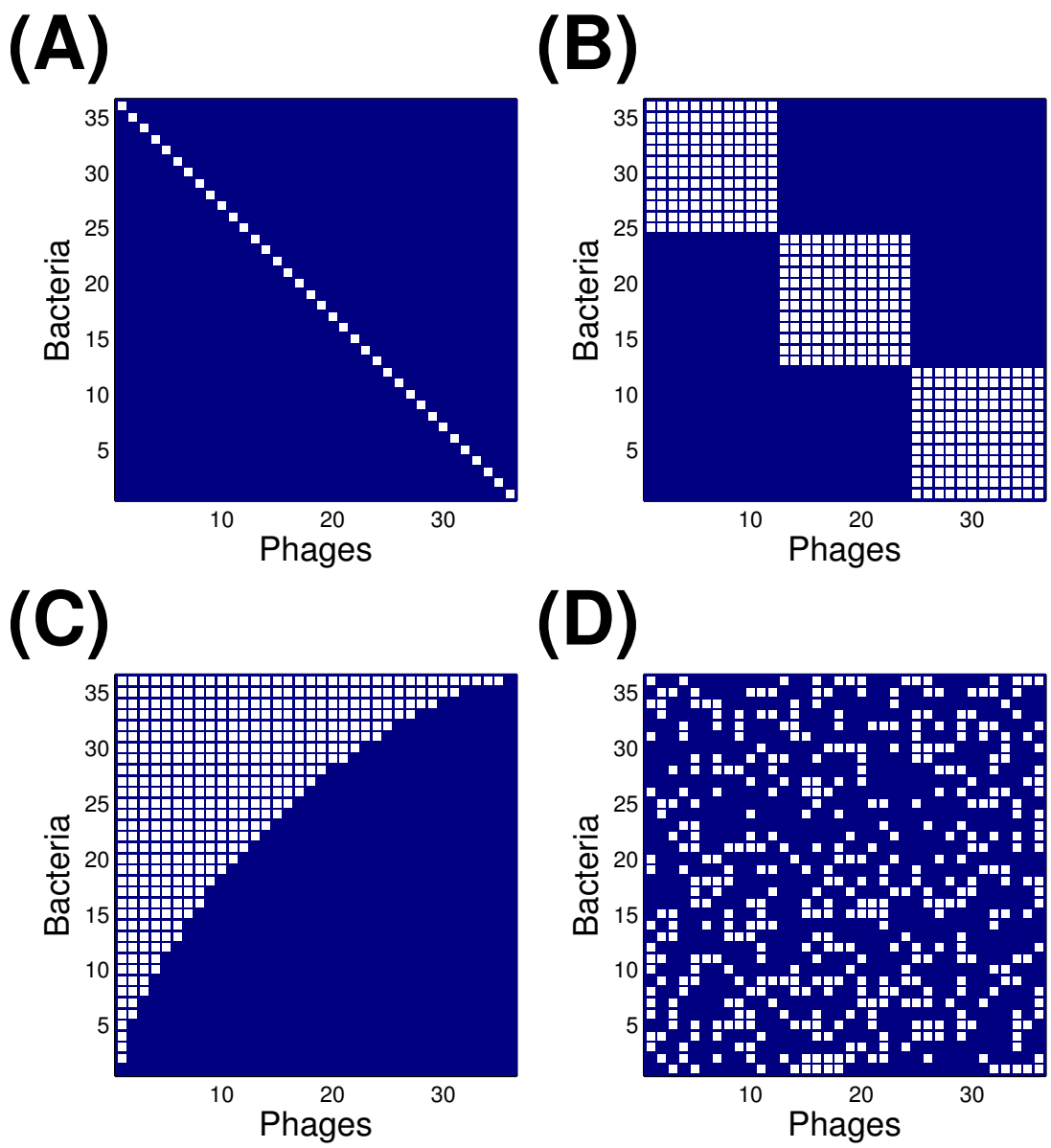

Figure 2: Schematic of expected host–phage interaction matrices (white cells denote infection). **(A)** Host–phage interactions are unique (i.e., only one phage infects a given host, and only one host is infected by a given phage). **(B)** Host–phage interactions are modular (i.e., blocks of phages can infect blocks of bacteria, but cross-block infections are not present). **(C)** Host–phage interactions are nested (i.e., the generalist phage infects the most sensitive and the most resistant bacteria, whereas the specialist phage infects the host that is infected by the most viruses). **(D)** Host–phage interactions are random and lack any particular structure. For **(B–D)**, a connectance of 0.33 was used so that the expected total number of interactions was the same in each case.

97, 101, 102, 104, 114, 118, 120, 122, 131, 133, 139, 140, 147, 152, 161, 163, 164, 168, 182, 189, 197]. The method of evaluating infection ability in assembling a host–phage infection matrix varies; however, the most commonly used approach is that of spot assays, in which a single virus type is combined with a population of bacteria cells from a single strain. Infection is considered to have occurred given evidence that the phage has infected and lysed (part of) the bacterial population. Hence, the result of each study is a matrix of the infection ability for each phage on each host. The studies included in the host–phage infection assays analyzed here were isolated from one of three sources: co-occurring isolates within natural communities taken directly from the environment and then cultured, coevolutionary laboratory experiments where a single bacterial clone and a single phage clone were allowed to coevolve for a fixed amount of time and then, their evolved progenitors examined, and laboratory stocks of phages and hosts that were artificially combined. Some of the matrices used were composed of bacteria and phage acquired from two separate isolation strategies. For these studies, we classified the matrix by which isolation strategy represented the majority of matrix cells and made a note of the other sources (Appendix A, Table 8). The criterion by which we searched and cataloged these studies is explained in more detail in Appendix A. Overall, we identified and analyzed a wide range of infection networks for organisms that varied in their phylogenic position, traits, and habitats. For example, the bacterial hosts included Gram–positives and –negatives, heterotrophs, and phototrophs as well as pathogens and nonpathogens.

Some of the assays include graded information about infection (for example, whether a phage simply inhibits bacterial growth or forms regions of complete bacterial mortality like plaques). In other studies, replicate phage populations were used to deduce whether phages always or only sometimes cause plaques. Details of the criteria for the interactions can be found in the original works [1, 14, 24, 29, 31, 32, 33, 41, 42, 47, 49, 54, 70, 77, 83, 86, 97, 101, 102, 104, 114, 118, 120, 122, 131, 133,

139, 140, 147, 152, 161, 163, 164, 168, 182, 189, 197], and the experimental methods for the experimental study of host–phage infection can be found in Materials and Methods. Because graded information about infection was not uniformly available in all studies, assays were standardized using hand–curated extraction of original data into a single matrix of ones and zeros with H rows (one for every bacterial host) and P columns (one for every phage), where a 1–valued cell represents evidence for infection (either full or partial) and a 0–valued cell represents no evidence for infection (Figure 3 shows a visual depiction of all host–phage interaction matrices).

2.2.2 Host–Phage Infection Statistics Do Not Vary with Study Type or Show Significant Cross-Correlations

We calculated a variety of global properties of these matrices: number of hosts (H), number of phages (P), number of interactions (I), number of species ($S = H+P$), size ($M = HP$), connectance ($C = I/M$), mean number of interactions across host species ($L_H = I/H$), and mean number of interactions across phage species ($L_P = I/P$) (Appendix A, Tables 7, 8, and 9 show values of each property within each of the 38 studies). Importantly, on a per-study basis, we find that the average number of phages infecting a given host is 4.88 (median = 3.04), whereas the average number of hosts that a phage can infect is 10.91 (median = 6.13). Both results are inconsistent with the hypothesis that phages only infect one host and that hosts are only infected by one phage (Figure 2A).

We first sought to establish whether the source type (natural communities taken directly from the environment and then cultured, coevolutionary laboratory experiments where a single bacterial clone and a single phage clone were allowed to coevolve for a fixed amount of time and then, their evolved progenitors examined, and laboratory stocks of phages and hosts that were artificially combined) had any influence on basic characteristics of the matrices. We performed a principal component analysis (Appendix A, Table 10, and Figure 20) using these eight global properties. Despite

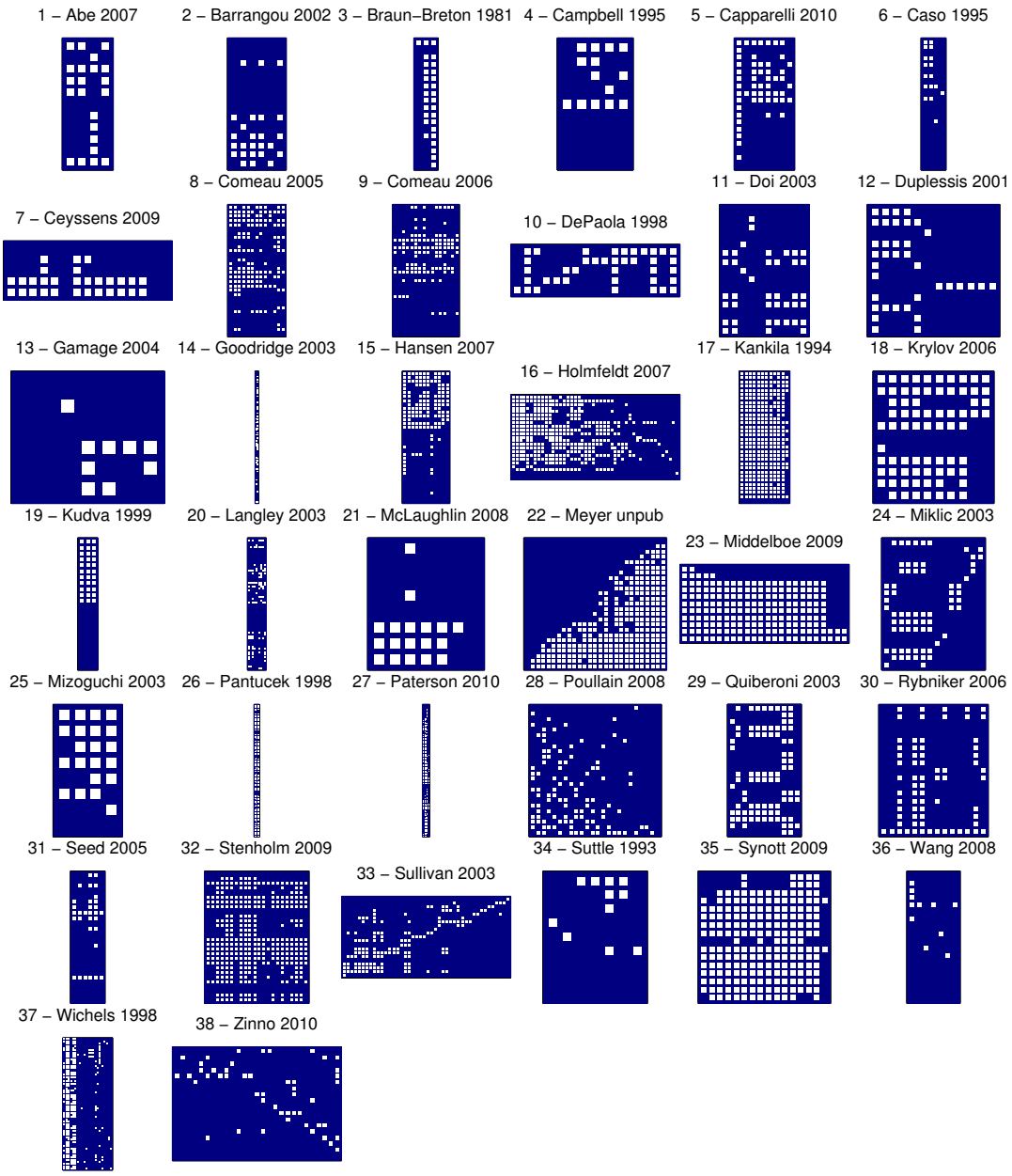

Figure 3: Matrix representation of the compiled studies. The rows represent the hosts, and the columns represent the phages. White cells indicate the recorded infections. Note the diversity in the size of these matrices.

the significant variation in global properties, we find no statistically significant distinction between the three different types of studies. For example, the distributions of type-specific matrices do not cluster into three groups. We apply a Jaccard clustering validity index [91] and find that the degree of clustering validity is 0.26 (indicating

poor separation of labeled classes into distinct clusters), which is not significantly different from random ($P = 0.33$) (Appendix A, Figs. 22 and 23).

Not only do we not find evidence for clustering, we also do not find evidence for significant and biologically meaningful correlations among the global properties of all matrices when grouped together. For example, previous work on the analysis of bipartite networks within plant and pollinator systems found inverse relationships between the total number of species in the network and the fraction of interactions that actually occurred [115, 18]. We do not find this relationship here. Appendix A, Figure 21 plots connectance (C) vs. number of species (S). The observed slope is small and nonsignificant (Appendix A, Table 11). Moreover, the other correlations between connectance and the size of host–phage infection matrices are not significant (Materials and Methods has details and Appendix A, Table 11 shows the correlation values).

2.2.3 Host–Phage Infection Assays Are Typically Nested and Not Modular

We measured higher–order properties of the host–phage interaction matrices, specifically modularity and nestedness. In this context, modularity is determined by the occurrence of groups of phages that infect groups of hosts significantly more often than they infect other hosts in the system. Modularity is typically found in biological systems in which groups of organisms preferentially interact with organisms within the group (e.g., plant–pollinator network) [115, 18] and is thought to be an important feature underlying the maintenance of biodiversity [173]. Likewise, nestedness is determined by the extent to which phages that infect the most hosts tend to infect bacteria that are infected by the fewest phages [179, 6]. Nestedness has been used to characterize species interactions because it is predicted to affect important properties of communities such as stability and extinction potential [18, 20]. Both modularity and nestedness may emerge because of coevolutionary adaptation of hosts

and phages [4, 150]. The individual host–phage infection studies collected here were not subjected to a network analysis with one exception [139]. Hence, we examined each study to see if previously unrealized patterns existed within each host–phage interaction network (Figure 4 and Appendix A, Figure 24 have an example of how network properties are extracted from two matrices, Datasets S1 and S2 shows data corresponding to each matrix, and Materials and Methods has additional details on how to calculate modularity and nestedness).

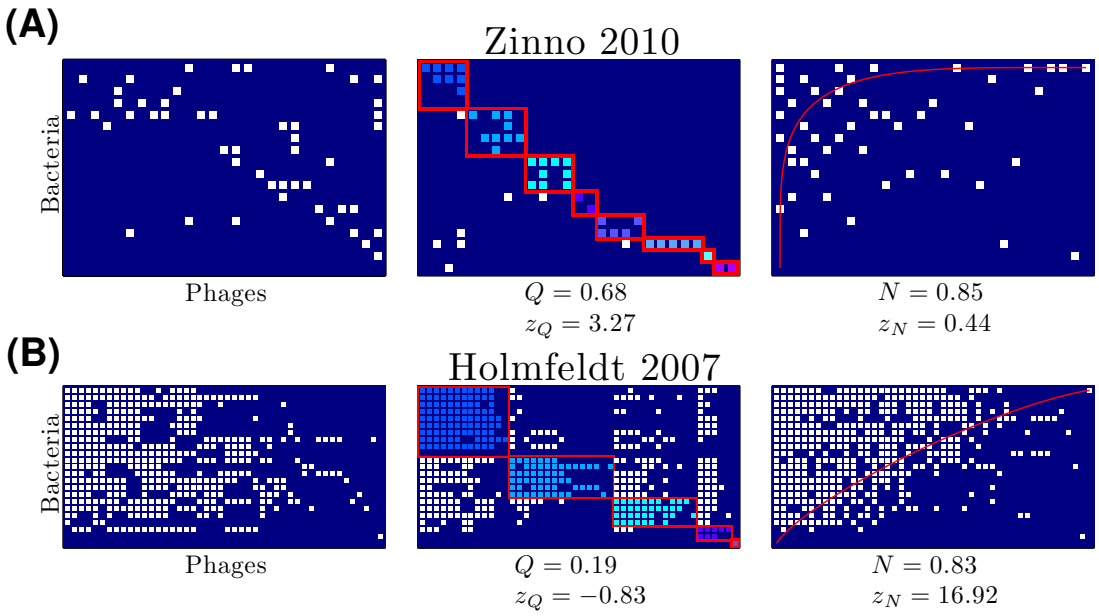

Figure 4: Two example matrices were resorted to maximize modularity and nestedness. ((**A**) and (**B**)) The matrix in Left is the original data, the matrix in Center is the output from the modularity algorithm [13], and the matrix in Right is the output from the modified nestedness algorithm [11, 143]. Colors represent different communities within the maximal modular configuration. **(A)** An example of a matrix with significantly elevated modularity and insignificant nestedness. **(B)** An example of a matrix with insignificant modularity and significantly elevated nestedness.

For the 38 matrices shown in Figure 3, the maximally modular relabeling of each matrix is displayed in Figure 5 and the maximally nested resorting of each matrix is displayed in Figure 6. To evaluate the statistical significance of the modularity and nestedness values of observed host–phage matrices, we have to compare the observed values to those values of random matrices. We generate random matrices that have

the same size and number of interactions as the original data (Appendix A, Materials and Methods). In that way, we constrain our null model to have exactly the same global properties as detailed in Appendix A, Table 7 for each study, whereas the nestedness and modularity will vary between realizations.

The titles of the study in Figure 5 (the maximally modular configuration) are red if they are significantly modular, blue if they are significantly antimodular, and black if they are nonsignificantly modular. The majority of studies are significantly antimodular (where we used a p–value = 0.05 and 10^5 random matrices as our null). Our findings stand in contrast to expectations that groups of phages adsorb to non–overlapping groups of hosts, which would be expected if groups of phages had specialized on groups of hosts within the study systems. The titles of each study in Figure 6 (the maximally nested configuration) are red if they are significantly nested, blue if they are significantly antinested, and black if they are nonsignificantly nested. The majority of studies are significantly nested ($p < 0.05$), where we used 10^5 random matrices as our null. Overall, we find 27 of 38 studies to be significantly nested, and when broken down by type, we find significant nestedness in 13 of 19 ecological, 7 of 10 experimental, and 7 of 9 artificial studies. Our findings corroborate, in one case, an earlier effort to characterize nestedness by Poullain et al. [139] using a different nestedness metric. It is also apparent that some matrices are almost perfectly nested [e.g., matrices in the works of Ceyssens et al. [33], McLaughlin and King [114], and Seed and Dennis [152]]. In some cases, like the work of Middelboe et al. [118], the data came from a mix of ecological and experimental studies in that the bacteria were derived from environmental and experimentally evolved isolates, whereas the phages were wild from the same environment as the host. Does the finding of a strongly nested matrix mean, in this case, that in vitro evolution mimics selection in nature, suggesting that there exists robust principles underlying the emergence of nestedness?

Hence, given the number of studies, we ask what evidence is there that host–phage

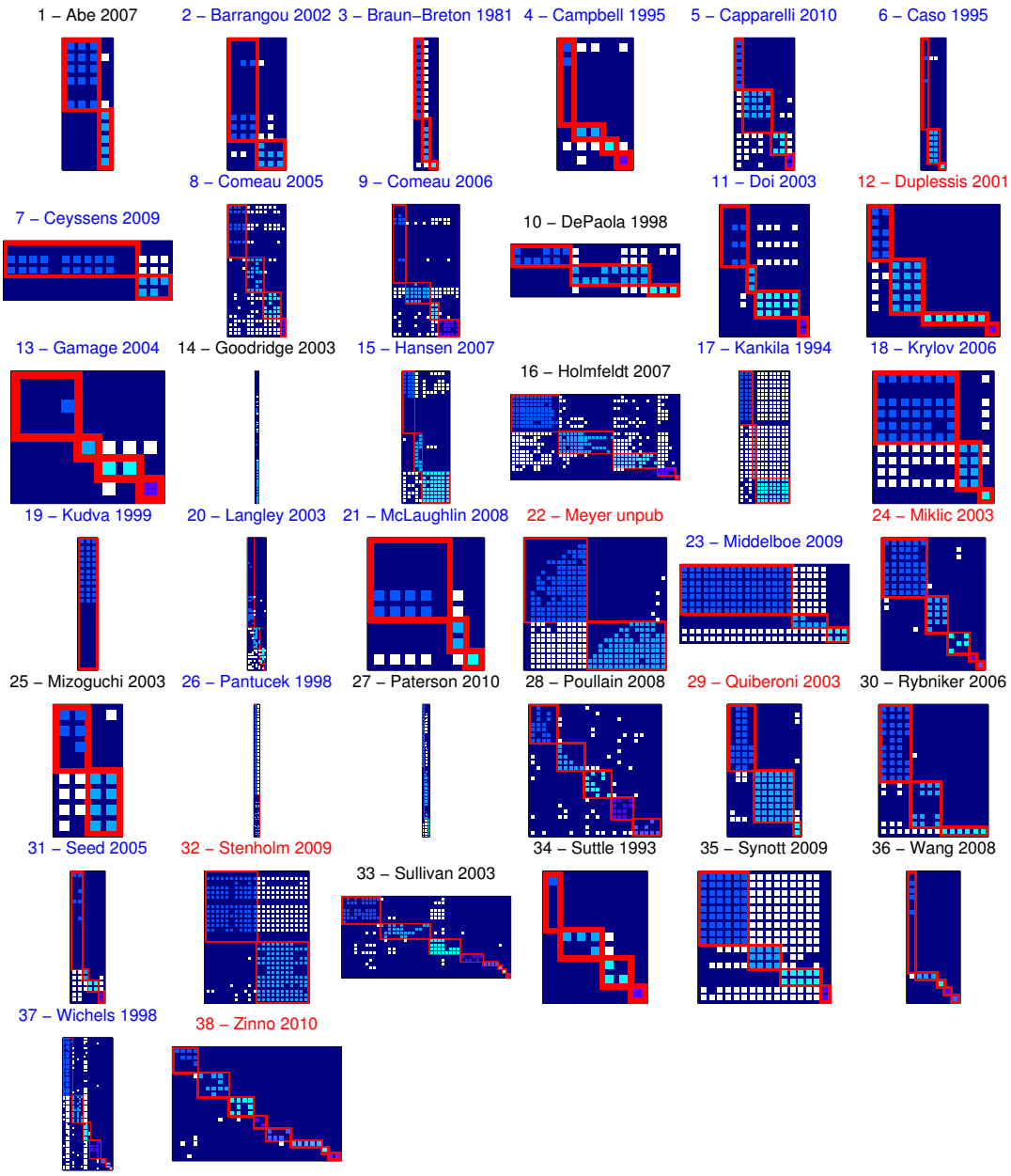

Figure 5: Modularity sorts of the collected studies. Blue labels (20/38) represent studies statistically antimodular, and red labels (6/38) represent studies statistically modular.

matrices are, as a whole, nested and not modular. We rank all 38 matrices from lowest to largest modularity and lowest to largest nestedness (Figure 7 A and B). It is evident that matrices tend to be more nested than their random counterparts but not more modular (and apparently, antimodular) than their random counterparts. How often

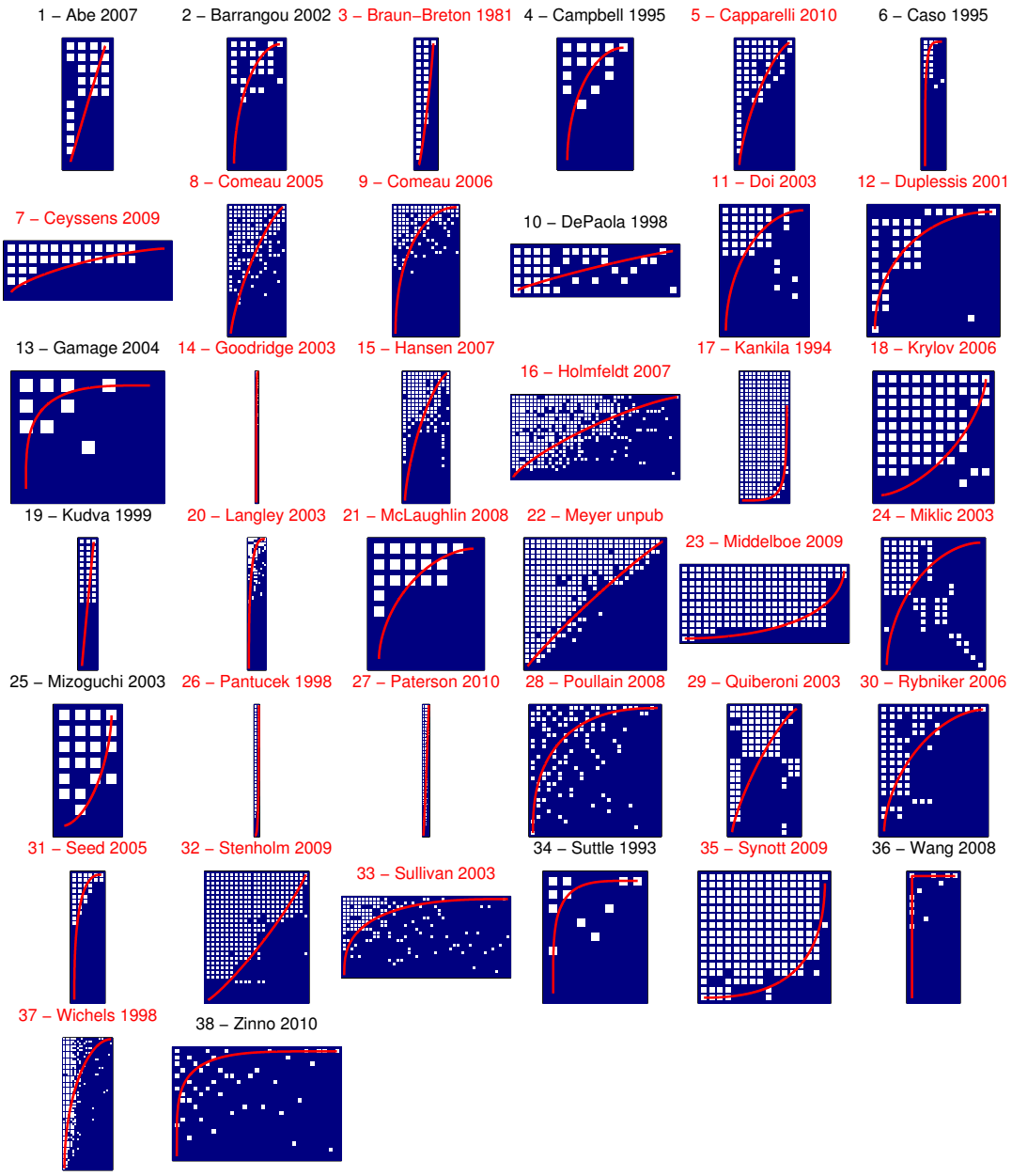

Figure 6: Nestedness sorts of the collected studies. Red line represents the isocline (see Equation 18 of Appendix A). Blue labels (0/38) represent studies statistically antinested, and red labels (27/ 38) represent studies statistically nested. See

do we expect to find 27 significantly nested matrices in a sample of 38 random matrices if each of the significantly nested matrices has a $p < 0.05$? Combinatorically, such a result is highly improbable and given by a binomial distribution with resulting $p \ll 10^{10}$. Likewise, the finding of an excess of antimodular matrices (20 of 38) compared

with a small number of modular matrices (6 of 38) is a highly improbable result. Moreover, most of the significantly modular matrices have low values of modularity, suggesting that, although modularity may be deemed significant in a few cases, it is not a driving mechanism underlying the structure of most of these matrices and may be incidental to other patterns. Together, these results imply that currently available host–phage infection networks are typically nested and not modular.

2.2.4 Previously Overlooked Nested Patterns Uncovered

An additional power of subjecting host–phage infection networks to a unified analysis is that, by doing so, we can extract meaningful biological information about the organization of a system that may not have been possible given the original placement of hosts and phages in matrix format. For example, the work by Zinno et al. [197] mentions variability in phage infection; however, Zinno et al. [197] make no mention of the fact that there are evidently groups of phages that preferentially infect groups of hosts (Figure 4A). Such block-like variability suggests that resistance mechanisms are less haphazard than they seem when network characteristics are not analyzed. Similarly, the work by Holmfeldt et al. [86] highlighted the variability and possibly unique signature of infection for each host and phage. However, reordering hosts according to the number of infecting phages while also reordering phages based on the number of hosts that they can infect leads to a nested pattern, suggesting that specific forms of infection rules may underlie infection variability (Figure 4B). To what extent is our finding of nestedness novel? As a reminder, nestedness is a property of a host–phage infection matrix as calculated for a given row and column ordering. Hence, we calculated nestedness for all of the matrices in the format as they were first reported in the literature and then compared these results to the nestedness calculated from our reshuffled matrices. We found that, in 35 of 37 cases of the previously published studies, the reshuffled matrix had a nestedness value higher than that of the original

publication, whereas in 2 of 37 studies, the nestedness was equal [102, 118] (Appendix A, Figure 25). Hence, our results suggest that, by and large, prior efforts did not identify the extent to which their matrices were nested or whether such nestedness was significant.

2.2.5 Addressing Sample Composition Biases as Potential Drivers of Network Structure

We report a set of analyses to quantify the extent to which potential biases might impact our results. One potential bias in our study derives from the methods some researchers used for phage isolation. Phages require a bacterial host to reproduce, and therefore, the bacterial host(s) chosen by the researcher can affect the form of the interaction matrix. For instance, if researchers used a single host to isolate phages and included this host in the matrix, then their matrix will necessarily possess a full row of positive infections, thereby introducing the first element of a perfectly nested matrix. We found only six studies that used such an approach [101, 102, 114, 118, 147, 161]. To determine if phage isolation strategy biased our results to nestedness, we reanalyzed all six of these matrices after removing the isolation host(s). We found no significant difference between the nestedness and modularity for each of these six matrices with or without the excluded host (Appendix A, Table 12).

Another potential bias is that studies included zero rows and columns, which implies that there are hosts that no phages infect and phages that do not infect hosts, respectively. Note that inclusion of zero rows and columns has the potential to bias the structure to a nested pattern. However, such zero rows and columns may be biologically meaningful if hosts or phages have evolved resistance that leads to noninteraction between particular sets of strains. Nonetheless, we performed the entire analysis again by generating alternative matrices such that hosts and phages were only included if they had had at least one nonzero element in their row or column, respectively. Then, we recalculated nestedness for the modified matrices

and compared it to the nestedness of appropriately resized null matrices. We found that 26 of 38 studies were nested compared with 27 of 38 using the original analysis (Appendix A, Figure 26). Moreover, although the quantitative value of nestedness did decrease in one case, that particular study [49] was, in fact, still highly nested and marginally significant at a $p = 0.067$ level. We also recalculated modularity for the modified matrices and found that 9 of 38 are modular compared with 6 of 38 in the original analysis (Appendix A, Figure 27). Hence, although there are minor changes in the number of significantly nested and modular networks, our finding that matrices have a characteristic nested structure is robust to either of these sources of bias.

Finally, we ask whether there are certain characteristics of matrices that defy the general pattern of nestedness and if it is possible to learn from these outliers? Interestingly, the three matrices with the most significant modular structures [54, 140, 197] were determined for a single bacterial species, Streptococcus thermophilus, and its phages. This finding seems robust, because different laboratories performed the studies and the microbes were isolated from three separate continents. Additionally, we did not find an example where a matrix that included S. thermophilus did not have the modular structure. We examined bacteria from the same taxonomic order (Lactobacillales) and isolated from the same environment (dairy products), but these bacteria lacked a modular structure. The consistent modularity observed for this species suggests that species-specific traits may have strong deterministic effects on the form that their interactions with parasites take. We are unsure of which traits produce the modular interactions; however, additional research may help reveal if and what resistance mechanisms determine the shape of microbial interaction networks.

2.2.6 Possible Scale Dependence of Host–Phage Interactions: From Nestedness to Modularity?

The data that we analyzed included almost 12,000 separate attempts to infect a host isolate with a phage isolate. Although the scale of the current data is beyond the scope of any individual project, it still pales compared with the number of possible interactions in a community at local or regional levels. Scaling up to larger assays presents technical challenges aside from increasing the depth of sampling. Studying many host strains beyond the species (or genus) level often requires distinct culture conditions, a prerequisite for studies that many laboratories cannot or do not want to reach. Here, we present an analysis of what such a hypothesized study may reveal. Consider an experiment in which the hosts from two groups of experiments were combined in a large cross-infection assay with the phages from the same two groups of experiments. If the original matrix sizes were $H_1 \times P_1$ and $H_2 \times P_2$, then the final matrix size is $(H_1 + H_2)(P_1 + P_2)$. A total of $H_1 \times P_2 + H_2 \times P_1$ new experiments would need to be performed. If the hosts were of sufficiently distant types (e.g., E. coli and Synecoccocus), we should expect that nearly all of the new cross-infection experiments would lead to no additional infections. Hence, if the original matrices were nested, then the new matrix would have two modules, each of which was nested (Figure 8 has the results of such a numerical experiment). In other words, we predict that, at larger, possibly macroevolutionary scales, host–phage interaction matrices should be typified by a modular structure, even if there is nested structure at smaller scales.

2.3 Discussion

2.3.1 Summary of Major Results

We have established a unified approach to analyzing host–phage infection matrices. In so doing, we find that a compilation of 38 empirical studies of host–phage interaction networks is nested on average and not modular (Figures 5 and 6). In most cases, our finding of higher–order structure such as nestedness within an individual study was not previously observed, in that prior analyses of host–phage interaction matrices usually did not attempt to estimate the network characteristics examined here. We found that host–phage interaction networks are not perfectly nested and that interactions that defy perfect nestedness are typical throughout nearly all of the data. Additionally, we found no significant difference in nestedness or modularity based on taxa, sources, or isolation method. This dataset, although far larger than any individual study, is limited to (largely) microevolutionary scales, an issue that we addressed in Results and will return to later in Discussion. Considering the large range of taxa, habitats, and sampling techniques used to construct the matrices, the repeated sampling of a nested pattern of host–phage infections is salient, although the process driving the nestedness is not obvious. It could result from multiple mechanisms or a single principle. Here, we examine three hypotheses to explain the nestedness pattern based on biochemical, ecological, and evolutionary principles. Note that these hypotheses are not mutually exclusive and that we have only limited ability to test them given our comparative approach. However, each of these hypotheses can be tested with additional laboratory- based or field experiments.

2.3.2 Mechanisms Responsible for Nestedness: Biophysical, Ecological, and Evolutionary

Phage and bacterial infection matrices at microevolutionary scales may be constrained to a nested shape by the nature of their molecular interactions. Phages infect bacteria

by using specialized proteins that target and bind to molecules on the outer membranes of bacteria (receptor molecules). Nested infection matrices have been shown for T-phages, which infect strains of E. coli, to be the result of the interactions of the phage proteins and receptor molecules [63]. T-phages bind to the lipopolysaccharide (LPS) chains on the cell surface. Mutant E. coli has been observed with shortened LPS chains that confer resistance to some but not all T-phages. There are T-phages that are able to infect these mutants, because they require fewer segments of the LPS molecule to bind. If phage–bacterial molecular interactions are dominated by single traits and variation in these traits is constrained along a single hierarchical dimension such as LPS, then one should expect the nested pattern to arise. There are other examples of traits with physical characteristics that behave similarly: bacteria that evolve a thicker and thicker protective coating [103], phages that evolve increased host range by continually reducing tail length [63], bacteria that reduce their number of receptors, and phages that target fewer receptors [89]. Although there are many examples of this type of one-dimensional interaction, the problem with this finding being a universal explanation for the form of bacterial–phage interactions are that host–phage interactions are governed by hundreds of other genes [113], bacteria can use multiple strategies for resistance [103], and phages have complex mechanism to evade bacteria defenses [103, 87]. Moreover, a recent discovery of an adaptive immune system, where bacteria acquire targeted sequences to prevent phage infection and phages evolve to evade such immunity, suggests a complex interaction space [98]. Given the diversity of host–phage interactions, it seems unlikely that the molecular details alone would constrain the form of their relationship [78]. Instead, we turn to the potential guiding forces of community assembly and coevolution to explain this reoccurring pattern.

The nested pattern may be common, because the processes of microbial community assembly select for species with nested relationships. One could imagine that

communities may settle into this pattern if this interaction structure is more stable than others [18, 20], noting that the stability of host–phage interaction structures may depend on ecological factors such as resource availability [137]. Cohesive interaction structures such as nested patterns have been shown to be more stable than other structures for mutualistic networks [16, 15]. The regularity of the interactions and redundancies make these communities less susceptible to the random removal of nodes. However, these networks are thought to be susceptible to invasion by new species that violate the nested pattern, suggesting that migration of a species would perturb the nestedness. Furthermore, the spatiotemporal complexity of microbial and viral communities suggests that prior theoretical efforts that consider community addition as a process in which invasions occur infrequently may not be widely applicable. Moreover, community assembly models rarely invoke the influence of evolutionary change at similar time scales as ecological change–an issue highly relevant to the study of microbial and viral communities.

Indeed, there may be an evolutionary explanation for nestedness. Most attempts to characterize the form of coevolution with host–phage experiments to date have shown a form of antagonistic evolution called expanded host range (or gene for gene) coevolution [122, 105, 28]. Under this model, bacteria evolve ever-increasing resistance to more and more phage genotypes, and phages evolve broader host ranges. If one were to sample a community of bacteria and phages coevolving under this model, they would uncover a diversity of phages and bacteria that exhibit a nested interaction pattern. At any time point, the most-derived bacteria should exist, which is either completely resistant or depending on the timing, sensitive to the most-derived phage. Given that selection by phage may be slow to alleviate the more sensitive ancestral variants or that there may be a trade-off between resistance and competitiveness, there will exist a diversity of bacteria with ever decreasing sets of phages to which they are resistant. Similarly, the most-derived phages will have the broadest host

range, and by the same logic as for the bacteria, its ancestors are likely to persist in the community and display ever-decreasing host ranges. The nested pattern could be a product of taking a snapshot of a dynamically evolving community. Although the majority of experimental results observed in artificial laboratory settings support this hypothesis, there is a single laboratory experiment [75] and models of bacterial host–parasite coevolution that suggest that other forms of coevolution are possible when there are bottom-up costs for modifications to resistance [186, 159]. Furthermore, if coevolution provided the only explanation, then the artificially assembled matrices would not have the nested pattern.

2.3.3 Dispelling and Recognizing Potential Biases

Three sources of sampling bias challenge the generality of our findings. First, the taxa sampled may poorly represent microbial diversity given that they are subject to both human and methodological biases. If, for instance, only taxa associated with humans were selected or all taxa were cultured similarly, then our results would only be relevant for a small group of microbes. Indeed, the majority of microbial studies were performed on the family Enterobacteriaceae, which lives within human digestive systems; however, the spectrum of bacteria that we examined is much broader and includes both heterotrophic and photosynthetic species. Further, gram-negative and -positive bacteria examined here were isolated from six continents and many disparate environments from the extreme conditions of hot springs, the rich resource conditions of sewage, depauperate marine environments, and the complex matrix of soil to the simplified laboratory environment. Although this study cannot feasibly test the full microbial diversity of the globe, it does include examples from much of it (Appendix A, Tables 7 and 8).

Second, as previously discussed, the number of hosts used to isolate phages and the inclusion of noninteracting hosts and phages have the potential to alter the nestedness

of a matrix. Ideally, the same number of hosts studied in the matrix would be used to isolate phages, or if only a subset of hosts was used, then these hosts would not be included in the matrix. This finding is important to ensure that the pattern of infection is independent of how the parasites were isolated. We found that these biases were not a problem by (*i*) testing matrices that were created by isolating phages on a single host and (*ii*) removing hosts and phages that were not interacting. We found that whether the matrices were significantly nested was not affected by including the isolation host in the matrix or by removing noninteracting hosts and phages, which is strong support that the isolation method did not enrich for nestedness.

The last category of bias, phylogenetic, is likely to mean that our results define a pattern at relatively narrow taxonomic scales. The majority of our studies was of closely related genotypes and species. As described in Results, we anticipate that more complex patterns of infection may form at larger phylogenetic scales that likely include increasing compartmentalization. Hence, we hypothesize that a multiscale view of host–phage infection networks will reveal nestedness at small scales and modularity at large scales. Our finding of nested interaction matrices is still relevant for characterizing patterns at short phylogenetic distances; they are, arguably, the most relevant for many ecological and evolutionary scenarios, because they likely share the richest connections.

2.3.4 Prospective View

Whatever the limitations of this dataset, it is important to point out that viewing host–phage interaction networks through a unifying lens will likely unveil other commonalities of microbial and viral communities. By way of analogy, over 25 y ago, the study of food webs was radically altered by the compilation of many small food webs that were subject to a unified analysis [38, 39, 134, 37]. The key finding of the earliest food web studies was that the members of a community could be ranked, and

that larger species would eat a random fraction of those species smaller than them. From this stage, there were two ways forward. First, by studying larger food webs, the original pattern was refined such that species ranking was found to be correlated with body size (but not equivalent to body size); therefore, individuals eat prey that are smaller, although they are a part of a well-defined size class [5, 191]. Second, the topology of food webs was then used as a target and basis for dynamic models of community behavior (i.e., what mechanisms can explain the patterns and how do the patterns influence community function) [132]. We hope and envision that a similar process unfolds here in that the finding of a general pattern in the current dataset will stimulate the collection of more and larger host–phage infection networks to continue to provide a fuller picture of who infects whom across an entire community. In so doing, we caution that data completeness can alter the observed patterns of connectivity and refer readers to a number of recent papers that address this topic [116, 129, 183, 66, 72].

What do we expect to find when analyzing ever larger host–phage interaction networks collected from an ecological community, evolution experiments, or culture collections? We hypothesize that host–phage interaction matrices are likely characterized by modularity at larger taxonomic scales even if there is structure (e.g., nestedness) at small taxonomic scales (Figure 8). What would such a multiscale phenomenon inform us about the structure and function of microbiological communities? First, it would suggest the existence of diversifying coevolutionary-induced selection that gave rise to (largely) independent host–phage communities. The molecular basis of such diversification could then be explored. Second, cross-infection assays or similar laboratory-based strategies [36] that test whether phages can infect or at least transmit their genes between phylogenetically divergent hosts have the potential to provide significant advances in understanding patterns of global gene transfer. Such phages (and the bacteria that they infect) may be critical to understanding the direct

transfer of genes on a global scale. Instead of phages acting locally (in a taxonomic sense) to shuttle genes between closely related bacteria, a few rare links would permit greater cross-talk between bacterial taxa. Quantifying the frequency of such events may represent the small- world links that connect distant microbial populations [184], and it is in need of experimental testing.

Furthermore, infections of distantly related groups by the same phages would imply that the bacteria are in indirect competition with one another, even if they do not seem to compete directly for the same set of carbon and nutrient sources. Although whole genome-based approaches to infer host range and phage susceptibility may help provide candidates for such rare links, they are not the only solution. Rather, we suggest that the continued use of laboratory-based assays to catalog the life history traits of culturable host–phage pairs is essential if we are to improve our understanding of the population dynamics of host–phage communities in the wild. Of course, many (if not most) bacteria and phages are not currently culturable. Hence, in parallel, we recommend attention be given to the development of inverse methods to catalog the life history traits of phages based on community infection assays in those circumstances in which culturing is impossible or yet intractable.

2.4 Materials and Methods

2.4.1 Network Statistics

Modularity is estimated by reshuffling the rows and columns of the matrix to find groupings of highly interconnected phages and bacteria, labeling these groups and assessing matrix-wide the ratio of the number of within to outside group connections. This calculation is done using a heuristic called the BRIM algorithm [13] to efficiently find the configuration that maximizes this ratio. We ported the BRIM algorithm to MATLAB from the original code in Octave and used the adaptive BRIM algorithm for all calculations here. By this definition, a perfectly modular matrix is comprised

of clusters of completely isolated groups, and modularity declines as the number of cross-group connections increases. Nestedness is estimated by reordering the rows and columns [11, 143] to determine whether phages that infect fewer hosts are only able to infect a subset of bacteria that are susceptible to many phages. This reordering tries to maximize the position of ones in the matrix such that they clusters above a nullcline (Figure 2C shows a perfectly nested matrix). The value for nestedness depends on how frequently ones fall above rather than below this nullcline. Complete details are provided in Appendix A, Materials and Methods.

2.4.2 Host–Phage Infection Assay

Matrix 22 is the only dataset not previously published.Weconstructed thematrix by coevolving an obligately lytic phage- strain with its host E. coli. The E. coli studied was of strain REL606, a derivative of E. coli B acquired from Richard Lenski (Michigan State University, Lansing, MI) and described in ref. [45], and phages were of strain cI21 (vir) provided by Donald Court (National Cancer Institute, Frederick, MD). The phages and bacteria were cocultured in 50-mL Erlenmeyer flasks with 10 mL liquid medium, shaken at 120 rpm, and incubated at 37 °C (New Brunswick Innova 4300 Incubator Shaker). This flask was incubated, and the cycle of transfer and incubation was continued one more time. Three 24-h incubations were long enough for the bacteria to evolve resistance and the phages to counter it; however, it was not long enough for a second round of coevolution. We randomly selected 150 bacteria and 150 phage isolates. We determined which of the 150 bacteria isolates were resistant to the 150 phage isolates. To do this task, we performed spot plate assays. All bacterial– phage combinations were replicated five separate times, and a total of 28,125 spots were assayed. To make this processes more efficient, we placed up to 96 separate phage stocks onto a single dish (150 mm radius). Phage stock replicates were never placed on the same plate to reduce the signal of any stochastic plating

effects. The five replicates were combined, and a phage was only determined to be able to infect a bacterium if three of five replicates were given ones. Lastly, phages or bacteria that had identical infection or resistance profiles as their ancestors were removed from the matrix. Complete details are provided in Appendix A, Materials and Methods.

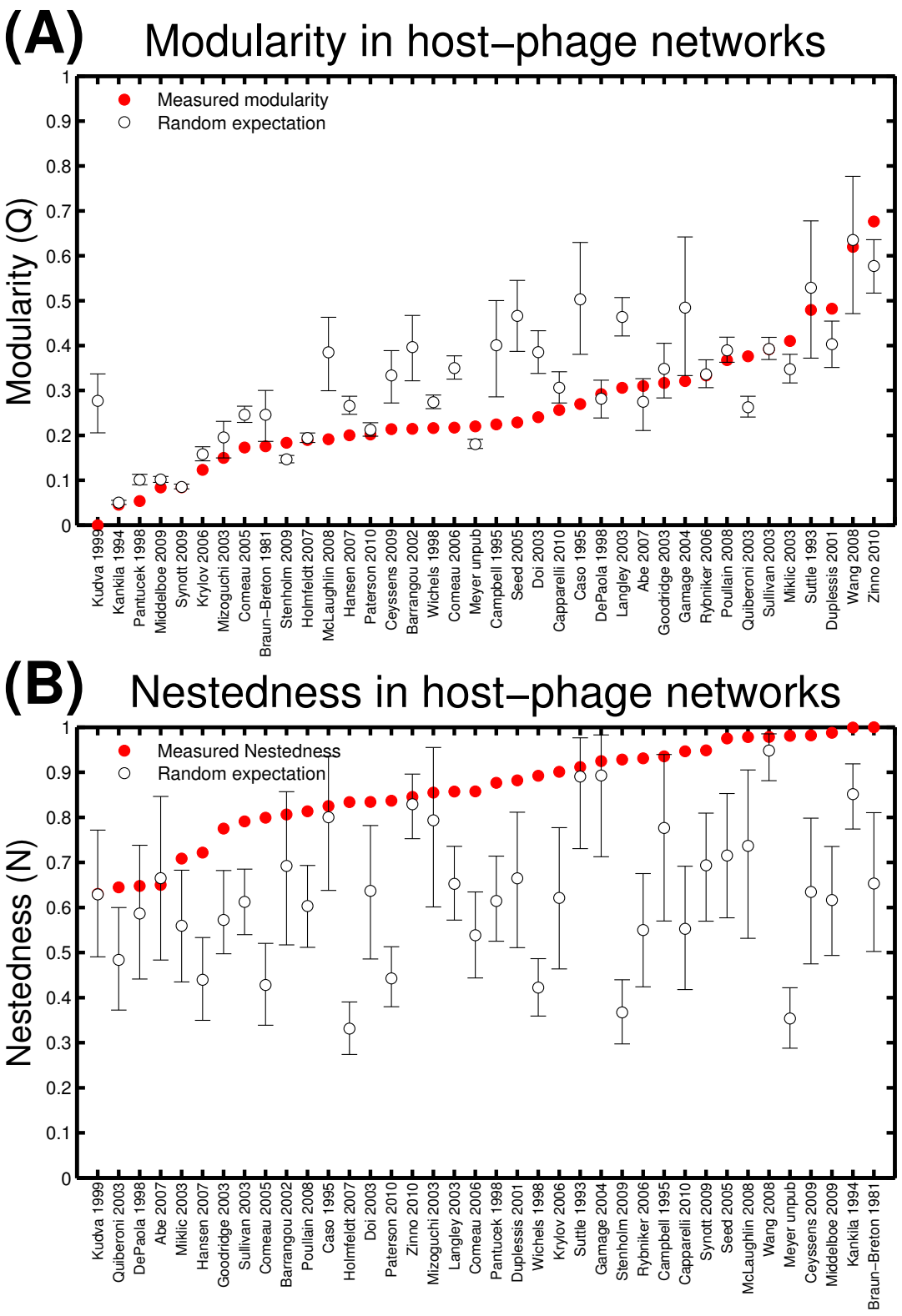

Figure 7: Statistical distribution of modularity and nestedness for random matrices compared with that of the original data. **(A)** Sorted comparison of modularity of the collected studies vs. random networks. **(B)** Sorted comparison of nestedness of the collected studies vs. random networks. In both cases, error bars denote 95random-izations.

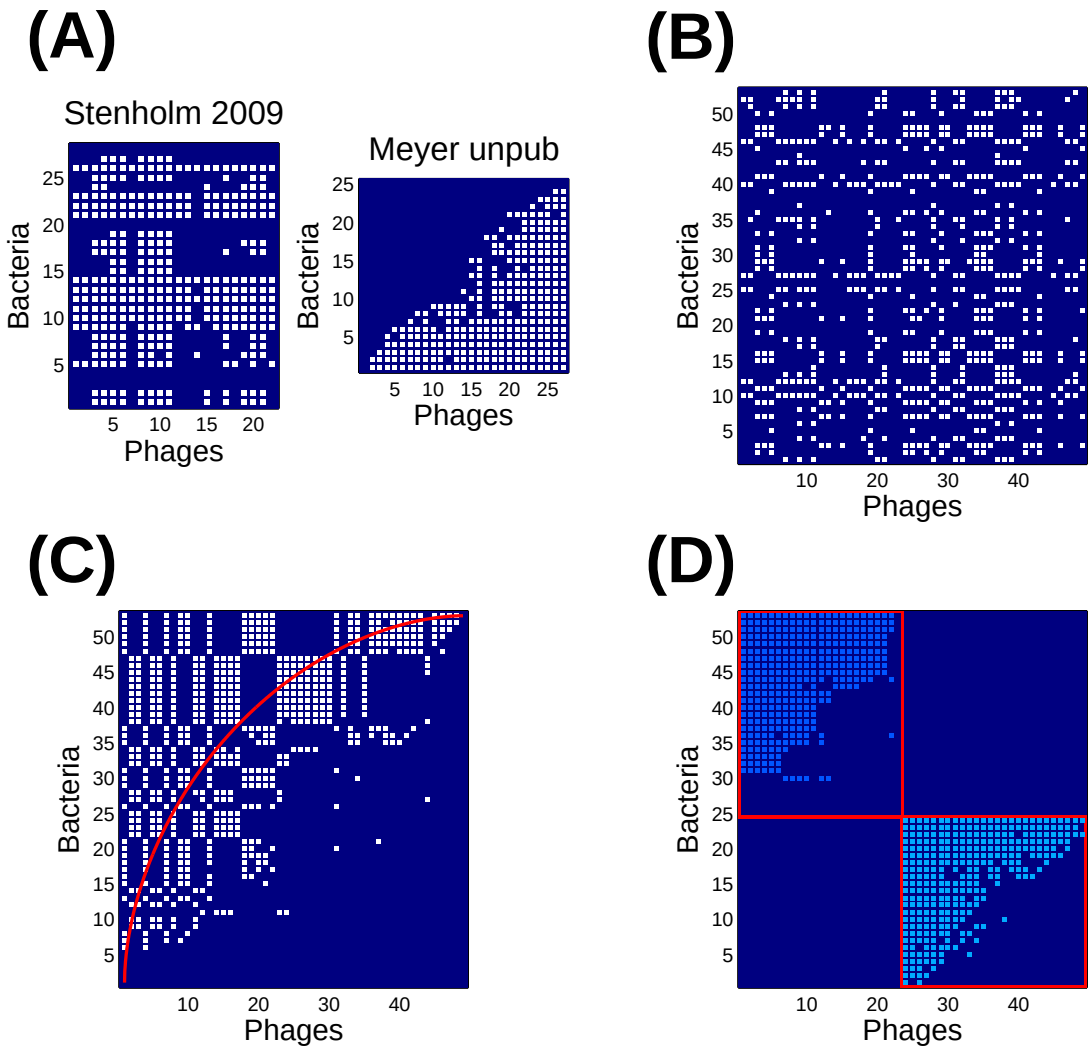

Figure 8: Union of two nested matrices indicates possible host–phage interaction structure at larger, possibly macroevolutionary scales. In this figure, we selected two of the most nested studies and performed a union while presuming that there were no cross-infections of hosts by phages of the other study. In this case, E. coli and cyanobacteria were the host types. **(A)** Depiction of the original matrices. **(B)** Randomization of the union matrix. **(C)** Nested sort of the union matrix. **(D)** Modularity sort of the union matrix with a nested sort of each module.

CHAPTER III

MULTI-SCALE STRUCTURE AND GEOGRAPHIC DRIVERS OF CROSS-INFECTION WITHIN MARINE BACTERIA AND PHAGES

Adapted from Cesar O. Flores, Sergi Valverde, and Joshua S. Weitz. Multi-scale structure and geographic drivers of cross–infection within marine bacteria and phages. ISME Journal (2013) [62].

Bacteriophages are the most abundant biological life forms on Earth. However, relatively little is known regarding which bacteriophages infect and exploit which bacteria. A recent meta-analysis showed that empirically measured phage-bacteria infection networks are often significantly nested, on average, and not modular. A perfectly nested network is one in which phages can be ordered from specialist to generalist such that the host range of a given phage is a subset of the host range of the subsequent phage in the ordering. The same meta-analysis hypothesized that modularity, in which groups of phages specialize on distinct groups of hosts, should emerge at larger geographic and/or taxonomic scales. In this paper, we evaluate the largest known phage-bacteria interaction data set, representing the interaction of 215 phage types with 286 host types sampled from geographically separated sites in the Atlantic Ocean. We find that this interaction network is highly modular. In addition, some of the modules identified in this data set are nested or contain sub-modules, indicating the presence of multi-scale structure, as hypothesized in the earlier meta-analysis. We examine the role of geography in driving these patterns and find evidence that the host range of phages and the phage permissibility of bacteria is driven, in part, by geographic separation. We conclude by discussing approaches to disentangle

the roles of ecology and evolution in driving complex patterns of interaction between phages and bacteria.

3.1 Introduction

Bacteriophages can have a significant effect on microbial communities and ecosystems [190, 194, 165, 166, 27]. Bacteriophages are responsible for a significant fraction of bacterial mortality [167, 185], engage in coevolutionary arms races with their hosts [28, 9, 85, 110], and redirect organic material to the microbial loop via a process known as the viral shunt [190, 119, 94]. A key event in all of these ecological functions is the interaction with and exploitation of a bacterium by a phage. It is widely hypothesized that phages can infect a very limited subset of bacteria in a given environment. However, given the high diversity of bacteria in natural environments [146, 141], even infecting a limited subset can nonetheless represent a heterogeneous range of hosts. Indeed, there is a long record of evidence to suggest that phages commonly infect multiple distinct bacterial types in natural environments (for example, [189, 86], including examples where individual phages can infect hosts from distinct genera (for example, cyanophages infecting hosts from *Prochlorococcus* and *Synechococcus* [163]). Recently, we utilized a network-based approach in order to identify and characterize patterns within published data sets of infection and exploitation of bacteria by phages [60].

The key interaction patterns examined in Flores et al., (2011) were nestedness [143, 179, 7, 178] and modularity [128, 13]. In the context of phage-bacteria interactions, nestedness indicates the extent to which the host ranges of phages are subsets of one another. In a maximally nested network, the most specialized phage could infect hosts most permissive to infection. Then, the next most specialized phage could infect the host most permissive to infection as well as one additional host, and so on. Nestedness is thought to emerge in coevolutionary arms race dynamics in which hosts

evolve resistance to current and past pathogens, while pathogens evolve counter resistance that enables them to infect past hosts [4], for example, as observed between the bacterium Pseudomonas fluorescens SB25 and the DNA phage SBW25F2 [28]. Similarly, modularity indicates the extent to which interactions, in this case an infection of a bacterium by a phage, can be partitioned into groups with many interactions within them and few interactions between them. These groups are referred to as modules. In a maximally modular network, there would be no cross-infections between phages of one module and hosts of another module. There are many possible drivers of modularity, including geographic isolation, which can facilitate the divergent coevolution of interacting species [174, 75].

In our re-analysis of published studies, we found that infection networks tended to be nested and not modular [60]. However, we hypothesized that modularity should be expected when a greater diversity of bacteria and phages interact. The work described here follows up on our earlier study by analyzing a previously published cross-infection data set [124] not included in our earlier analysis. The Moebus and Nattkemper (1981) data set is the largest phage–bacteria infection network available in the literature (as far as we are aware), representing interactions between marine phages and bacteria in the Atlantic Ocean. The data set contains cross-infection and geographic information but no sequence information. As such, we focus our analysis on the following questions: (i) how do patterns of infection change at different scales, that is, when examining the entire network (large scale) vs subcomponents of the network (small scale); (ii) what role does geographic separation have in shaping cross-infection? Despite the cosmopolitan nature of viruses [26, 10] (for an exception see [48], multiple lines of evidence suggest that phages are often better adapted to hosts from the same location than they are to hosts from a different location [85, 181, 75, 100]. Hence, by examining explicit cross-infections among many microbes isolated across a large geographic range, we hope to shed light on the structure of

phage–bacteria infection networks.

3.2 Materials and methods

3.2.1 Data set

We analyzed the cross-infection data set of Moebus and Nattkemper (1981). This data include phage and bacteria collected from February to April 1979 in the Atlantic Ocean between the European continental shelf and the Sargasso Sea [123]. Bacteria were cultured and isolated using seawater-based media and bacteriophages were enriched from the same water sample [123]. In the original analysis of cross-infection [124], the authors describe cross-reaction tests among 733 bacteria and 258 phage strains collected at 48 stations separated, in some cases, some 200 miles apart (Appendix B, Figure 28). However, the authors do not report results from strains, which have both (i) identical infection patterns and (ii) that were isolated from the same station. The reported data set is included as a fold-out table in the main text (see Appendix B, Figure 29). We digitized and automatically extracted the positive infection results and then manually curated the results, yielding a network of 286 bacteria strains and 215 phage strains with 1332 positive infection outcomes out of a possible $61,490 = 286 \times 215$ interactions (see Appendix B.1 for more details). The interactions were classified in the original study as either (i) 'More or less clear spots due to lysis of bacteria'; (ii) 'More or less turbid spots'. We classified all interactions as either positive (either clear or turbid spots) or negative (neither clearing nor turbid spots). We refer to this data set as the MN (Moebus and Nattkemper) matrix. The resulting digitized data set is shown in Figure 9.

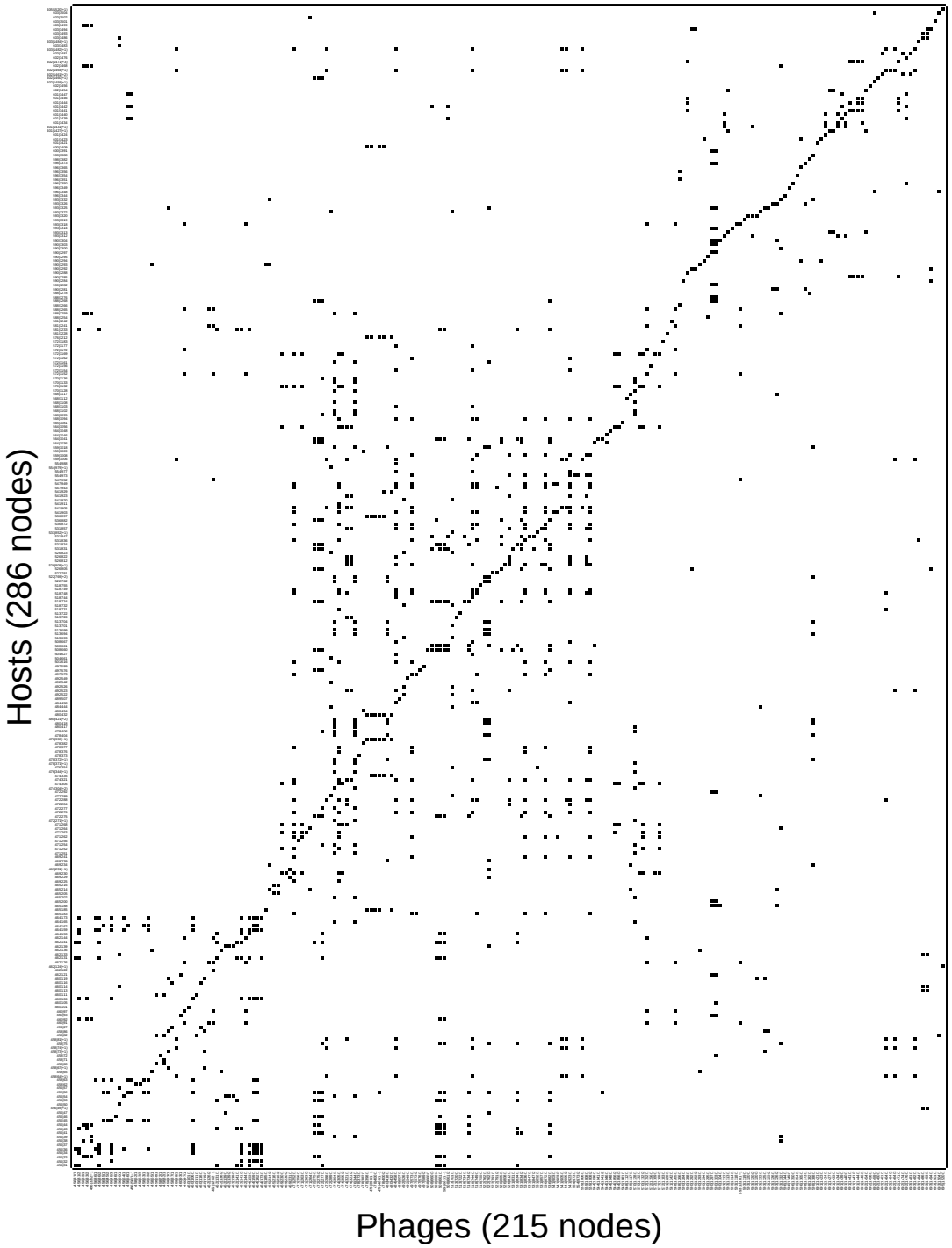

Figure 9: Digitized version of the MN matrix with 286 hosts (rows) and 215 phages (columns) in the same orientation as originally published (Moebus and Nattkemper, 1981). The 1332 black cells represent positive interactions between hosts and phages (see Materials and methods). The connectance of the network (interactions/total size) is approximately $0.022 \approx 1332/61490$.

3.2.2 Network Analysis

3.2.2.1 Disjoint components

An interaction network is considered bipartite when it contains two types of agents that interact, for example, bacteria and phages. Any bipartite network can be decomposed into disjoint components such that no cross-infections are found between components. Formally, each disjoint component in a bipartite network of host-viral cross-infection is defined in terms of a set of hosts, **H**, and viruses **V**, such that: (i) there is no virus V outside of **V** that can infect any host in **H**; (ii) there is no host H outside of **H** that can be infected by any virus in **V**; (iii) for each virus in **V** there is at least one host in **H** that it can infect.

3.2.2.2 Modularity

We used the standard BRIM (Bipartite Recursively Induced Modules) algorithm [13], which utilizes a local search heuristic to maximize a bipartite modularity value Q (see Appendix B.2 for more details). The value of Q represents how often a particular ordering of phages and bacteria into modules corresponds to interactions that are primarily inside a module ($Q \approx 1$ or modular), primarily outside of modules ($Q \approx -11$ or antimodular) or somewhere in between ($-1 \leq Q \leq 1$). BRIM helps find the arrangement of phages and bacteria in modules that maximize Q. We used two different approaches of the BRIM algorithm depending on the size of the matrix. For the entire matrix, we extended the BRIM algorithm to first partition the network into different isolated modules and then subsequently recursively subdivide the network as has been done in the case of unipartite networks [127, 128], that is, networks with only one type of node. Our approach (described in Appendix B.2) yields higher values of Q than both BRIM and LP-BRIM [108]. Within each module, we used the adaptive heuristic of the BRIM algorithm [13], which has been verified to perform well in small matrices [108].

3.2.2.3 Nestedness

We utilized two algorithms to measure the extent to which hosts and phage interactions have a nested pattern.

3.2.2.4 Nestedness Temperature Calculator

The nestedness temperature calculator (NTC) algorithm was originally developed by [11] and has been reviewed elsewhere [143]. In the present context, the 'temperature', T, of an interaction matrix is estimated by resorting the row order of hosts and the column order of phages such that as many of the interactions occur in the upper left portion of the matrix. In doing so, the value of T quantifies the extent to which interactions only take place in the upper left ($T \approx 0$), or are equally distributed between the upper left and the lower right ($T \approx 100$). Perfectly nested interaction matrices can be resorted to lie exclusively in the upper left portion and hence have a temperature of 0. The value of temperature depends on the size, connectance and structure of the network. Because the temperature value quantifies departures from perfect nestedness, we define the nestedness, N_{NTC}, of a matrix to range from 0 to 1, $N_{NTC} = (100 - T)/100$, such that $N_{NTC} = 1$ when $T = 0$ (perfect nested pattern) and $N_{NTC} = 0$ when $T = 100$ (chessboard pattern).

3.2.2.5 Nestedness metric based on overlap and decreasing filling

NODF is a nestedness metric introduced by Almeida-Neto et al. (2008) [7]. NODF is independent of row and column order. This algorithm measures the nestedness across hosts by assigning a value M_{ij}^{H} to each pair i, j of hosts (rows) in the interaction matrix, which is defined as:

$$M_{ij}^{H} = \begin{cases} 0, & \text{if } k_i = k_j \\ n_{ij}/\min(k_i, k_j), & \text{otherwise} \end{cases} \tag{1}$$

where k_i and k_j are the degree of hosts i and j respectively, and n_{ij} is the number of common interactions between them. 'Degree' is a standard network science term that is defined as the number of interactions that a given type has [126]. For example, in this context, the degree of a host is the number of viruses that can infect it and the degree of a virus is the number of hosts it can infect. The same method is used to calculate nestedness across phages, such that the total nestedness value is:

$$N_{NODF} = \frac{\sum_{i<j} M_{ij}^{H} + \sum_{i<j} M_{ij}^{P}}{\frac{H(H-1)}{2} + \frac{P(P-1)}{2}} \tag{2}$$

The meaning of nestedness as calculated by NODF is that higher values denote matrices whose (i) pairs of rows are typically subsets of each other, that is, host pairs share some, but not all, viruses that can infect them; (ii) pairs of columns are typically subsets of each other, that is, viral pairs share some, but not all, hosts that they can infect.

3.2.2.6 Null Models

We utilized two null models in order to measure the statistical significance of modularity and nestedness. The first is a Bernoulli random null model in which the null matrix has the same total number of interactions as the original matrix, albeit randomly positioned. The second is a probabilistic degree null model in which each interaction between host i and phage j in the null matrix is assigned with a probability p_{ij} according to:

$$p_{ij} = \frac{1}{2}\left(\frac{k_i}{P} + \frac{d_j}{H}\right) \tag{3}$$

where the degree k_i is the number of phages that infect host i, the degree d_j is the number of hosts infected by phage j, P is the number of phages and H is the number of hosts. In all cases, we utilize $100,000$ random matrices to evaluate the statistical significance of modularity and nestedness. Finally, given the two null models, we evaluate modularity using two significant tests, and we evaluate nestedness using four significance tests (two each for the NTC and NODF).

3.2.3 Multi-scale Analysis

Nestedness metrics may overestimate the statistical significance of nestedness, particularly when the fraction of realized interactions of a network becomes either very large or very small, for example, [59]. In addition, in cases where a network is comprised of nested modules, we expect that some nestedness measures will spuriously identify the entire network as nested (see for example, Figure 7 of Flores et al. (2011)). We developed two approaches to characterize nestedness given a large, sparsely connected network. These two approaches are consistent with recent calls to take a local, rather than a strictly global, approach to identifying community structure [66]. First, in the case of nestedness as calculated using NTC, we identify modules in the original matrix, and then constrain the row/column re-ordering so that rows and columns cannot break the modular structure. Hence, we still sort the rows and columns, but only inside modules. In addition, we permit random permutations of the modular blocks along the main matrix diagonal and select the configuration that minimizes temperature (maximizes nestedness). Second, in the case of nestedness as calculated using NODF, we again identified modules and then restricted the comparisons of overlap to rows and columns across modules. In this way, we can evaluate the overall nestedness of the original matrix without considering the nestedness contribution that comes from inside of modules. More details are found in Appendix B.3.

3.2.4 Geographical Analysis

Modules identified in our network analysis include hosts and phages collected at potentially different sample sites. The sample site of each phage and host corresponds to different 'stations' in the Atlantic Ocean. We estimated the geographic diversity of stations within a given module using Shannon (H_k) and Simpson indices (D_k) [153, 157] where the subscript k denotes the module number. Both indices measure the variability in the stations of isolation of phages and hosts within a given module.

In addition, both indices were applied to hosts and phages separately. The diversity indices of a given module are:

$$H_k = -\sum_{i=1}^{R} \frac{n_i}{N} \log \frac{n_i}{N}, \text{ and } D_k = 1 - \sum_{i=1}^{R} \frac{n_i(n_i-1)}{N(N-1)} \tag{4}$$

where N are the number of different strains inside the module, R are the number of stations inside the module, and n_i are the number of strains from station i. Low values in both indices indicate low geographical diversity. We determined the significance of a measured diversity value by comparing observations with an ensemble of randomized matrix assignations of station labels to modules (see Appendix B.4 for details).

3.3 Results

3.3.1 Characteristics of a large-scale phage-bacteria infection network

The network properties of the MN phage-host infection data set are shown in Table 1. We find that only a small percentage of the cross-infections yield a positive result (2.17% = 1332/61490), in contrast to a previous meta-analysis where many cross-infections yielded positive results (36.6% =4365/11944) [60]. However, in agreement with the prior meta-analysis we find that phages can infect multiple hosts (average of 6.20, median of 4 in the present study, average of 8.75, median of 6 in the prior meta-analysis). Similarly, we find that hosts are infected by multiple phages (average of 4.66, median of 3 in the present study, average of 4.34, median of 3 in the prior metaanalysis). These averages and medians were calculated over all strains in the current study and by aggregating strains from the prior analysis. Importantly, the degree distribution of this network is not unimodal, that is, it does not have a single peak. Instead, we find long-tailed distributions of the number of hosts that a phage can infect, and similarly, the number of phages that can infect a host (see Appendix B, Figure 30). Hence, there exists a spectrum of viral types spanning specialists to generalists; we find there are many more specialists than generalist viral types in this

Table 1: General properties of a large-scale phage–bacteria infection network

General properties	Definition	Value
N_c	Number of components	38
H	Number of hosts	286
P	Number of phages	215
I	Number of interactions	1332
$S = H + P$	Number of species	501
$M = HP$	Size	61490
$C = I/M$	Connectance or fill	0.0217
$L_H = I/M$	Mean host degree	4.6573
$\max(k_i)$	Max host degree	20
$\min(k_i)$	Min host degree	1
$L_P = I/M$	Mean phage degree	6.1953
$\max(d_i)$	Max phage degree	31
$\min(d_i)$	Min phage degree	1
N_c	Number of components	38

study. Similarly, hosts can span a spectrum of types from permissive to resistant types; we find there are many more resistant types than permissive types in this study.

3.3.2 Evaluating modularity at the whole-network scale

The MN matrix is comprised of 38 disjoint components, that is, sets of phages and bacteria, which have cross-infections within a component but no crossinfections between components (see Figure 10). Given the finding of disjoint components, we expect that the MN matrix is significantly modular. We confirm this via a modularity analysis using the BRIM algorithm in which we identify 49 separate modules (see Appendix B, Table S2). The 49 modules include the subdivision of some of the 38 disjoint components as identified in the BRIM analysis such that the overall modularity value Q is increased. These results enable in-depth resolution of the specialization within the system, in contrast to the conclusion by Moebus and Nattkemper (1981) via visual inspection that 'two large groups of bacteriophage–host systems were encountered'

and '8 small ones were found'. Figure 11 shows the modularity sorting of the MN matrix resulting from the BRIM algorithm, in which rows and columns inside modules were sorted in order to highlight the possible nested structure within modules. Remarkably, 1219/1332 = 91.52% of the interactions occurs within modules rather than between modules. The calculated modularity of the MN matrix ($Q = 0.7950$) is larger than any of the 105 realizations in either null model ($p < 10^{-5}$), which is a conservative upper bound). As a point of reference, the highest value of any of the random matrices was $Q = 0.4503$. The z-score, representing the relative number of standard deviations the actual modularity is larger than the mean of the random ensemble, as calculated for modularity was 87.55 using the Bernoulli null model and 51.02 using the probabilistic degree null model. It is important to note that although most interactions occur within a module, these modules include phages and bacteria from multiple stations. Hence, we find that 76% ($\sim$1012/1332) of infections transcend the site of isolation.

3.3.3 Evaluating nestedness at the whole-network scale

We evaluated the nestedness of the MN matrix using a combination of algorithms and null models. First, we resorted the row and columns in order of increasing degree, a heuristic that tends to maximize nestedness using the temperature calculator. Visually, it would seem that the MN matrix is not nested (see Figure 11 and Appendix B, Figure 31). We showed in a previous study that a community of nested modules can lead to apparent nestedness at the whole-matrix scale [60]. Indeed, for the four nestedness tests (two null models and two algorithms) we find that the MN matrix is apparently significantly nested in all cases except for the NODF algorithm using the probabilistic interaction null model. We argue that the apparent finding of nestedness is driven by the fact that the matrix contains nested modules, rather than a nested arrangement of hosts and phages spanning the entire matrix. We applied

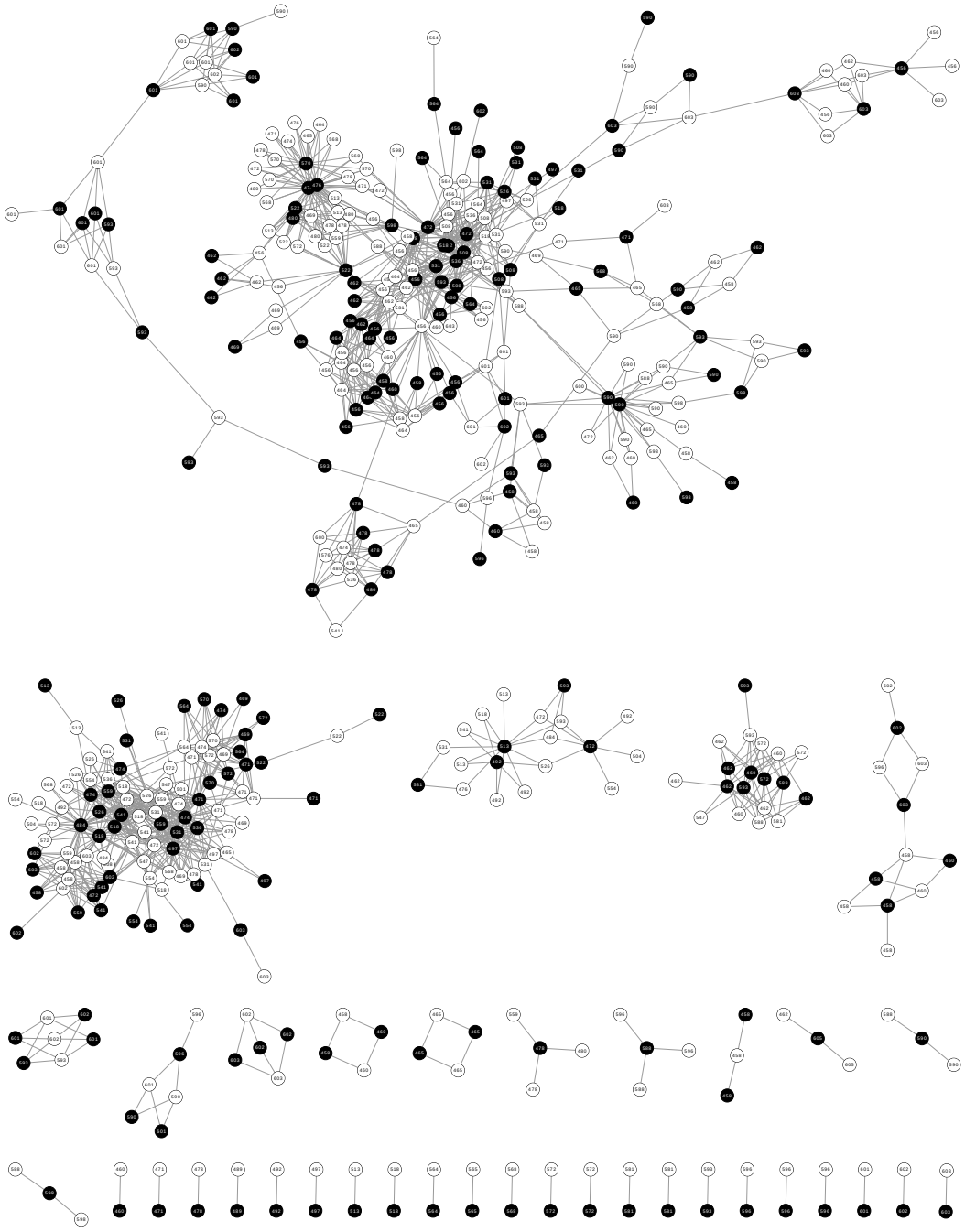

Figure 10: Network representation of the study. We observe 38 isolated components. Black nodes represent phages, and white nodes represent hosts. The station IDs of each host and phage are contained in the center of each node.

a multi-scale network analysis to evaluate this hypothesis (see Materials and methods and Appendix B.3). The results of the conventional and multi-scale nestedness

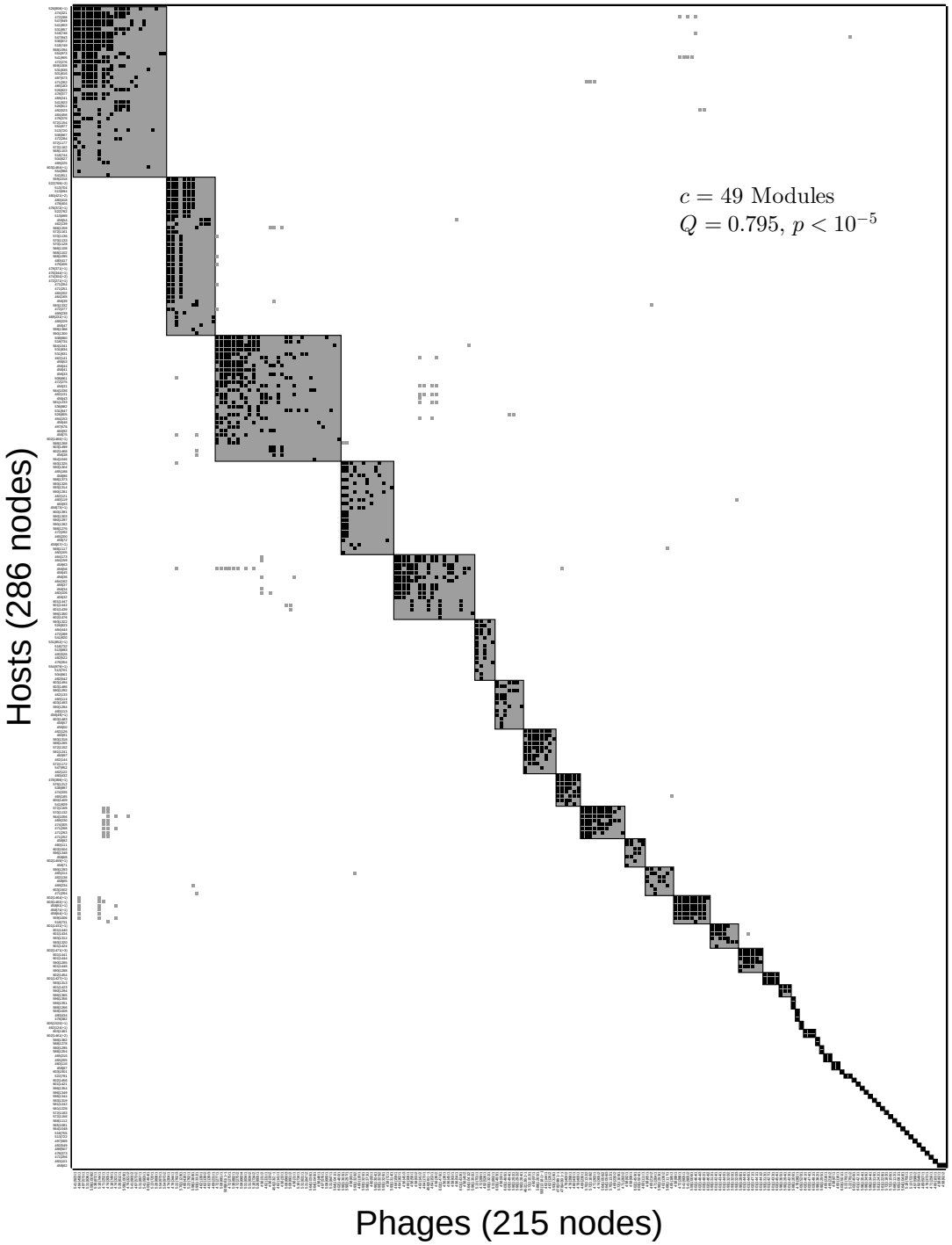

Figure 11: Modularity sorting of the network. We detect 49 modules (shaded rectangles). The 15 largest modules discussed in the main document begin at the left of the matrix. Black symbols represent those interactions within a module. Gray symbols represent those occurring between modules. The p-value for the observed modularity is smaller than 10^{-5}.

analysis are summarized in Table 2. The multi-scale analysis enables us to reject the finding of nestedness for both algorithms when using the probabilistic degree null model. Nestedness can also be rejected even in the case of the Bernoulli null model for NODF and for one of the multi-scale analysis methods using NTC.

Table 2: Significance of the nestedness of the MN matrix using alternative algorithms

	NTC algorithm			NODF algorithm		
	N_{NTC}	Bernoulli	Probabilistic degree	N_{NODF}	Bernoulli	Probabilistic degree
Normal analysis	0.9541	$p < 1e-5$	$p < 1e-5$	0.0341	$p < 1e-5$	$p = 0.2336$
Multi-scale analysis	0.93588	$p < 1e-5$	$p = 1$	0.0062	$p = 1$	$p = 1$
	0.9263	$p < 1e-5$	$p = 1$			
	0.8568	$p = 1$	$p = 1$			

Abbreviations: MN matrix, Moebus and Nattkemper matrix; NODF, nestedness metric based on overlap and decreasing filling; NTC, nestedness temperature calculator; The P-value denotes the fraction of random matrices that have a larger value of nestedness, N, than the observed MN matrix. In the 'normal' analysis, the NTC algorithm and NODF algorithms are used to estimate nestedness using alternative null models (see Materials and methods). For the multi-scale analysis three values have been reported for analyzing the significance of nestedness using the NTC algorithm: (1) Modules are sorted according to the sort heuristic described in Appendix B.3; (2) Modules are sorted in descending order of the number of phages; (3) Modules are sorted in ascending order of the number of phages. See Appendix B, Figure 33 for the details of sorting. Note that the values of nestedness can differ depending on the algorithm used, it is their relative value to the null model that determines significance.

3.3.4 Network analysis at the intra-module scale

Table 3: Network properties of the largest 15 modules identified using the modularity analysis (see Table 1 for definitions of all quantities)

#	H	P	S	I	M	C	Lp	Lh
1	42	23	269	65	966	0.28	6.4	11.7
2	39	12	138	51	468	0.29	3.54	11.5
3	31	31	233	62	961	0.24	7.52	7.52
4	23	13	61	36	299	0.2	2.65	4.69
5	16	20	114	36	320	0.36	7.13	5.7
6	15	5	30	20	75	0.4	2	6
7	12	7	27	19	84	0.32	2.25	3.86
8	11	8	52	19	88	0.59	4.73	6.5
9	8	6	38	14	48	0.79	4.75	6.33
10	8	11	57	19	88	0.65	7.13	5.18
11	7	5	15	12	35	0.43	2.14	3
12	7	7	17	14	49	0.35	2.43	2.43
13	7	9	49	16	63	0.78	7	5.44
14	6	7	21	13	42	0.5	3.5	3
15	6	6	27	12	36	0.75	4.5	4.5
Mean	**15.87**	**11.33**	**76.53**	**27.2**	**241.47**	**0.46**	**4.51**	**5.82**
Median	**11**	**8**	**49**	**19**	**84**	**0.4**	**4.5**	**5.44**

We performed a network analysis of the 15 largest modules extracted from the modularity sort (see Table 3 for summary statistics and Appendix B, Table 14 for information on all 49 modules). Figures 12 and 13 show the modularity and nestedness sorting, respectively. We detected that 9/15 modules are statistically modular in at least one of the two null models, whereas 5/15 are modular using both of the null models. In addition, we find that 8/15 of the modules are statistically nested in at least one combination of NTC/NODF vs Bernoulli/Probabilistic degree null models. The fact that 8 of 15 modules are statistically nested in at least one case is an indication that nestedness is present at smaller scales. This supports the hypothesis that modularity may be characteristic at large scales (the scale of the entire network), whereas nestedness may be observed at small scales (at the scale of an individual module) [60]. However, here we note that small-scale structure includes nestedness and modularity.

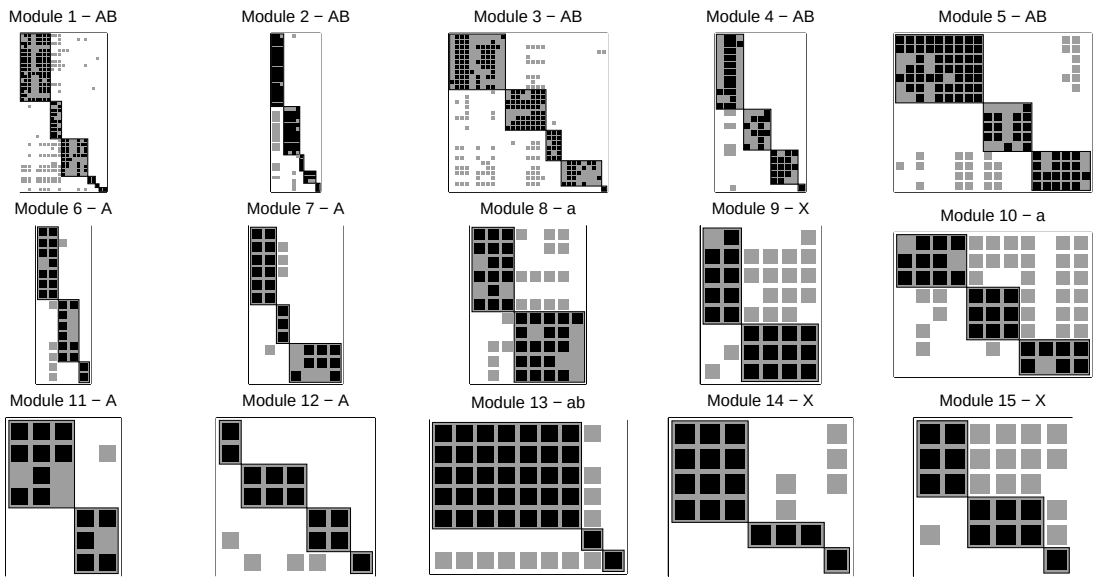

Figure 12: Modular sort of the internal structure of the 15 largest modules, in the same order as they appear in Figure 3. The significance of modularity is denoted as follows: A/a = statistically modular/antimodular using Bernoulli null model, B/b = statistically modular/ antimodular using probabilistic degree null model. X = no significant modular or antimodular.

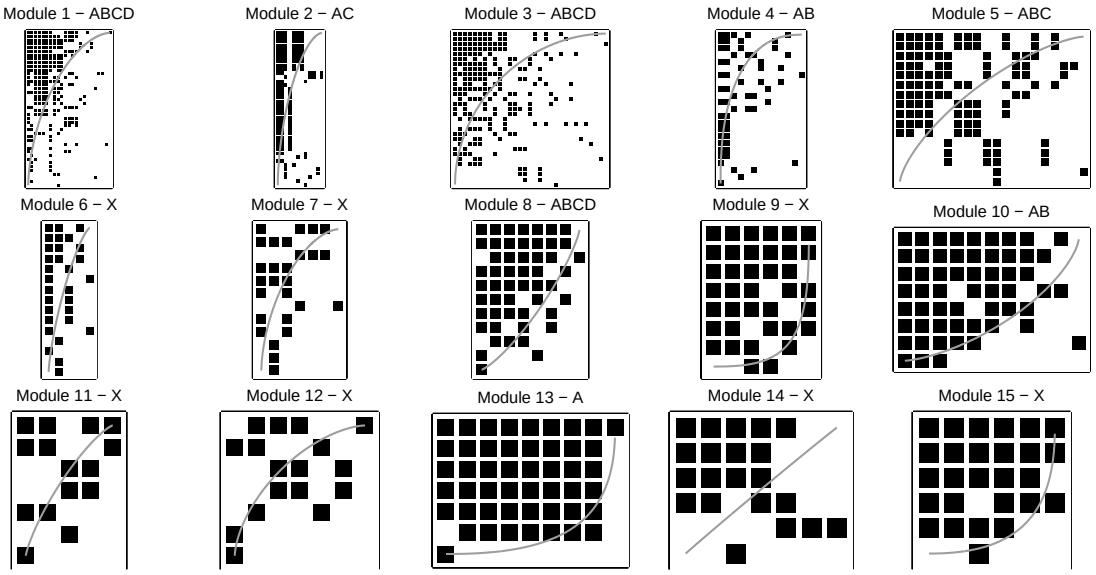

Figure 13: Nestedness sort of the 15 largest modules. The gray line represents the isocline of the NTC algorithm. A/B = statistically nested using NTC and Bernoulli/probabilistic degree null model, C/D = statistically nested using NODF and Bernoulli/ probabilistic degree null model. X = no significance was found.

3.3.5 Geographical diversity of interactions

We find that, on average, there is less geographic diversity in each of the largest 15 modules identified in Figure 11 than would be expected by chance. The result of the geographic diversity test is shown in Figure 14. Specifically for phages, 11 of 15 modules exhibit statistically significant lower diversity than is expected by chance using Simpson diversity, and 12 of 15 modules are found to be statistically significant when using Shannon diversity (see Appendix B, Figure 34 and Appendix B, Table 15). Moreover, the two largest modules have lower geographic diversity of phages than average, but not significantly lower than might be expected by chance. Similar results hold for hosts, where 10 of 15 modules exhibit statistical significant lower diversity using Simpson and 11 of 15 using Shannon diversity (again see Appendix B, Figure 34). These results imply that strains within modules are overrepresented by phages and hosts that belong to the same subset of stations. However, it is important to point out that this data set includes many positive infections (1012 of 1332) of

hosts by phages that were not isolated from the same sample site.

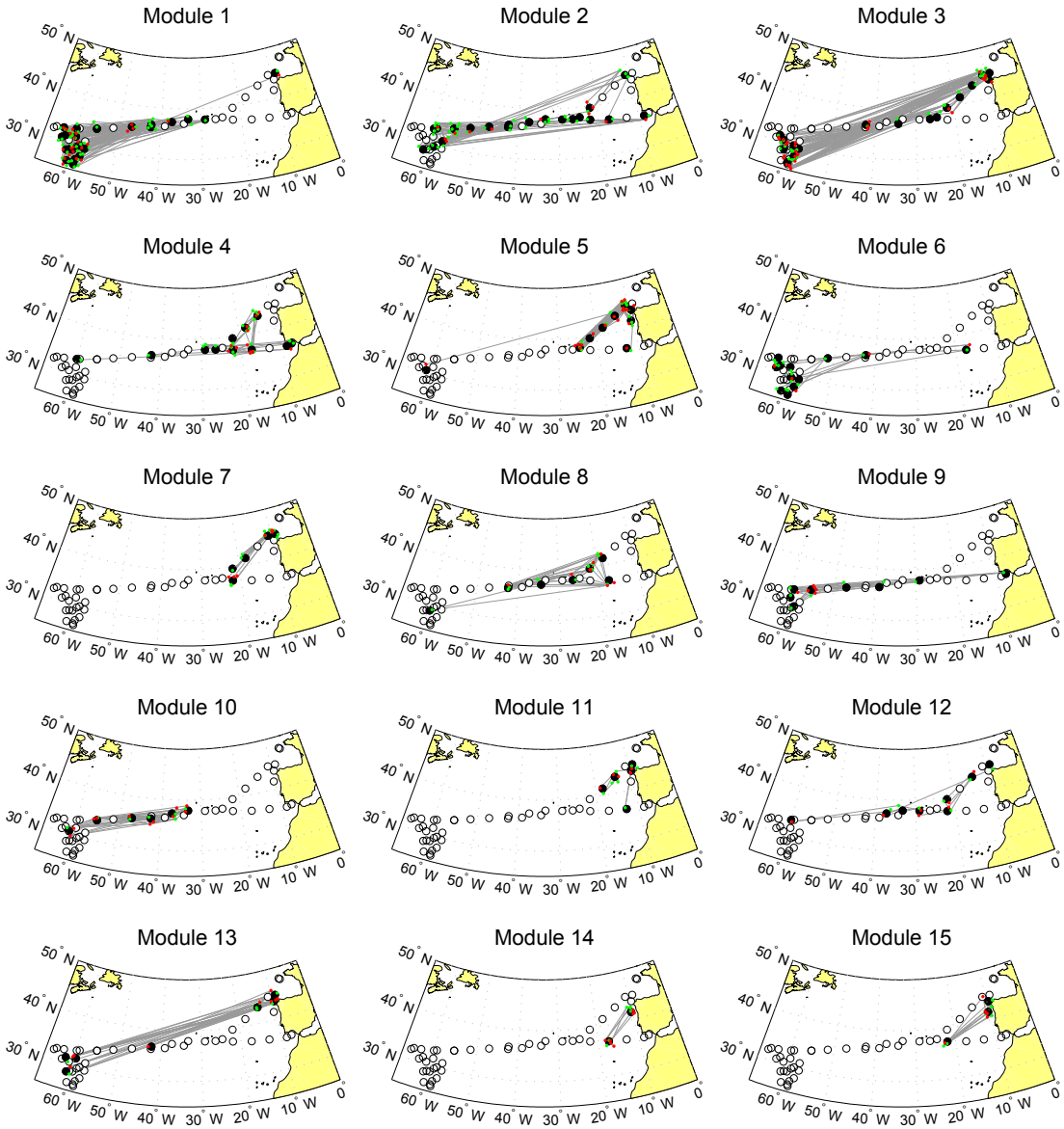

Figure 14: Geographical representation of the 15 largest modules. Each module is considered in a separate panel. Large filled circles represent the stations included in the corresponding module; open circles represent the stations not included in the corresponding module. Red and green small circles representing phages and bacteria, respectively, were randomly placed around their corresponding station for improved visibility. A gray line between a red and green circle denotes an interaction between a virus and bacteria.

To what extent are the interactions between phages and hosts at a given site more likely to occur than those between sites? First, we find that the probability of a

phage infecting and exploiting a host from a different station is lower (0.017) than it is of infecting and exploiting a host from the same station (0.17). This is a 10-fold effect in geographic isolation. We caution that the isolation procedures for phages are heavily biased toward obtaining this effect as phages were isolated from hosts at a given station. As one means to control for this effect, we reduced the number of internal station interactions by the total number of viruses and re-perform this analysis. In doing so, we find a revised probability of 0.061 within modules, which is a 3.6-fold increase when compared with interactions between modules. Finally, in Appendix B, Figure 35, we show that the fraction of shared interactions for both hosts and phages is larger within stations than it is between stations. Altogether these results show geographic location, whether at a given site or among a subset of sites, have an important role in driving infection patterns.

3.4 Discussion

We performed the first multi-scale analysis of a phage-bacteria infection network, comprised of 286 bacteria and 215 phages isolated from the Atlantic Ocean. First, we found that bacteria and viruses were highly variable in their interactions, corresponding to a spectrum of generalist and specialist viruses as well as hard-to-infect to permissive bacteria (Appendix B, Figure 30). Second, we found that the infection network was modular at a large scale and had multi-scale structure such that modules were themselves nested and/or had further modular organization. Network studies have suggested that modularity can be topological, for example, functional modularity as found in proteinprotein interaction networks [142] or transcriptional regulatory networks [90]. Here, a geographic diversity analysis revealed that the modular signal observed was driven, in part, by geographic isolation. However, it is important to point out that cross-infections that transcend site of isolation were common, indeed

approximately 76% of observed interactions occurred between a phage and a bacterium isolated at different sites.We discuss the relevance and implications of each of these results below.

The observation has been made on multiple occasions that the number of hosts a virus can infect can vary substantially, (for example, [124, 189, 42, 86, 118]. Variability in the host range of phages is consistent with the notion that phages have evolved evolutionary strategies ranging from specialists to generalists. Similarly, variability in the number of viruses that can infect a given host is consistent with the notion that hosts have evolved evolutionary strategies ranging from well defended to permissive. It is thought that the relative ecological success of such strategies depends on environmental conditions, for example, bacterial defense specialists may be favored when resources are abundant and competition strategists may be favored when resources are limited [193]. However, such conclusions are often based on models of interaction dynamics, such as Kill-the-Winner [172, 170], that do not include significant cross-infection. Combining cross-infection networks into dynamic models could help develop predictions relating infection structure to community composition [188].

Although we identified generalist viruses, the most generalist virus could infect 31 of the 286 total hosts in the network, suggesting that nestedness at the whole-network scale is unlikely. Indeed, the MN matrix is comprised of disjoint components (Figure 10) of which some of these components exhibit additional modular structure within a component (Figure 11). These modules may themselves have further modularity and/or nestedness (Figures 12 and 13). This is the first instance, of which we are aware, of detection of such multi-scale structure in microbial interaction networks. This result can be interpreted in a number of ways. First, the finding of modules within modules suggests multiple levels of specialization that may be present in the community. Second, the finding of nestedness and modularity are not exclusive. In our prior study [60], we found nearly perfectly nested networks that appear 'modular'

using the standard BRIM metric [13]. This warrants separate examination to develop metrics that can disentangle these two network properties. We developed one such approach here, by suggesting that estimates of nestedness could be performed under modular constraints, and in so doing find that modularity at the scale of the entire MN network and observe nestedness at a local scale (that is, within modules).

What is the biological basis for modules? Given the data available, we evaluate the role of geography in structuring infection. Moebus and Nattkemper (1981) hypothesized, based on visual inspection, that geographic location drove part of the interaction signal. Recent work has suggested that viruses are more likely to infect hosts from the same site than they are hosts isolated at different sites [181, 75, 100]. We found a similar result, in that viruses were at least three times more likely to infect a host isolated from the same location than a host isolated from a different location, even after accounting for isolation bias. However, infection across sample sites was observed frequently, and modules typically contained hosts and phages from multiple sample sites. Using a geographic diversity method, we found that modules tend to have phages and hosts from a much smaller number of sample sites than would be expected by chance. Hence our study is consistent with recent calls for greater attention to spatial structure to viral biogeography [48, 85]. One interpretation of our results is that interactions between phages and host may be endemic despite a consensus that viruses are usually cosmopolitan, that is, they can be observed across a broad range of locations [26, 10]. This may be the case because geographically separated sites are comprised of relatively distinct microbes (for example, microbes differ at the genus level or higher) so that isolated viruses are unlikely to infect the taxa of microbes across sites. Or, it may be that geographically separated sites have relatively similar microbial isolates (for example, communities are dominated by culturable microbes related at the species level or lower) but that their geographic separation facilitated local coevolution to take place, which enabled divergences in functional interactions

[85, 133, 25]

The finding of multi-scale structure also suggests that different processes may drive the emergence of functional interactions at different scales. For example, in the gene-for-gene model of coevolutionary adaptation [4], hosts and phages accumulate differences in defense and counter defense that are consistent with the emergence of nestedness. However, innovations by hosts may also have an important, albeit less frequent, role in permitting hosts to escape from phage infection and selective pressure. Similarly, innovations by phages may also permit them to re-establish access to a host population [117]. A number of evolutionary models of phages and hosts have proposed mechanisms by which coevolutionary dynamics unfold [170, 186, 144, 35]. We suggest that examining resultant phagebacteria interaction networks will be an important means to quantify functional complexity in natural systems and to identify signatures that could discriminate between alternative coevolutionary models.

Ecological patterns depend on the scale of inquiry [107]. In the case of phage-bacteria infection networks, relevant scales may be taxonomic, environmental and/or geographic. Hence, measurements of interaction networks coupled with information on geography, taxa and environmental conditions (for example, [137]) could help disentangle the relative importance of drivers of microbial interactions, in much the same way that biogeographic studies are beginning to quantify the relative importance of drivers of microbial species distributions [112]. Of course, in doing so, new methods to measure cross-infection will be needed. First, our discussion of phage-host interactions in this paper has largely focused on the antagonistic mode. However, the MN matrix includes turbid plaques, which could be interpreted as indicative of infection by temperate phages. Followup studies on the differences and similarities between virulent vs temperate phages in natural environments are worthwhile. Second, it was recently noted that 'the true host range for most marine phages is completely uncharacterized' [25]. Previously published cross-infection assays, including the MN matrix

examined here, use traditional spot-assay or plaque-assay based methods for assessing interactions between cultured bacteria and phages. In moving forward, we suggest that methods to evaluate the functional interaction between hosts and phages that do not rely on cultured isolates [169, 46] will represent an important step to assessing the general structure of interactions in natural communities. We hope that the network approach developed here will be of use in such an effort.

CHAPTER IV

BIMAT : A MATLAB® PACKAGE TO FACILITATE THE ANALYSIS AND VISUALIZATION OF BIPARTITE NETWORKS

Adapted from Cesar O. Flores, Timothée Poisot, Sergi Valverde, and Joshua S. Weitz. BiMAT: a MATLAB (R) package to facilitate the analysis and visualization of bipartite networks. arXiv:1406.6732 [61].

The statistical analysis of the structure of bipartite ecological networks has increased in importance in recent years. Yet, both algorithms and software packages for the analysis of network structure focus on properties of unipartite networks. In response, we describe BiMAT, an object-oriented MATLAB package for the study of the structure of bipartite ecological networks. BiMAT can analyze the structure of networks, including features such as modularity and nestedness, using a selection of widely-adopted algorithms. BiMAT also includes a variety of null models for evaluating the statistical significance of network properties. BiMAT is capable of performing multi-scale analysis of structure - a potential (and under-examined) feature of many biological networks. Finally, BiMAT relies on the graphics capabilities of MATLAB to enable the visualization of the statistical structure of bipartite networks in either matrix or graph layout representations. BiMAT is available as an open-source package at http://ecotheory.biology.gatech.edu/cflores.

4.1 Background

Biological and social systems involve interactions amongst many components. Such systems are increasingly represented as networks, where nodes denote the interacting objects, and the edges denote the interactions between them [126]. Of course, not all networks are alike. For example, networks are often differentiated based on whether or not individual nodes have the same types of incoming and outgoing links. A network is termed unipartite if any node can potentially connect to any other node, as in metabolic networks [93], food webs [40, 52], or friendships/contacts in a social network [184]. The interactions between nodes in such networks are often highly structured, i.e. they differ from idealized networks in which the probability of interacting between any two nodes is constant (i.e. the so-called Erdös-Renyi graph [56]). Evaluating the structure of a unipartite network has spurred the development of concepts such as modularity, small-world structure, and hierarchy [126]. Measuring these structures has in turn, led to efficient implementations of algorithms meant to quantify and characterize network structure, primarily that of unipartite networks [19, 154, 82, 44].

In contrast, a network is termed bipartite if nodes represent two distinct types such that interactions can only occur between nodes of different types [34]. The canonical example of bipartite networks is that of interactions among plant and pollinators, where links represent pollination [52]. Indeed, an abundant literature has emerged on the use of bipartite networks and associated analysis techniques for analysing plant–pollinators systems [17, 162, 18, 20, 95]. However, the concept of bipartite networks (and the specific methodology it carries) can be applied in different domains, including the study of antagonistic networks such as host-parasite interactions [137, 138, 187, 60, 62]. Bipartite networks, like unipartite networks, are rarely random in their structure, i.e. the probability of any potential link between each pair of nodes of different types is not equal. Studies of both plant-pollinator and host-parasite systems

have shown that bipartite networks can be (i) modular, i.e. subsets of nodes often preferentially connect to each other, rather than to other nodes [130]; (ii) nested, i.e. the interaction between nodes can be thought of as subsets of each other [60, 18]; (iii) multi-scale, i.e. the structural properties of the network differ depending on whether the whole or components are considered [62]. As an example, Figure 15 shows Memmot [115] plant-pollinator network, such that the nested and modular structure only becomes apparent when the appropriate sorting is used. Besides the importance of these metrics to quantity the structure of bipartite empirical data, there is still not a self-contained library or software package for analysing the structure of bipartite networks.

In response, we describe `BiMAT` , an open-source software for the analysis of bipartite networks. `BiMAT` is written in `MATLAB`® . Although `MATLAB`® is proprietary software, its use has increased among ecological research groups due to the fact that producing results and plots is easy and quick. The library includes implementations of the most commonly used algorithms for characterizing the extent to which a bipartite network exhibits modular, nested and multi-scale structure. In addition to measuring the structure of a network, `BiMAT` also evaluates the statistical significance of this structure given a suite of null models. Finally, `BiMAT` provides a range of visualization tools for exploring bipartite network structure in either matrix or graph layouts. Here, we first describe the core definitions and methods used in the analysis of bipartite networks. Then, we describe the implementation of `BiMAT` and its application to a number of examples drawn from virus-host interaction data.

4.2 Methods

4.2.1 Bipartite ecological network

A bipartite network, **B**, is a network in which nodes can be divided in two sets R (row nodes) and C (column nodes) such that edges exist only across R and C.

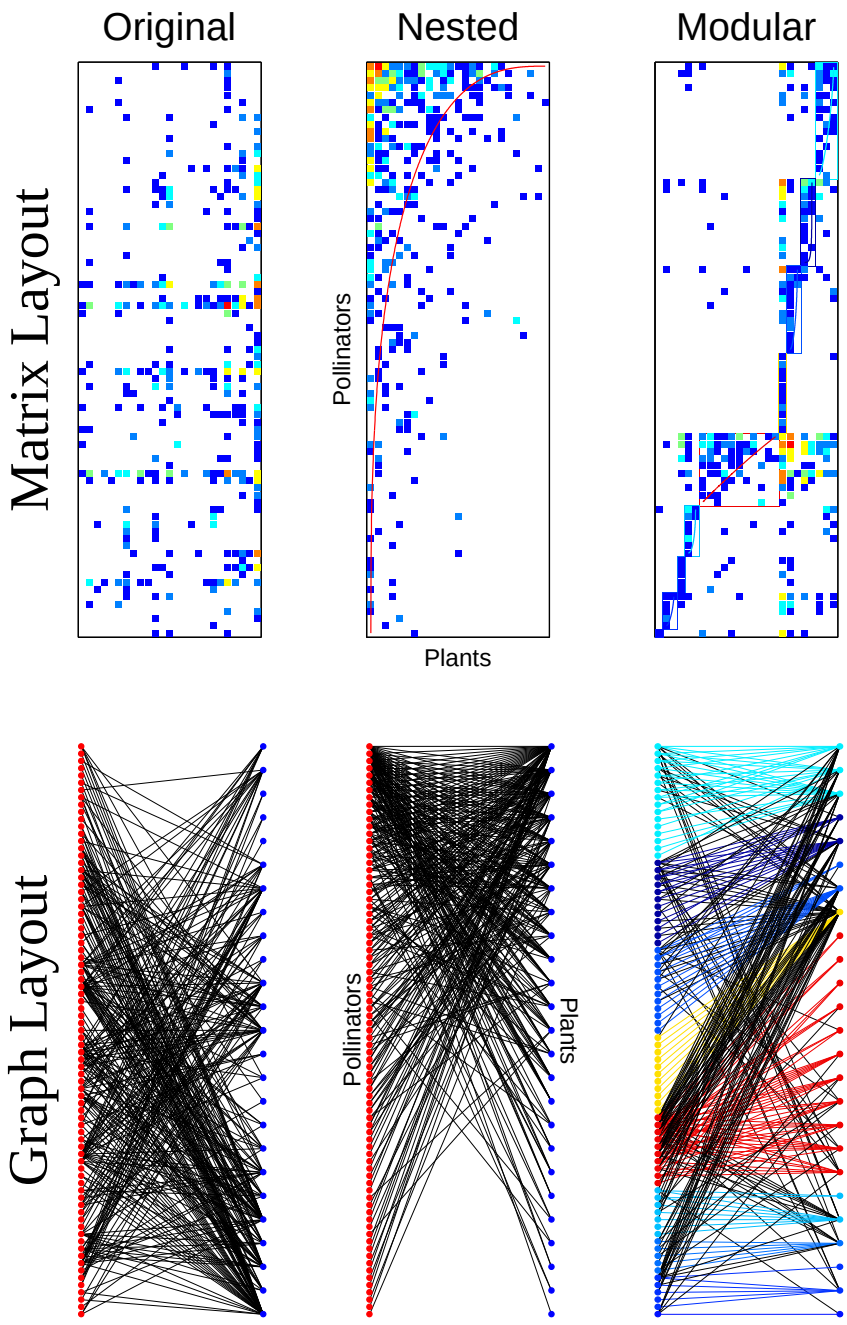

Figure 15: Schematic of an empirical bipartite network (plant-pollinator [115]) in matrix and graph layout using the original, nested and modular sorting of plant and pollinator nodes. Color of cells are frequency of visits mapped to log scale, from small number of visits (darker blue) to large number of visits (dark red). While in the left panels no structure is apparent, the middle and right panels show the opposite. Through visual inspection of the panels, we may infer that the network is nested.

This type of network can be represented as a bipartite adjacency matrix $\mathbf{B}$ of size $m \times n$, where m is the number of nodes in set R and n is the number of nodes in set C. In our implementation, R and C are the node sets that are represented by the rows and columns of the bipartite adjacency matrix in a `Bipartite` object. Although `BiMat` takes quantitative matrices as input, all algorithms implemented in `BiMAT` first threshold these values such that interactions are either present (1) or absent (0). The number of links can be defined as $E = \sum_{ij} B_{ij}$. Finally $k_i = \sum_j B_{ij}$ and $d_j = \sum_i B_{ij}$ define the degree (number of interactions) of the two kinds of nodes.

4.2.2 Algorithms

4.2.2.1 Modularity

`BiMAT` use the standard measure of modularity [128], which for a bipartite network can be defined as (following Barber [13]):

$$Q_b = \frac{1}{E} \sum_{ij} \left(B_{ij} - \frac{k_i d_j}{E} \right) \delta(g_i, h_j), \tag{5}$$

where g_i and h_i are the module indexes of nodes i (that belongs to set R) and j (that belongs to set C). The idea behind the last equation is to maximize Q by choosing the appropriate indexes for vectors $\mathbf{g}$ and $\mathbf{h}$. Significant debate concerns identifying the optimal set of modules in the case of bipartite networks [65, 151]. In order to provide multiple options, `BiMAT` uses three different algorithms to maximize Equation 5: Adaptive BRIM [13], LP-BRIM [108] and the leading eigenvector method [128].

- `AdaptiveBRIM`: The standard BRIM (for Bipartite Recursively Induced Modules) algorithm works in the matricial notation version of Equation 5 given by:

$$Q_b = \frac{1}{E} \operatorname{Tr} \mathbf{R}^T \tilde{\mathbf{B}} \mathbf{T}, \tag{6}$$

  where $\tilde{B}_{ij} = B_{ij} - \frac{k_i d_j}{E}$ is often called the modularity matrix. Further, we replaced the delta function and vectors $\mathbf{g}$ and $\mathbf{h}$ by the $m \times c$ index matrix

$\mathbf{R} = [\mathbf{r}_1|\mathbf{r}_2|...|\mathbf{r}_m]^T$ and the $n \times c$ index matrix $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2|...|\mathbf{t}_n]^T$, for row and column nodes, respectively, with c denoting the number of modules [13]. Notice that nodes cannot be classified into more than one module. Hence, vectors $\mathbf{r}_i$ and $\mathbf{t}_i$ consist of a single one (corresponding to the chosen module) with all the other entries being zero. For example, $r_{ik} = 1$ if the i-th row node belongs to the k-th module with $r_{ij} = 0$ for all $j \neq k$. Using the last expression, the standard BRIM algorithm computes the optimal modularity by inducing the division of one set of nodes (say vector $\mathbf{T}$) from the division in the other set of nodes (say vector $\mathbf{R}$). At each step, BRIM assigns nodes of one type to modules in order to maximize the modularity. BRIM iterates this process until a local maximum is reached. However, the choice of a predefined number c of modules limits the efficacy of the algorithm. Hence, we use an adaptive heuristic [13] to identify the optimal set of modules (and associated modularity Q). This heuristic assumes that there is a smooth relationship between the number of modules c and the modularity $Q_b(c)$. For continuous and smooth landscapes, a simple bisection method ensures that we will find the optimal value of $c = c^\star$ corresponding to maximum Q_b. Starting at $c = 1$ (and modularity $Q_b(1) = 0$ because all nodes belong to the same module) the adaptive BRIM searches for optimal c by repeatedly doubling the number of modules while modularity increases, $Q_b(2c) > Q_b(c)$. At some point, the search crosses a maximum in the modularity landscape, i.e. $Q_b(2c) < Q_b(c)$, and we interpolate the number of modules $c^\star$ to some intermediate value in the interval $(c/2, 2c)$.

- `LP&BRIM`: The algorithm is a combination between the BRIM and LP (Label Propagation) algorithms. The heuristic of this algorithm consists in searching for the best module configuration by first using the LP algorithm. This algorithm initially assigns each node to a different module (label). At each interaction the module of each node is reassigned to the module to which the

majority of its neighbours belong to. The order of node reassignment is random and ties are broken randomly. The algorithm continues until convergence is achieved. The standard BRIM algorithm is used at the end to refine the results.

- `LeadingEigenvector`: This algorithm works with the unipartite adjacency matrix $\mathbf{A}$ of size $m+n \times m+n$ instead of the bipartite adjacency matrix $\mathbf{B}$. The modularity using this notation can be defined for two modules in matrix notation as [128]:

$$Q = \frac{1}{4E}\mathbf{s}^T \tilde{\mathbf{A}} s, \tag{7}$$

where $\tilde{A}_{ij} = A_{ij} - \frac{k_i k_j}{2E}$ is the modularity matrix expressed using the unipartite adjacency matrix with no distinction for degrees or rows and columns. Further, for a particular division of the network into two modules, $s_i = 1$ if node i belongs to module 1 and $s_i = -1$ if it belongs to module 2. The idea of this algorithm is that we can decompose the previous equation in a linear combination of the normalized eigenvectors u_i of $\tilde{\mathbf{A}}$ so that $\mathbf{s} = \sum a_i \mathbf{u}_i$ with $a_i = \mathbf{u}_i^T \cdot \mathbf{s}$:

$$Q = \frac{1}{4E}\sum a_i \mathbf{u}_i^T \tilde{\mathbf{A}} \sum a_j \mathbf{u}_j = \frac{1}{4E}\sum (\mathbf{u}_i^T \cdot \mathbf{s})^2 \alpha_i, \tag{8}$$

where α_i is the eigenvalue of $\tilde{\mathbf{A}}$ corresponding to eigenvector $\mathbf{u}_i$. The leading eigenvector name comes from the fact that in order to maximize the last equation what we can do is to focus only in the sum term with the maximum eigenvalue α_{max} which corresponds the leading eigenvector $\mathbf{u}_{max}$. This term can be maximized by trying to maximize $\mathbf{u}_{max} \cdot \mathbf{s}$. Because s_i can only have the values ± 1, this can be solved by assigning $s_i = 1$ and $s_i = -1$ when $\mathbf{u}_{max_i} > 0$ and $\mathbf{u}_{max_i} \leq 0$, respectively, which completes the core of the leading eigenvector algorithm. After performing the first iteration of the last process we will have a subdivision of just two modules. Newman [128] then explain that this process

can be applied recursively in each of the subdivisions. However, instead of isolating each subdivision of each other, we apply this heuristic in the expression ΔQ which defines the change of modularity that a new subdivision in an specific module will give us. The subdivision is only accepted if $\Delta Q > 0$. For mode details about ΔQ we recommend to read [128]. Finally, it is worth to mention that in `BiMAT` by default each subdivision is refined using the Kernighan–Lin algorithm [99] too. The essence of this algorithm is swapping nodes between the two modules such that at each step the node that gives the biggest increase in Q or the smallest decrease (if increase is not possible) is swapped. In a complete iteration all nodes are swapped with the constraint that a node is swapped only once. The intermediate state during the iteration that has the biggest Q is selected as the new configuration and the process repeats using this new configuration until no improvement is possible.

In addition to optimize the standard modularity Q_b `BiMAT` also evaluates (after optimizing Q_b) an a posteriori measure of modularity Q_r introduced in [136] and defined as:

$$Q_r = 2 \times \frac{W}{E} - 1 \tag{9}$$

where $W = \sum_{ij} B_{ij}\delta(g_i, h_j)$ is the number of edges that are inside modules. Alternatively, $Q_r \equiv \frac{W-T}{W+T}$ where T is the number of edges that are between modules. In other words, this quantity maps the relative difference of edges that are within modules to those between modules on a scale from 1 (all edges are within modules) to -1 (all edges are between modules). This measure allows to compare the output of different algorithms.

4.2.2.2 Nestedness

Nestedness is a term used to describe the extent to which interactions form ordered subsets of each other. Multiple indices are available to quantify nestedness (see [178]

for details about many of these measures). Two of the most commonly used methods are: NTC (Nestedness Temperature Calculator) [11, 143] and NODF (for Nestedness metric based on Overlap and Decreasing Fill) [7]. Both of these are implemented in `BiMAT` and are summarized below:

- `NestednessNTC` (NTC): A 'temperature', T, of the interaction matrix is estimated by resorting rows and columns such that the largest quantity of interactions falls above the isocline (a curve that will divide the interaction from the non-interaction zone of a perfectly nested matrix of the same size and connectance). In doing so, the value of T quantifies the extent to which interactions only take place in the upper left ($T \approx 0$), or are equally distributed between the upper left and the lower right ($T \approx 100$). Perfectly nested interaction matrices can be resorted to lie exclusively in the upper left portion and hence have a temperature of 0. The value of temperature depends on the size, connectance and structure of the network. Because the temperature value quantifies departures from perfect nestedness, we define the nestedness, N_{NTC}, of a matrix to range from 0 to 1, $N_{NTC} = (100 - T)/100$, such that $N_{NTC} = 1$ when $T = 0$ (perfect nestedness) and $N_{NTC} = 0$ when $T = 100$ (checkerboard).

- `NestednessNODF`: NODF is independent of row and column order. This algorithm measures the nestedness across rows by assigning a value M_{ij}^{rows} to each pair i, j of rows in the interaction matrix[7]:

$$M_{ij}^{\text{rows}} = \begin{cases} 0 & \text{if } k_i = k_j \\ n_{ij}/\min(k_i, k_j) & \text{otherwise} \end{cases} \tag{10}$$

  where n_{ij} is the number of common interactions between them. A similar term is used for the column contributions, such that the total nestedness is defined as:

$$N_{NODF} = \frac{\sum_{i<j} M_{ij}^{\text{rows}} + \sum_{i<j} M_{ij}^{\text{columns}}}{m(m-1)/2 + n(n-1)/2}. \tag{11}$$

However, `BiMAT` redefined Equation 10 (and its column version), such that the last equation can be more easily vectorized:

$$M_{ij}^{\text{rows}} = \frac{(\mathbf{r}_i \cdot \mathbf{r}_j)\delta(k_i, k_j)}{\min(k_i, k_j)}, \tag{12}$$

where $\mathbf{r_i}$ is a vector that represents the row i of the bipartite adjacency matrix. Equation 11 can be rewritten in terms of adjacency matrix multiplications (see code for details). This new vectorized version of calculating the N_{NODF} value outperforms the naive one (using loops) by a factor over 50 in most of the matrices that we tested.

Note that a new eigenvalue-eigenvector approach to evaluating nestedness has recently been introduced [160], which will be introduced in a future `BiMAT` release.

4.2.3 Statistics

4.2.3.1 Null Models

We propose four null models to test the significance of measured nestedness and modularity (see [18, 179, 60, 138] for more details). These null models generate random networks through a Bernoulli process, where the probability of interactions are determined following different rules. Define k_i as the degree of a node i of the column class and d_j as the degree of a node j of the row class. Then, the probability that two nodes (of distinct classes) interact, P_{ij} is:

EQUIPROBABLE , $P_{ij} = E/(mn)$ – the connectance of the network is respected, but not the number of interactions in which each node is involved.

AVERAGE , $P_{ij} = (k_i/n + d_j/m)/2$ – the connectance, and the expected number of interactions in which each node is involved, are respected

COLUMNS , $P_{ij} = k_i/n$ – the connectance, and the expected number of interactions of row nodes, are respected

ROWS , $P_{ij} = d_j/m$ – the connectance, and the expected number of interactions of column nodes, are respected

By default, `BiMAT` generate networks that can have disconnected nodes (i.e. nodes with no edges to any other nodes in the network). However the user can impose a constraint that all nodes must be connected to at least one other node (if possible) in the null model generating process. Note that `BiMAT` does not include some of the most constrained null models, *e.g.*, random networks that respect not only the expectation of connectance and degree but also the *exact* degree sequences as the original network [160],

4.2.3.2 Statistic Values

Once an ensemble of random networks is specified, `BiMAT` will return the following values:

- `value`: value to be tested (*e.g.* nestedness or modularity).
- `random_values`: the values of all random replicates.
- `replicates`: number of replicates used during testing.
- `mean`: mean of the replicate values.
- `std`: standard deviation of the replicate values (note that distributions of network values are not necessarily well described by a normal distribution).
- `zscore`: The z-score of `value` assuming that the replicate values represent the entire population.
- `percentile`: The percentage of replicate values that are smaller than `value`.

4.2.3.3 Extended statistics

As described above, `BiMAT` enables the evaluation of the statistical significance of modularity and nestedness. Additional statistical evaluation is possible, including the capability to conduct a meta-analysis and a multi-scale analysis.

Meta analysis : `BiMAT` can simultaneously analyse the network structure of a set of related bipartite networks (*e.g.* plant-pollinator networks or virus-host interaction networks). In which case, the distribution of network properties of the set of networks can be analysed (see example I in the Examples section for more details).

Multi-scale analysis : Individual modules need not always be homogeneous. Hence, `BiMAT` offers functionality to evaluate whether or not the network has different structures at different scales, *e.g.*, the overall network may be modular, but individual modules may be nested (see example II in the Examples section for more details).

4.3 The `BiMAT` package

`BiMAT` is a open-source package (see Figure 16) written in `MATLAB`® . It is primarily designed for the analysis and visualization of bipartite ecological networks, thought it may be used for any type of bipartite networks. The package aims to consolidate some of the most popular algorithms and metrics for the analysis of bipartite ecological networks in the same software environment. Specifically, the core features examined are bipartite modularity [13] and nestedness [11, 7]. Further, `BiMAT` include the necessary tools for analysing the statistical significance of these values, together with tools for visualizing bipartite networks in such a way that these properties becomes apparent to the user. `BiMAT` utilizes an object-oriented framework which enables users to extend the package.

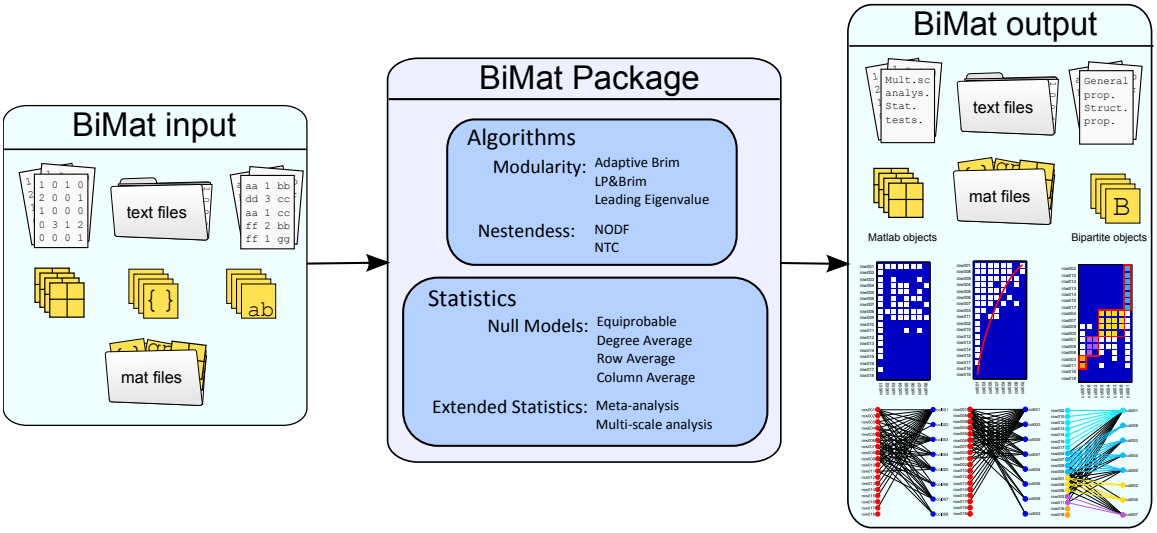

Figure 16: `BiMAT` Workflow. The figure shows the main scheme of the `BiMAT` package. `BiMAT` can take matlab objects or text files as main input. The input is analysed mainly around modularity and nestedness using a variety of null models. The user may also perform an additional multi-scale analysis on the data, or if he have more than one matrix to perform a meta-analysis in the entire data. Finally, the user can observe the results via matlab objects, text files and plots.

4.3.1 Usability

Users are expected to be familiar with the `MATLAB`® environment. However, `BiMAT` has been designed so that even `MATLAB`® beginners or those with very limited expertise can easily carry out a comprehensive analysis and visualization of their data, in many cases with a single command. Despite an emphasis on simplicity, `BiMAT` still retains all of the functionality and flexibility provided by the `MATLAB`® environment (*e.g.*, all the results are returned to the current session workspace, the results can be stored in `MATLAB`® files, and the class properties can be used for `MATLAB`® plotting). A complete start guide is distributed with the library.

4.3.2 Comparison with other software

Current and popular available tools for the analysis of complex networks include implementations that are predominantly: (i) visually oriented (*e.g.* Gephi [19], Cytoscape [154]) or (ii) library-package oriented (*e.g.* networkx [82], iGraph [44]). Unfortunately, these tools have a strong focus on the analysis of unipartite networks, i.e. bipartite networks are treated as a special case of a unipartite network. As a consequence, algorithms for the analysis of unipartite networks, when applied to bipartite networks, are not intended to be optimal, neither where designed to the study of ecological bipartite networks. In contrast, specialized tools for the analysis of bipartite ecological networks are available but they are very specific (*e.g.* ANINHADO [79], WINE [69], and recently FALCON [158] focus only in nestedness analysis).

However, the authors acknowledge the existence of `bipartite` [50], a software library written in R. Thought this library initially included only nestedness analysis regarding internal network structure, they just recently aggregated modularity analysis too [51]. `BiMAT` does not intent to replace this library but to complemented by bringing similar tools to the `MATLAB`® ecology community. Further, `BiMAT` also includes tools for the analysis of many related networks (meta analysis) and for the analysis of different levels of the same network (multi-scale analysis), which will facilitate the statistical analysis of bipartite ecological networks. Whereas `bipartite` strives for exhaustivity, `BiMAT` focuses on implementing a well-documented core of statistical procedures in an optimized way.

In summary, `BiMAT` provides a broad selection of tools required for the analysis and visualization of bipartite ecological networks. As such, `BiMAT` is aimed towards empiricists seeking to apply a network perspective to their data, and is particularly suited to exploratory analyses of data derived from ecological, evolutionary, and environmental datasets. Table 4 show the current tools of current libraries.

Table 4: Bipartite Ecological libraries

Software	Language	Open Source	Visualization	Nestedness	Modularity
ANINHADO [79]	Executable	✗	✗	✓	✗
WINE [69]	MATLAB® /R/C++	✓	✓	✓	✗
FALCON [158]	MATLAB® /R	✓	✓	✓	✗
bipartite [50]	R	✓	✓	✓	✓
BiMAT	MATLAB®	✓	✓	✓	✓

4.3.3 Installation

BiMAT stable version can be downloaded directly from the main author webpage: http://ecotheory.biology.gatech.edu/cflores. Last updated version can be downloaded from https://github.com/cesar7f/BiMat.

4.3.4 License and bug tracking

The software is distributed using FreeBSD license, which basically means that the user can redistribute it, with or without modification for any kind of purpose as long as its copyright notices and the licence's disclaimers of warranty are maintained. Though the license do not force users to do so, we encourage them to cite this paper if the use of BiMAT library leads to any kind of scientific publication.

Users can report bugs directly in the github repository (see URL above), provided they have a github account.

4.3.5 Configuration

The BiMAT directory should be added to the MATLAB® paths. At this point, BiMAT can be executed without any additional configuration. The default parameters for algorithms implemented in the BiMAT package are available in file main/Options.m. Additional details are available in the Start Guide, including as part of the BiMAT package (and released here as Supplementary File X).

4.3.6 Objected-Oriented Programming Scheme

`BiMAT` has been coded using the Objected-Oriented Programming (OOP) paradigm. Note that understanding of OOP is not required for use of `BiMAT` . Nonetheless, the use of OOP is meant facilitate maintainability and extensibility of the codebase. Access to `BiMAT` functions is granted (with the exception of some static classes) using instances of the class that implements the functions.

The main package class is the `Bipartite` class, whose only function is to work as a common interface to all of the available statistical, algorithmic, plotting, and input/output classes. Because of this OOP design pattern, most of the `MATLAB`® functionality will be granted using the following syntax:

```
bip.class_instance_in_bip.method_name(arguments)
```

where `bip` is a `bipartite` instance created by the user, `class_instance_in_bip` is a property of the `bipartite` class which represents an instance of the class which has access to the method `method_name`. The method that is called will frequently have direct read and writeable access to other properties inside `bip`. Table 5 shows the main calls from the Bipartite object, assuming that the user call its bipartite instance `bip`.

Note that the OOP capabilities of `MATLAB`® are not as extensive as those of OOP focus languages (*e.g.* python, Java, C++). As such, certain behaviours have been emulated in `BiMAT` , *e.g.* static classes were emulated using private constructors. However, in contrast to other languages that enable OOP, `MATLAB`® enables users to store created instances as `MATLAB`® objects in files. This ensures that users can save, and subsequently load, the results of partial analysis.

4.3.7 Input/Output

The class `bipartite` is the main class of the package. Hence, a user will usually need to work with at least one instance of this class. An instance of this class requires a

Table 5: Some useful calls using the OOP approach

Call	Class	Description
bip.community.Detect()	BipartiteModularity	Calculate Modularity
bip.nestedness.Detect()	Nestedness	Calculate nestedness
bip.statistics.DoCompleteAnalysis()	StatisticalTest	Executes the required comm a complete analysis of nest
bip.statistics.DoNulls()	StatisticalTest	Create the null model matr
bip.statistics.TestCommunityStructure()	StatisticalTest	Perform the statistical test
bip.statistics.TestNestedness()	StatisticalTest	Perform the statistical test
bip.internal_statistics.TestDiversityRows()	InternalStatistics	Perform diversity analysis a
bip.internal_statistics.TestDiversityColumns()	InternalStatistics	Perform diversity analysis a
bip.internal_statistics.TestInternalModules()	InternalStatistics	Perform an statistical test for modularity and nestedn
bip.plotter.PlotMatrix()	PlotWebs	Plot a matrix layout of the
bip.plotter.PlotModularMatrix()	PlotWebs	Plot a matrix layout of the
bip.plotter.PlotNestedMatrix()	PlotWebs	Plot a matrix layout of the
bip.plotter.PlotGraph()	PlotWebs	Plot a graph layout of the
bip.plotter.PlotModularGraph()	PlotWebs	Plot a graph layout of the
bip.plotter.PlotNestedGraph()	PlotWebs	Plot a graph layout of the

boolean `MATLAB`® matrix object, representing the bipartite adjacency network. Alternatively, a `integer` matrix can be provided e.g., when the values represent categorical levels of interactions, and these categorical levels can be included in the visualization tools. Optional arguments that can be passed are the row and column node labels and classification classes. These arguments need to be passed directly to the properties of the `Bipartite` object. In practice, an object of the class `Bipartite` can be created as follows:

```
bip = Bipartite(matrix);
bip.row_labels = rowLabels;
bip.col_labels = colLabels;
bip.row_class = rowClasses;
bip.col_class = colClasses;
```

in which the variables `matrix, rowLabels, colLabels, rowClasses` and `colClasses` are previously defined variables. Network information, including adjacency matrix and node labels, can be read directory from data files using the static class `Reading`:

- `bip = Reader.READ_BIPARTITE_MATRIX(filename)`: The file should be in the following format:

  ```
  1 0 0 2 0 0 0
  1 2 0 0 0 2 1
  1 1 0 0 1 2 1
  1 2 3 0 0 1 1
  2 1 1 1 0 0 0
  ```

  Each row in the file represents a different outgoing set of interaction from a node (in set A) to a different set of nodes (in set B) in the columns. All values different from 0 are counted as interactions, such that evaluation of network structure utilizes Boolean information whereas visualization can leverage the non-negative strengths of interactions:

- `bip = Reader.READ_ADJACENCY_LIST(filename)`: The file should be an ordered list of triples:

  ```
  row_label_1 1 col_label_1
  row_label_1 1 col_label_2
  row_label_1 2 col_label_3
  row_label_3 1 col_label_2
  row_label_3 3 col_label_1
  row_label_2 3 col_label_2
  ```

  such that the first and third columns represent nodes from sets A and B, respectively, and (an optional) second column denoting the strength of interactions.

Table 6: Useful calls in the functional approach

Call	Description
BipartiteModularity.ADAPTIVE_BRIM(matrix)	Calculate the modularity values usin
BipartiteModularity.LP_BRIM(matrix)	Calculate the modularity values usin
BipartiteModularity.LEADING_EIGENVECTOR(matrix)	Calculate the modularity values usin
Nestedness.NODF(matrix)	Calculate the NODF values
Nestedness.NTC(matrix)	Calculate the NTC values
PlotWebs.PLOT_MATRIX(matrix)	Plot the data in matrix layout
PlotWebs.PLOT_NESTED_MATRIX(matrix)	Plot the nested sorted data in matrix
PlotWebs.PLOT_MODULAR_MATRIX(matrix)	Plot the modular sorted data in mat
PlotWebs.PLOT_GRAPH(matrix)	Plot the data in graph layout
PlotWebs.PLOT_NESTED_GRAPH(matrix)	Plot the graph sorted data in matrix
PlotWebs.PLOT_MODULAR_GRAPH(matrix)	Plot the graph sorted data in matrix
Printer.PRINT_GENERAL_PROPERTIES(matrix)	Print to screen the general properties
Printer.PRINT_STRUCTURE_VALUES(matrix)	Print the modularity and nestedness

4.3.8 Functional alternative

Static functions can be used as an alternative to interacting with the `BiMAT` package in an OOP framework. For example, the network can be visualized in a graph or matrix layout as follows:

```
PlotWebs.PLOT_MATRIX(matrix);
PlotWebs.PLOT_GRAPH(matrix);
```

instead of:

```
bp = Bipartite(matrix);
bp.plotter.PlotMatrix();
bp.plotter.PlotGraph();
```

Table 6 shows some of the most important static functions that provide access to part of the `BiMAT` functionality.

4.3.9 Plotting

The class `PlotWebs` provides the required functions to visualize a bipartite network in a matrix or graph layout. Visualization can utilize (i) the original sorted version of the data, (ii) the nested sorted version of the data, or and (iii) the modular sorted version of the data. `BiMAT` represents the interaction data with colored cells when a matrix layout is used. Rows and columns denote members of the the two sets and cells denote interaction strength. The format of the matrix is specified by modifying the `PlotWebs` class properties before calling the plotting functions. Further, the format of the matrix will depend on what kind of sorting is used. For example, the modular sorting plot can color the cells according to the module to which they belong to or the type of interaction. Some features are restricted to particular sortings, *e.g.*, plotting an isocline (see Methods) is available only in the nested and modular sorting. Alternatively, the `PlotWebs` can plot the data in a graph layout in which members of the two sets A and B are depicted using a stacked set of circles to the left and right, respectively. Lines are draw between sets that interact. As for matrices, many of the properties of `PlotWebs` can be used to specify the format of the plot (see documentation).

In addition to this main class, `BiMAT` has an additional plot class called `MetaStatisticsPlotter` that is used for plotting meta-analysis results (analysis of many networks). This class can plot statistical results of the structural quantities of the algorithms, together with visual graph or matrix layout representations of the networks (see Examples section).

4.3.10 Performance

The `BiMAT` packages leverages optimization tools of `MATLAB`® . For example, algorithms implemented in `BiMAT` were vectorized to improve performance. In addition, a version of `BiMAT` that uses the `MATLAB`® Parallel Computing Toolbox can be requested to the corresponding author.

4.4 Examples

We present here two examples to illustrate the potential use of `BiMAT` for visualization and analysis of bipartite complex networks: (i) a meta-analysis of 38 different phage-bacteria interaction networks; (ii) a multi-scale analysis of the largest phage-bacteria interaction network. Scripts and data for these examples are included in the `BiMAT` release and additional documentation is included in the start guide.

4.4.1 Example I: Meta-analysis

The study of virus-host interactions includes examination of whom infects whom. Exhaustive assays of cross-infection of a set of phages (viruses that infect and kill bacteria) and a set of bacteria are generally reported as a bipartite cross-infection matrix. These matrices can be standardized such that rows and columns represent bacteria and phages, respectively. The cell enrties in these matrices represent the level of infection between phages and bacteria. In a previous study, Flores et al [60] re-examined 38 such networks extracted from the published literature between 1950 and 2011. In doing so, the authors found that phage-bacteria infection networks (as published) tend to be nested and not modular. `BiMAT` can reproduce these results using the `MetaStatistics` module.

First, the user should begin by creating an instance of the `MetaStatistics` class. This class takes, as input, a cell array `matrices` containing either a set of `MATLAB`® matrices or a set of `Bipartite` objects. An automatic meta-analysis, using default parameters, can be performed by the commands:

```
mstat = MetaStatistics(matrices);
mstat.names = matrix_names %Labels for networks
%chosing the algorithms:
mstat.modularity_algorithm = @AdaptiveBrim
mstat.nestedness_algorithm = @NestednessNTC
```

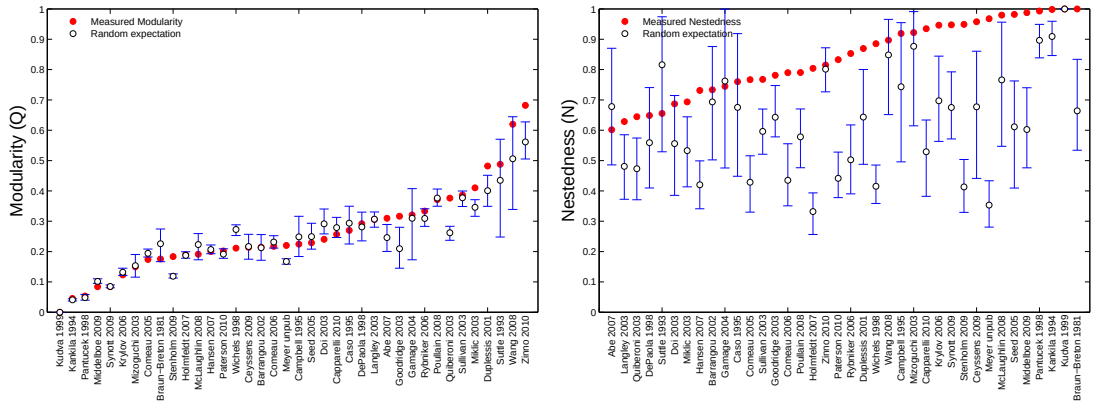

Figure 17: Visual representation of the statistical tests in the set of matrices. Red circles represent the value of the analyzed networks. White circles represent the mean of the null model, while the error bars represent the networks that falls inside a two-tailed version of the random null model values. The margin of the error bars are $(p,1-p)$, where p is the p-value that is an optional argument of the plot functions.

```
mstat.DoMetaAnalyisis();
```

Results of the meta-analysis are stored in the object `gstat`, for subsequent examination. The meta-analysis class (MetaStatistics.m) also has additional plot functions. *e.g.*, to compare network structures against a null model values:

```
mstat.plotter.PlotModularValues(0.05);
mstat.plotter.PlotNestednessValues(0.05);
```

where the argument represent the p-value threshold in determining the variation about the network statistics generated from the ensemble (lower values denote wider variation). The output for the modular and NTC values can be observed in Figure 17. As is apparent, the majority of studies have modularity *below* that of the networks in the random ensemble. In contrast, the majority of studies have nestedness *significantly above* that of the networks in the random ensemble.

In addition, it is possible to plot all the matrices at once using any of the next functions:

```
%Grid of 5 x 8
```

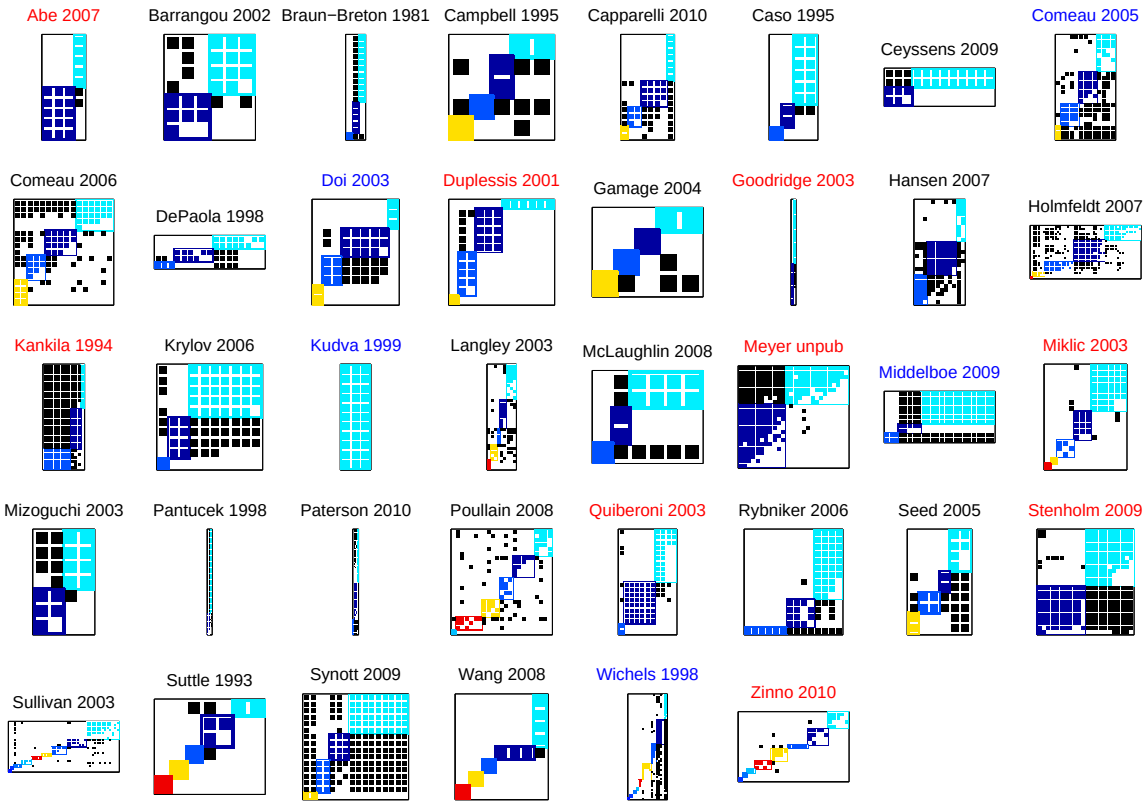

Figure 18: The meta-set collected on Flores et al [60] plotted using the modularity algorithm of the `BiMAT` library. Red and blue labels represent significant modularity ($p \geq 0.975$) and anti-modularity ($p \leq 0.275$), respectively. For bibliographic information about these matrices see [60].

```
mstat.plotter.PlotMatrices(5,8);

mstat.plotter.PlotNestedMatrices(5,8,0.05);

mstat.plotter.PlotModularMatrices(5,8,0.05);
```

where the first and second arguments are the number the matrices along horizontal and vertical axis of the plot. If the statistical test have been already performed, red and blue labels are used for indicate the statistical significance of the corresponding structure (red for significance, and blue for anti-significance), where the third argument (optional) is used to assess a critical p-value the significance. Figure 18 shows the plot for the case of modularity.

4.4.2 Example II: Multi-scale analysis

Moebus and Nattkemper [124] published the largest phage-bacteria infection network. The individual phage and bacteria were extracted from different locations across the Atlantic Ocean. In a previous study we developed a multi-scale analysis of network structure in this dataset[62]. Here, we demonstrate how such a multi-scale analysis can be automated. The first objective is to analyze the global-scale structure of a bipartite network, i.e. to quantify if the overall network has significantly elevated or diminished modularity and/or nestedness. Assuming that our matrix is called `moebus.weight_matrix` left panel of Figure 29 shows a visual representation of this data in matrix layout after typing:

```
bp = Bipartite(moebus.weight_matrix);
bp.community.Detect();
bp.plotter.font_size = 2.0;
figure(1);
bp.plotter.PlotModularMatrix();
```

It becomes apparent that the network is modular. However, what is really important to observe is that internal nodes seems to have nested structure (triangular pattern with most of the links above the isocline). Hence, the Moebus network may have multi-scale structure properties. We will confirm that this is the case for nestedness using the N_{NTC} values. In order to perform this test `BiMAT` make use of the `InternalStatistics` class in order to get the statistics of those modules by isolating them and treating them as independent networks:

```
%We are interested in only the first 15 modules
%from the most righ-top one.
bp.internal_statistics.idx_to_focus_on = 1:15;
%Perform a default internal analysis
```

```
bp.internal_statistics.TestInternalModules();
figure(2);
bp.internal_statistics.meta_statistics...
             .plotter.PlotNestednessValues();
figure(3);
bp.internal_statistics.meta_statistics...
             .plotter.PlotNestedMatrices();
```

where the last two plots are the ones on the right panels of Figure 29. The smart reader may already notice that `meta_statistics` property is in fact an instance of the class `MetaStatistics`, which translates to be able to use any of the methods inside `MetaStatistics` (including its property `plotter`) in the internal modules.

Finally, another multi-scale analysis that `BiMAT` can perform is to quantify if a relation exist between node classification and module distribution. If the extreme case, if this relation exist nodes inside the same module will share the same classification. If the such relationship does not exist, modules will have nodes with random classification. In order words, the relationship depends in how random is the node classification inside the each module. In order to perform this analysis `BiMAT` make use of both Shannon's and Simpon's indexes. And, for evaluating the significance we use a null model in which we randomly swap all node classifications. We will give here a simple example about how to print the significance of Simpson's index for the case of phage (column) nodes. In order to do so, we will use geographical location extraction as classification identifier of each node:

```
% We want to use geographical location
% as classification
bp.col_class = moebus.phage_stations;
% Perform the analysis
```

```
bp.internal_statistics.TestDiversityColumns();
% Print results
bp.printer.PrintColumnModuleDiversity();
```

The user must be able to visualize an output similar to:

```
Diversity index:       Diversity.SIMPSON_INDEX
Random permutations:                        100
Module,index value,  zscore,percent
     1,    0.94805, -1.3848,      6
     2,    0.91738, -5.0054,      0
     3,    0.95238,-0.42625,     11
     4,    0.81667,-13.1025,      0
     5,          1, 0.36742,     12
     6,    0.85714, -2.5808,      0
     7,    0.66667, -2.4661,      0
     8,    0.33333,-13.5825,      0
     9,    0.90909, -2.0933,      3
    10,        0.9, -1.1203,      2
    11,        0.5, -6.6773,      0
    12,    0.88889, -2.6493,      1
    13,        0.6, -7.0097,      0
    14,        0.6, -8.0336,      0
    15,    0.83333, -1.3318,      3
```

If we want to use the percentile as statistical test (using one-tail) and p-value=0.5 we have that 12 modules are not as diverse as the random expectation. Hence, these modules contain phages that come from similar geographical stations, which translate to potentially have a relationship between the geographical location and

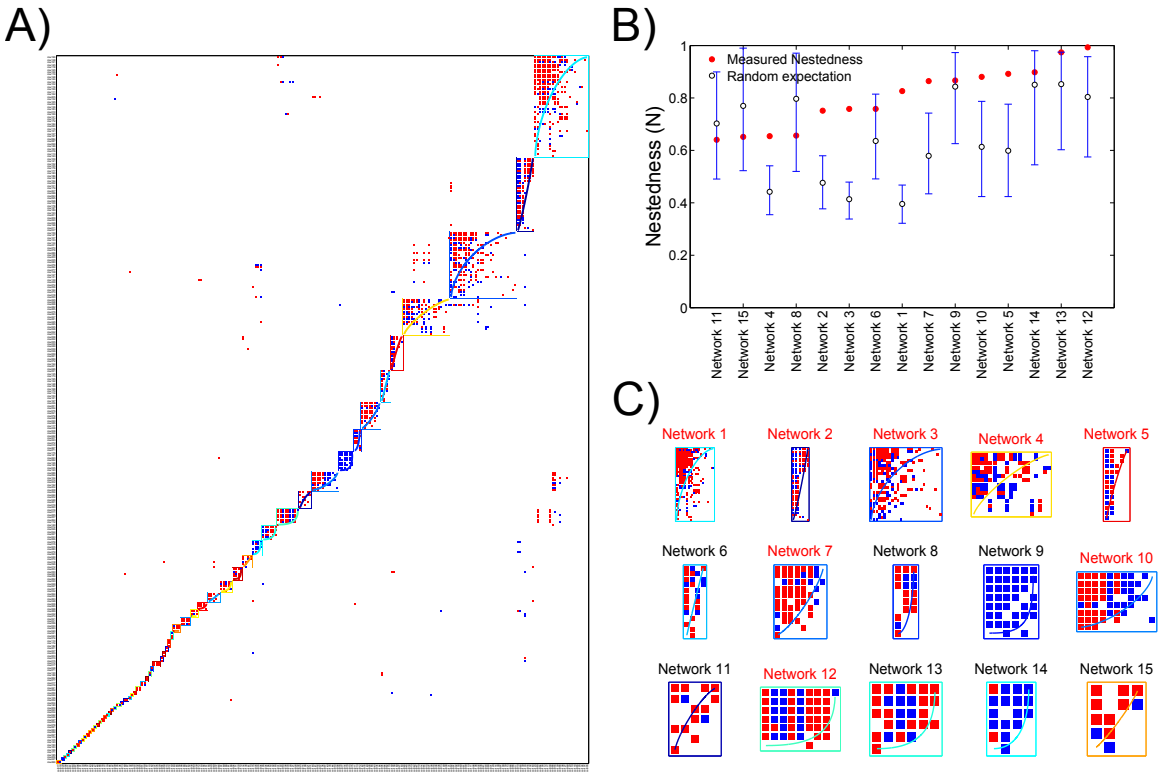

Figure 19: Standard plots that can be extracted using the multi-scale analysis capabilities of `BiMAT` . Here, we focus in the internal nested structure using N_{NTC} values, but we can also perform an internal study using Q_b and N_{NODF} values. **A)** The standard output using the modular matrix layout gives us a hint about the potential multi-scale structure. **B)** Here we focus on the study of N_{NTC} values with respect to random expectation. Error bars cover 95 % of the random replicate values. **C)** A more closer visual inspection on the analyzed matrices. Read labels indicate statistical significance of N_{NTC} values.

module formation for the phages case.

4.5 Future Work

We have developed `BiMAT` – an extensible `MATLAB`® library for the analysis of bipartite networks. `BiMAT` implements standard algorithms for the quantification of network structure, including multiple tools to facilitate the analysis of the significance of network structure at the whole network scale, across networks and within networks. The focus on two network features, modularity and nestedness, reflects the importance both have in analyses of bipartite network structure in ecological datasets. However, these are not the only potential features of a bipartite network nor are they necessarily independent.

Indeed, it has been suggested that modularity and nestedness can be strongly correlated [64]. Such correlations may, on the one hand, lead to spurious attempts at classifying a network as either network or modular. Poisot et al [135] have suggested that bipartite networks may be classified based on the degree to which a network is both nestedness and modularity – such classification may relate to the presence of functional groups in the network. Finally, both modularity and nestedness focus on structures of the entire network. However, non-random structures may be present at alternative scales (*e.g.*, see the work on biological network motifs within unipartite networks [8]). We have already made inroads in this direction with a prior proposal [62] and the current automation of a multi-scale bipartite network analysis. Future work is needed to evaluate the extent to which the projection of bipartite networks into a lower dimensional state space can help provide insights into distinct types of networks and, eventually, on connections between network structure and network function.

In moving forward, we hope that `BiMAT` will become a dynamic, extensible tool of use to scientists interested in bipartite networks. We are not the only group to propose such a comprehensive library. For example, a team of UK scientists recently proposed FALCON [158], a library of tools for the analysis of bipartite network structure in

MATLAB® and R. Similarly, we are aware of unpublished efforts to develop a code-base with similar toolsets in R[1]. The study of bipartite networks will necessarily involve those with distinct scientific and computatoinal backgrounds. Hence, so long as the code-bases are open-source, such efforts are likely to reduce barriers in the analysis of bipartite network structure, whether in the ecological, social or physical sciences.

4.6 Citation of methods implemented in BiMAT

The core algorithms implemented in BiMAT are thoroughly described in their original publications and discussed extensively by others. In the case of nestedness, for the NTC metric and implementation, see [11] and [143] and for the NODF metric and implementation, see [7]. In the case of modularity, the standard BRIM algorithms as well as its adaptive heuristic for module division are described by [13]. For a another heuristic using the standard BRIM algorithm, see [108]. For the leading eigenvector algorithm, which is one of the most popular algorithms in unipartite networks, see [128].

[1] L. Zaman, personal correspondence

CHAPTER V

CONCLUSIONS AND FUTURE DIRECTIONS

Many ecological relationships can be expressed as *bipartite complex networks*. The initial motivation of this dissertation came from the study of plant–pollinator networks and the fact that these networks have features that distinguish them from *random networks*. Contrary to these networks, that are mutualistic, phage–bacteria networks are antagonistic, in the sense that phage survive by killing bacteria. In this study we first performed the largest collection of phage–bacteria cross–infection studies and showed that these networks have features that distinguish them from random networks too.

5.1 Summary of major contributions

5.1.1 Phage–Bacteria cross infection data collection

We performed the largest collection and digitization of phage–bacteria cross infection studies. In order to perform that, we looked at papers that date back as far as 1950. The collection includes 38 laboratory–verified studies of phage–bacteria interactions representing almost 12,000 distinct experimental infection assays across a broad spectrum of taxa, habitat, and mode of selection. The collected studies included cross–infection assays that contained isolates related in one of three manners: co–occurring within natural communities and obtained directly from the environment and then cultured, evolved progenitors of a single bacterial clone and a single phage clone that were allowed to co–evolve for a fixed amount of time in a laboratory experiment, and phages and hosts that were artificially combined from laboratory stocks.

Finally we digitized the largest data set of cross–infection phage–bacteria network, from Moebus and Nattkemper [124]. Our digitization of this data set has already been

used by Beckett and Williams [21].

5.1.2 Phage–Bacteria networks are nested

We performed the very first meta–analysis in phage–bacteria networks and showed that independently of the type of study, these networks are in general nested. In doing so, we quantified both nestedness and modularity values of 38 collected data sets categorized in three different types of study. These values were later evaluated statistically using a null model based on the Erdös–Rényi random model [56], which in our case meant that we randomly redistributed the links inside the real networks. The features of these random networks were later compared with the real values in order to quantify the statistical significance.

The strong signal of nestedness is similar in nature to the results obtained by Bascompte and collaborators [17, 130, 18, 20]. The fact that not only mutualistic, but also antagonistic bipartite ecological networks have profound implications in ecology. For instance, how antagonistic networks could be stable was already discussed by Jover et al. [96], where they propose a simple Lotka–Volterra non-linear model in order to find the conditions for these two types of species to coexist. Further, the nestedness property of these networks has been already discussed in terms of co–evolution and species diversity [30, 171, 81].

5.1.3 Phage–Bacteria networks are modular as the study scale increase

Most of the 38 collected networks were studies between different strains of the same bacteria and phage species (*i.e.*, *E. Coli* vs. λ–phage). Hence, the genetic distance between the strains was short. After performing our meta–analysis, we asked ourselves if, by increasing the genetic diversity in a study (and by consequence the size of the study), it could become modular (as strains from long genetic distances will not interact with each other). And the answer is yes. In order to come to this answer we analyzed the largest cross–infection study, performed by Moebus and Nattkemper

[124]. This study is composed (after data curation) of 286 bacteria vs. 215 phages extracted from different locations along the Atlantic Ocean. Moebus and Nattkemper did not include information related to the taxonomy of the collected strains. However, they did observe preliminary evidence of a geographic signal to cross-infection.

After confirming that the study is modular, we also showed that geographical location was a statistically significant factor to explain this structure. In doing so, we compared the current geographical labeling inside modules to random permutations and discovered that for 11 of the 15 largest modules the labeling could be explained by their geographical location. Unfortunately, we could not test if taxonomy is a stronger signal for explaining the modular pattern.

5.1.4 Phage–bacteria network structure changes with size

Related to the previous contribution, we also tested for nestedness at local parts of the Moebus and Nattkemper network (*i.e.*, sub–networks of the entire one). In doing so, we analyzed the structure of each of the 15 largest modules of this study. We find these modules by using a variation of the Adaptive BRIM algorithm [13]. We showed that these internal modules have a nested structure (even when the total network is modular). The multi–scale structure in phage–bacteria networks have profound implications in stability and evolution mechanisms of these communities. Our results and conclusions were extended by Beckett and Williams [21], where they constructed a simple evolution model to explain our multi–scale structure discovery.

5.1.5 Release of `BiMAT`

We released all the code I did during my PhD as a standard `MATLAB`® library for the analysis and visualization of bipartite ecological networks. This software is already being used by people at some research labs around the world (*i.e.*, Sullivan's lab at the University of Arizona, Lennon's Lab at the Indiana University, Earth Systems Science Group at the University of Exeter, and many others). Further, it has been

already used for producing results in peer–reviewed publications [21].

5.2 Future Directions

5.2.1 Network Structure: Individual *vs* Species Density

Current metrics of modularity and nestedness when applied to bipartite ecological networks are limited to the species bipartite adjacency matrix. In other words each node in the network represent an entire species (all of its individuals). However, each species population density can vary over many orders of magnitude. And therefore, if we focus at the individual network level, the structure may look totally different. In order to study that effect, I started during my PhD work the translation of current metrics that depends only in the species adjacency matrix $\mathbf{B}$ ($F = f(\mathbf{B})$) to metrics that depends on the population densities too ($F' = f(\mathbf{B}, \rho_{bacteria}, \rho_{phages})$, where ρ's are vectors that represent the corresponding population densities of each species of bacteria and phages, respectively. Notice that this approach is not the same as ongoing discussions about how to adapt metrics to weighted edge networks. In weighted edge networks we focus on the weight of the edges, while we propose here is to focus on the weight of the nodes.

Such weighted node metrics might reveal a number of interesting aspects of phage–bacteria interaction networks (and any other bipartite ecological relationships). First, these metrics might lead to reinforcement of the patterns that we are already observing using standard metrics. But, on the contrary, if we observe that the patterns are lost when we go to the individual level, we will be able to explain these patterns from a different angle.

5.3 Conclusions

We used simple network metrics to investigate various aspects of phage–bacteria cross–infection networks. This study indicates that a strong nested structure exists in these kinds of networks. However, the nested signal starts to decrease as the scale

of the study increases. We also helped facilitate the research community to perform this type of analysis by releasing a library (`BiMAT`). Finally, we believe that a lot of work can be done in order to extend the robustness of the current network metrics.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Quantitative estimation of nestedness and modularity

We represent the host-phage network with a bipartite network consisting of three sets $G = (U, V, E)$, where U and V are disjoint sets of nodes and $E = \{\{u_i, v_j\}\}$ is the set of edges connecting nodes of different type. For example, Supplementary Figure 24A shows the host-phage network described in Quiberoni [140]. Define $P = |U|$ the number of phages and $H = |V|$ the number of hosts. The adjacency matrix of the bipartite network is $B_{ij} = 1$ if there is an edge $\{u_i, v_j\} \in E$ or $A_{ij} = 0$ otherwise (see Supplementary Figure 24b-c). The number of links attached to node u_i is the so-called degree $k_i = \sum_j B_{ij}$ (similarly, we can define the degree for v_j as $d_j = \sum_i B_{ij}$). Distinct colors indicate whether the node is a host (blue) or a phage (yellow) and bright (dark) shading depicts high (low) degree. Visual inspection of the network reveals significant structure, which can be rigorously detected by means of standard network measurements.

We have examined different properties of host-phage networks. Many real networks have a natural community structure, where disjoint subgroups of nodes exchange many internal connections among then than with the rest of nodes. Formally, we want to compute the optimal division of the network that minimizes the number of links between subgroups (also called communities). The raw number of links at the boundary does not give a good partition of the network. For example, the community structure can be a consequence of random variations in the density of links [80]. A more reliable approach uses a null 1 model to assess the quality of a given network partition. Newman and Girvan [125] defines the modularity for a unipartite networks

as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j) \tag{13}$$

where $2m = \sum_{ij} A_{ij}$ is the number of links and g_i gives the label of the community the node i belongs to. Notice that maximizing the above function yields a partition that minimizes the expected number of links falling between different communities, i.e., when $\delta(g_i, g_j) = 0$. Modularity Q takes values between 0 and 1: low modularity indicates the number of links between distinct communities is not significantly different from the random distribution and high modularity indicates there is a strong community structure.

Our networks are different from the networks studied with the standard modularity measure Q (see above). Here, we study bipartite networks, i.e., networks having two distinct types of nodes and there are no links between nodes of the same type. Barber defines a new modularity quantity $Q_{bipartite}$ using a specific null model for bipartite networks:

$$Q_{bipartite} = \frac{1}{m} \sum_{ij} \left(B_{ij} - \frac{k_i d_j}{m} \right) \delta(g_i, g_j) \tag{14}$$

where $B_{ij} = 1$ if nodes i and j are of different type and 0 otherwise. Related studies of modularity in plant-pollinator networks have used the standard modularity Q [64]. Empirical analyses of bipartite networks have shown that $Q_{bipartite} > Q$, that is, the bipartite modularity can often find better community divisions than the standard modularity when we do not consider the possibility to have links between nodes of the same type [13]. We use the BRIM [13] (Bipartite Recursive Induced Modules) algorithm to maximize this bipartite modularity in our host-phage networks (see the paper by Barber for full details on the BRIM algorithm). For example Supplementary Figure 24A and 24D show the matrix and network representations of the optimal community structure found in a host-phage network. Figure 5B maps

the four network communities found with BRIM into coherent matrix blocks of the (sorted) adjacency matrix. Alternatively, the network representation of community structure in Figure 7d suggests a geometrical interpretation of the maximization of bipartite modularity in terms of link crossing minimization, a hard problem that has been extensively studied in literature [71].

Fortunato and Barthélemy have pointed out that, in large networks, modularity optimization may fail to identify modules smaller than a characteristic size-dependent scale [66]. A check of the modularity obtained through modularity optimization is thus necessary. When modularity optimization finds a module S with l_s internal links, it may be that the latter is a combination of two or more smaller modules. In this case:

$$l_s = \sqrt{2L} \tag{15}$$

where L is the number of links in the full network (see the paper by Fortunato and Barthlémy [66] for full details on the derivation). Modules close to this resolution limit can result from the random merging of two or more sub-modules. Then, modularity optimization might fail to detect the fine modularity structure in these situations.

An important measurement of ecological networks determines to what extent they form a nested network, i.e., when the specialist species only interact with proper subsets of the species interacting with the generalists [18]. The computation of the degree of nestedness involves three steps: (i) computing the isocline of perfect order, which is the curve that separates all the non–zero entries in the adjacency matrix (above the isocline) from the absence of interactions (below the isocline) in a perfectly nested network, (ii) re-arrange all the rows and columns of the adjacency matrix in a way that maximizes the nestedness and (iii) compute the temperature T as the sum of distances dij between the expected and unexpected matrix entries and the isocline:

$$T = \frac{k}{HP} \sum_{ij \in \text{unexpected cells}} \left(\frac{d_{ij}}{D_{ij}} \right)^2 \tag{16}$$

where D_{ij} is the diagonal that cross the unexpected cell and $k = 100/U_{max}$ with $U_{max} = 0.04145$ is a normalization factor that makes $0 \leq T \leq 100$ [18, 143]. Finally, we have normalized the temperature T in such a way that the new range is $0 \leq N \leq 1$:

$$N = \frac{100 - T}{100} \tag{17}$$

Now, for the isocline of perfect order, basically any function that can separate all the non–zero entries from the absence of interactions in a perfect nested matrix can be used. However, in this case we chose the next function from [143]:

$$f(x, p) = \frac{0.5}{n} + \frac{n-1}{n}\left[1 - \left(1 - \frac{mx - 0.5}{m-1}\right)^{p}\right]^{\frac{1}{p}}, \tag{18}$$

where p is the fill of the matrix, and n and m the number of rows and columns, respectively. Before using this function, each cell in the matrix must be matched to a unit square, such that the function will cover the entire matrix using $x \in (0, 1)$.

Supplementary Figure 24C shows the sorted matrix corresponding to the optimal nestedness temperature. This matrix ordering indicates the network is highly nested.

A.2 Criterion for cataloging studies as Co-evolution (EXP), Natural communities (NAT) or Host-phage typing (TYP):

Representative host-phage studies were found using a literature search using ISI Web of Science and tracking references (both to and from the original article). Productive search terms were as follows:

- (phage or bacteriophage) and host and range
- (phage or bacteriophage) and host and typing
- (phage or bacteriophage) and host and infectivity
- (phage or bacteriophage) and characterization

Searching cross-references were also a useful means of collecting infectivity matrices. Web of Science also generated the BibTex reference information for each article. The criteria of inclusion of a study was as follows:

1. Data is available in a matrix/table format in the paper
2. The matrix included interpretable quantitative information on infection
3. The matrix had no missing values
4. The matrix could be manually verified at each cell.
5. The matrix included at least 2 hosts and 2 phages.

Thirty-eight matrices were included in the analysis. Infectivity was indicated either with shading or a (+/-) system. Different amounts of shading would indicate the degree of infection. In the (+/-) system, a '+' generally indicated a positive infection, while a '-' indicated no infection. According to these criterion, we excluded three datasets because of missing data [92, 156, 196]. The criterion for cataloging studies was as follows:

A.2.1 Natural communities (NAT) 19 studies:

This criterion was applied to studies in which both phages and hosts were isolated from the environment. These types of studies are indicative of community interactions within a natural network. These studies were then divided into one of four sub-classes: 1 aquatic, soil, microbiome, and food items. These sub-classes were based upon the environment from which the hosts and phages were isolated.

A.2.2 Co-evolution (EXP) 10 studies:

This criterion was applied to studies in which phages and/or hosts were allowed to evolve in the lab. After phages were allowed to evolve, their host ranges were then

tested. Sub-classes were based upon methodology of the study, and studies were classified as either serial dilution or chemostat experiments. Importantly, matrices of the EXP class need not be reflective of a given community at a fixed moment in time.

A.2.3 Artificial (ART) 9 studies:

This criterion was applied to studies in which almost all hosts and phages were either generated within the lab or came from a collection. Sub-classes indicated the origination of the host strains. Host strains were either environmental or pathogenic.

A.3 Principal component analysis

The objective of PCA is to find a new coordinate system such that the maximal variance is explained in order of each coordinate (i.e., the principal components). Each variable was normalized to have zero mean and a standard deviation of 1 so that each contributed equally to the PCA. Supplementary Figure 20 shows the projection of each study onto the first two principal axes and Supplementary Table 10 shows the detailed coordinates underlying the principal components. Roughly, principal component 1 (PC1) corresponds to the size of the matrix, and so those studies to the right-side of Supplementary Figure 22 tend to be large matrices and those to the left tend to be small matrices. Roughly, PC2 corresponds to the asymmetry between number of phages and number of hosts, so that the top-most studies of Supplementary Figure 22 have more hosts than phages, whereas the bottom-most studies have more phages than hosts. Finally, the third principal component (not shown) corresponds, roughly, to the connectance of the study.

A.4 Statistical analysis of clustering validity using a reshuffling approach

In order to find clusters the k-means algorithm [109] (with $k = 3$) has been applied to the two main components of the PCA analysis output. This 1 output is used as

benchmark for study the subdivision of the studies and compare with those of random labels. The way in which this algorithm works is the next.

Given a set of observations $(\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_n)$, where each observation in our case represents a point in the PCA-analysis output, the k-means aims to partition the n observations into k sets $(k \leq n)$ $S = (S_1, S_2, ..., S_k)$ so as to minimize the within-cluster sum of squares:

$$\arg\min \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{S}_i} ||\mathbf{x}_j - \mu_i|| \tag{19}$$

where μ_i is the mean of the points in Si. In our case $n = 38$ and $k = 3$. See Supplementary Figure 22 for the output of this algorithm.

In order to compare the three clusters found in this algorithm with the three real categories (NAT, EXP, ART) of our studies we used the Jaccard Index defined as:

$$J(C, K) = \frac{a}{a + b + c} \tag{20}$$

Where C represents the real labels and K the labels of the output in the k-means algorithm. a denotes the number of pairs of points with the same label in C and assigned to the same cluster in K, b denotes the number of pairs with the same label, but in different clusters and c denotes the number of pairs in the same cluster, but with different class labels. The index produces a result in the range [0,1], where a value of 1 indicates that C and K are identical.

We find that the three real categories when compared with the output of the k-means algorithm share a Jaccard Index of 0.26. This value indicates that there exist a poor clustering of labels of the studies with the labels of the k-means algorithm. And by consequence we can say (assuming that the k-means output is the perfect subdivision) that there is not significant subdivision in the three real categories (EXP, NAT and ART).

We subjected this index to a randomization test. We generated 10,000 trials where we relabeled the studies while retaining the number of each class (EXP, NAT and ART). The distribution of the Jaccard index of these random trials is showed in Supplementary Figure 23. We found a p-value = 0.34 in the Jaccard index of the real labels. This indicates that there is not a statistically significant difference between the real subdivision of the studies and those that are labeled randomly.

A.5 Statistical analysis of correlations among global properties using a Bonferroni correction

We study the correlations coefficients among the global properties. These values are show in Supplementary Table 11. In that table is showed also the statistical significance of those values. For evaluate the statistical significance we used a Bonferroni correction, using both, the number of combinations and the number of global properties. This correction is used in statistics when one needs to address multiple comparisons. And comes by the fact that even when there is not statistical significance, we can find just by probability that some of the comparisons are statistically significant. Therefore this correction aims to avoid this problem. We can see in the indicated table that among the statistically significant values there is only a strong correlation between the number of hosts and the number of species. Another interesting result is that there is almost no correlation (no statistical significance) between the connectance and the number of species. This is contrary to the plant–pollinator networks where the relation follows a power law.

A.6 Experimental assays of host–phage infection

A.6.1 Conditions and microbial cultures

The phage and bacteria were cocultured in 50ml Erlenmeyer flasks, with 10ml of liquid medium, shaken at 120 rpm, and incubated at 37 °C. The medium was an altered version of Davis Medium (15), in which we added 10 times the magnesium

sulfate (1g/L) to improve phage viability and 125 mg/L of maltotriose instead of glucose because *E. coli* and phage λ are predicted to undergo a coevolutionary arms-race when provided with maltodextrins as its only source of carbon [105, 159, 186] The medium was filtered and the magnesium was added just before use in order to stop crystallization of the magnesium during the experiment. 75 separate flasks were initiated with very small populations of bacteria ($\sim 1,000$ *E. coli* cells) and phage ($\sim$ 100 phage λ particles) to assure that the initial populations were isogenic and that all mutant bacteria and phage arose de novo, this is important to make sure that each community has the potential to follow its own coevolutionary path. The *E. coli* studied were of strain REL606, a derivative of *E. coli* B acquired from Richard Lenski (Michigan State University), described in [45] and phage were of strain cI21 (λvir) provided by Donald Court (National Cancer Institute). Most phage λ strains have two life cycles, lytic and lysogenic, the second includes a latent phase where the phage genome is incorporated into the bacterial chromosome at which time the bacteria acquires immunity to phage infection. Because the goal of this study was to characterize evolved phage resistance instead of acquired resistance, we used a phage that was 1 unable to create the resistant lysogenic bacteria. cI21 is only able to reproduce through the lytic phase because it has a chemically induced mutation in the cI gene which is a repressor protein required for lysogeny. Each flask was cultured for 24 hours and then a random subsample of 100ul of the culture was removed and transferred to 9.9ml of fresh medium. This flask was incubated and the cycle of transfer and incubation was continued once more. Three 24 hour incubations were long enough for the bacteria to evolve resistance and the phage to counter it, however not long enough for a second round of coevolution.

A.6.2 Isolation strategies

After 72 hours of coculturing, two bacterial clones were isolated from each flask by streaking on LB (Luria Burtani medium, recipe found in [148]) agar plates and picking single colonies. These colonies were restreaked twice more to assure the bacteria was separated from the phage. A mixed phage stock of all coevolved genotypes was created from each flask by adding 500 l of chloroform to the remaining culture in order to kill the bacterial cells, which were removed by centrifugation [3]. Two phage clones were isolated from each of these mixed phage stocks by applying an aliquot of diluted stocks onto soft agar plates and picking isogenic 'plaques'. Soft agar plates are created by suspending an isogenic population of bacteria combined with the diluted phage stock in a thin agar matrix on top of a petri dish. When a single phage particle infects a bacterial cell trapped in the agar, the phage reproduces and spreads to nearby bacteria, this continues for a number of rounds and a clearing known as a plaque is produced in the 'lawn' of viable bacteria after 24 hours of incubations at 37 °C. This plaque contains an isogenic population of phage that can be removed to create a clonal stock of phage. We made three plates for each coevolved viral population; one from each bacterial clone isolated from the same population and then one of the ancestral bacteria REL606. Clonal phage cultures were created by isolating single plaques from the soft-agar plates and following the procedure given by [49]. Plaques on the coevolved bacteria were chosen over ones grown on REL606 to increase the chance of isolating phage that had evolve specialized counter-resistance strategies that have the plietropic consequence of losing the ability to infect the ancestral REL606. Despite this effort, none of the phage isolated lost the ability to infect REL606. Besides favoring plaques on the evolved bacterial plates, we tried to choose plaques from separate 1 plates to improve our chances of picking different phage genotypes.

A.6.3 Evaluating patterns of infection and cross-resistance

We determined which of the 150 bacteria isolates were resistant to the 150 phage isolates. To do this we preformed 'spot' plate assays. Spot plates are created just as the soft agar plates above were, except instead of combining dilute samples of phage into the agar, one drops 2 ul of concentrated phage stock on top of the bacterial-agar matrix. If the phage is able to infect and reproduce on the bacterium, then a clearing or 'spot' larger than a single plaque will form in the bacterial lawn after 24 hours of incubations at 37 °C. If any clearing or inhibition of bacterial growth larger than a single plaque was observed a '1' was recorded. Plaque-sized clearings were excluded because they likely represent cross-contamination or a mutant phage that has a broader host-range than the originally isolated phage. All bacterial-phage combinations without '1's were given '0's. All bacterial phage combinations were replicated five separate times, a total of 28,125 spots were assayed. To make this processes more efficient, we placed up to 96 separate phage stocks onto a single dish (150mm radius). Phage stock replicates were never placed on the same plate in order to reduce the signal of any stochastic plating effects. The five replicates were combined and a phage was only determined to be able to infect a bacterium if 3 of 5 replicates were given '1's. Lastly, phage or bacteria that had identical infection resistance profiles as their ancestors were removed from the matrix.

Table 7: Characteristics of complete host-phage networks included in the present study

	Reference	*Source Type*	*H*	*P*	*S*	*I*	*M*	*C*	*Lp*	*Lh*
1	Abe (2007)	ecological	11	4	15	22	44	0.5	5.5	2
2	Barrangou (2002)	ecological	14	6	20	25	84	0.3	4.17	1.79
3	Braun-Brenton (1981)	experimental	18	3	21	30	54	0.56	10	1.67
4	Campbell (1995)	experimental	9	5	14	14	45	0.31	2.8	1.56
5	Capparelli (2010)	ecological	18	8	26	54	144	0.38	6.75	3
6	Caso (1995)	experimental	23	4	27	17	92	0.18	4.25	0.74
7	Ceyssens (2009)	artificial	5	15	20	29	75	0.39	1.93	5.8
8	Comeau (2005)	experimental	30	13	43	152	390	0.39	11.69	5.07
9	Comeau (2006)	experimental	32	16	48	118	512	0.23	7.38	3.69
10	DePaola (1998)	ecological	5	17	22	39	85	0.46	2.29	7.8
11	Doi (2003)	artificial	15	10	25	41	150	0.27	4.1	2.73
12	Duplessis (2001)	artificial	12	12	24	37	144	0.26	3.08	3.08
13	Gamage (2004)	ecological	6	7	13	9	42	0.21	1.29	1.5
14	Goodridge (2003)	ecological	93	2	95	60	186	0.32	30	0.65
15	Hansen (2007)	ecological	34	12	46	146	408	0.36	12.17	4.29
16	Holmfeldt (2007)	artificial	23	46	69	418	1058	0.4	9.09	18.17
17	Kankila (1994)	ecological	32	12	44	346	384	0.9	28.83	10.81
18	Krylov (2006)	ecological	11	10	21	73	110	0.66	7.3	6.64
19	Kudva (1999)	artificial	22	3	25	33	66	0.5	11	1.5
20	Langley (2003)	ecological	66	9	75	99	594	0.17	11	1.5
21	McLaughlin (2008)	ecological	8	7	15	18	56	0.32	2.57	2.25
22	Meyer (unpub)	experimental	25	27	52	314	675	0.47	11.63	12.56
23	Middelboe (2009)	experimental	11	24	35	202	264	0.77	8.42	18.36
24	Miklic (2003)	ecological	24	14	38	70	336	0.21	5	2.92
25	Mizoguchi (2003)	experimental	8	4	12	21	32	0.66	5.25	2.63
26	Pantucek (1998)	artificial	102	4	106	322	408	0.79	80.5	3.16
27	Paterson (2010)	experimental	100	5	105	267	500	0.53	53.4	2.67
28	Poullain (2008)	experimental	24	24	48	107	576	0.19	4.46	4.46
29	Quiberoni (2003)	ecological	20	11	31	89	220	0.4	8.09	4.45
30	Rybniker (2006)	artificial	17	14	31	70	238	0.29	5	4.12
31	Seed (2005)	artificial	24	6	30	31	144	0.22	5.17	1.29
32	Stenholm (2008)	ecological	28	22	50	348	616	0.56	15.82	12.43
33	Sullivan (2003)	ecological	21	44	65	148	924	0.16	3.36	7.05
34	Suttle (1993)	artificial	7	9	16	11	63	0.17	1.22	1.57
35	Synnott (2009)	ecological	16	16	32	207	256	0.81	12.94	12.94
36	Wang (2008)	ecological	18	7	25	11	126	0.09	1.57	0.61
37	Wichels (1998)	ecological	59	23	82	318	1357	0.23	13.83	5.39
38	Zinno (2010)	ecological	18	27	45	49	486	0.1	1.81	2.72
		Average	26.55	13.21	39.76	114.87	314.32	0.39	10.91	4.88
		Median	19.00	10.50	31.00	65.00	203.00	0.34	6.13	3.04
		Total	1009	502	1511	4365	11944			

First column: These ID's corresponds to indexes in supplementary figures 20–22.

Table 8: Characteristics of complete host-phage networks included in the present study, including additional information on biological context of each study

	Reference	*Bacteria*	*Phage*	*Majority source*	*Additional source*	*Isolation Habitat*	*Bacterial association*	*Bacterial trophy*	*Geography*
1	Abe (2007)	*Escherichia coli*	*T2 and PP01*	ecological	artificial		human pathogen	heterotrophic	
2	Barrangou (2002)	*Leuconostoc*	*Caudovirales*	ecological	artificial	sauerkraut	free	heterotrophic	North Carolina, USA
3	Braun-Brenton (1981)	*Escherichia coli*	λ	experimental		lab-agar plates	human symbiont	heterotrophic	
4	Campbell (1995)	*Pseudomonas*	*Myoviridae*	experimental	ecological	barley roots	plant symbiont	heterotrophic	Hojbakkegaard, Denmark
5	Capparelli (2010)	*Salmonella*		ecological		gastroenteritis patients	human pathogen	heterotrophic	Europe
6	Caso (1995)	*Lactobacillus*	*Siphoviridae*	experimental		food, fresh water, soil, sewage	free	heterotrophic	Spain
7	Ceyssens (2009)	*Pseudomonas aeruginosa*		artificial		hospital sewage, fresh water	human pathogen	heterotrophic	global
8	Comeau (2005)	*Vibrio*		experimental		marine	human pathogen / oysters	heterotrophic	British Columbia, Canada
9	Comeau (2006)	*Vibrio*	*Siphoviridae and Podoviridae*	experimental		marine	human pathogen	heterotrophic	British Columbia, Canada
10	DePaola (1998)	*Vibrio vulnificus*	*Podoviridae, Styloviridae, and Myoviridae*	ecological		marine	human pathogen / oysters	heterotrophic	Gulf of Mexico
11	Doi (2003)	*Lactobacillus*	*Siphoviridae and Myoviridae*	artificial		silage (fermented bovine feed)	free	heterotrophic	Japan
12	Duplessis (2001)	*Streptococcus thermophilus*	*Myoviridae and Siphoviridae*	artificial		Industrial cheese plants	free	heterotrophic	Quebec, Canada
13	Gamage (2004)	*Escherichia coli*		ecological		human and animal fecal isolates	human pathogen	heterotrophic	Ohio, USA
14	Goodridge (2003)	*Enterobacteriaceae*	*Myoviridae*	ecological		human and animal	human pathogen	heterotrophic	global
15	Hansen (2007)	*Campylobacter*	*Myoviridae*	ecological		poultry intestine	human pathogen	heterotrophic	Denmark
16	Holmfeldt (2007)	*Flavobacteriaceae*	*Myoviridae, Siphoviridae, and Podoviridae*	artificial	ecological	marine	free	heterotrophic	Scandinavia
17	Kankila (1994)	*Rhizobium*		ecological		soil	free	heterotrophic	Finland
18	Krylov (2006)	*Escherichia and Salmonella*	*T-even superfamily*	ecological		sewage	human pathogen	heterotrophic	
19	Kudva (1999)	*Enterobacteriaceae*		artificial		bovine and ovine feces	human pathogen	heterotrophic	North West USA
20	Langley (2003)	*Burkholderia*	*T-even and λ-like*	ecological	artificial	soil, freshwater, plant	mycorrhizal	heterotrophic	global
21	McLaughlin (2008)	*Salmonella*		ecological	artificial	swine lagoon	human pathogen	heterotrophic	Mississippi, USA
22	Meyer (unpub)	*Escherichia*		experimental		lab - batch culture	human symbiont	heterotrophic	
23	Middelboe (2009)	*Cellulophaga baltica*	*Myoviridae, Siphoviridae, and Podoviridae*	experimental	ecological	marine	free	photosynthetic	Scandinavia
24	Miklic (2003)	*Lactococcus lactis*	*Siphoviridae*	ecological		dairy products	free	heterotrophic	Solvania
25	Mizoguchi (2003)	*Escherichia coli*	*PP01*	experimental		lab-chemostat	human pathogen	heterotrophic	
26	Pantucek (1998)	*Staphylococcus*	*polyvalent staphylophage*	artificial		clinical isolates	human pathogen	heterotrophic	Brno, Czech Republic
27	Paterson (2010)	*Pseudomonas fluorescens*	*ϕ2*	experimental		lab - batch culture	plant symbiont	heterotrophic	UK
28	Poullain (2008)	*Pseudomonas fluorescens*	*ϕ2*	experimental		lab - batch culture	plant symbiont	heterotrophic	UK
29	Quiberoni (2003)	*Streptococcus thermophilus*	*Siphoviridae*	ecological		yogurt industrial plant	free	heterotrophic	Argentina
30	Rybniker (2006)	*Mycobacterium*		artificial		soil	human pathogen	heterotrophic	global
31	Seed (2005)	*Burkholderia*	*Myoviridae*	artificial		soil, freshwater, plant	human pathogen	heterotrophic	
32	Stenholm (2008)	*Flavobacterium psychrophilum*	*Siphoviridae, Myoviridae, and Podoviridae*	ecological		fresh water	fish pathogen	heterotrophic	Denmark
33	Sullivan (2003)	*Prochlorococcus* *Synechococcus*	*Myoviridae and Podoviridae*	ecological		marine	free	photosynthetic	Atlantic Ocean
34	Suttle (1993)	*Synechococcus and Anacystis*	*Siphoviridae, Myoviridae, and Podoviridae*	artificial	ecological	marine	free	photosynthetic	Texas, USA
35	Synnott (2009)	*Staphylococcus aureus*	*Myoviridae*	ecological		sewage, dairy products	bovine pathogen	heterotrophic	Tokyo, Japan
36	Wang (2008)	*Synechococcus and Prochlorococcus*	*Myoviridae and Podoviridae*	ecological		marine	free	photosynthetic	Chesapeake Bay, USA
37	Wichels (1998)	*Pseudoalteromonas*	*Siphoviridae, Myoviridae, and Podoviridae*	ecological		marine	free	heterotrophic	North Sea, Germany
38	Zinno (2010)	*Streptococcus thermophilus*		ecological		dairy products	free	heterotrophic	Italy

Table 9: Global properties

Property	***Definition***
H	number of hosts
P	number of phages
I	number of interactions
$S = H + P$	number of species
$M = HP$	size
$C = I/M$	connectance
$LH = I/H$	mean number of interactions across host species
$LP = I/P$	mean number of interactions across phage species

Table 10: PCA Analysis

	1st	***2nd***	***3rd***	***4th***	***5th***	***6th***	***7th***	***8th***
H	0.352	0.446	-0.179	0.131	0.389	-0.131	-0.097	0.67
P	0.247	-0.534	-0.203	0.474	-0.461	-0.14	-0.279	0.279
I	0.47	-0.138	0.143	-0.474	0.008	0.517	-0.498	0
$S = H + P$	0.444	0.218	-0.257	0.32	0.192	-0.184	-0.208	-0.688
$M = HP$	0.397	-0.239	-0.359	-0.542	-0.078	-0.373	0.466	0
$C = I/M$	0.188	0.062	0.743	-0.093	-0.112	-0.601	-0.164	0
$LH = I/H$	0.281	-0.449	0.359	0.313	0.504	0.224	0.435	0
$LP = I/P$	0.353	0.431	0.177	0.177	-0.571	0.335	0.434	0
	48.95%	*27.98%*	*18.55%*	*2.03%*	*1.30%*	*1.07%*	*0.11%*	*0*

Table 11: Correlation analysis

	H	***P***	***S***	***I***	***M***	***C***	***Lp***	***Lh***
H	1	-0.146	*0.916	$^{+}$0.458	0.394	0.125	*0.847	-0.133
P		1	0.264	*0.535	*0.744	-0.11	-0.191	*0.697
S			1	*0.664	*0.686	0.077	*0.748	0.154
I				1	*0.752	$^{+}$0.466	*0.553	*0.716
M					1	-0.109	0.204	$^{+}$0.449
C						1	*0.501	*0.517
Lp							1	0.035
Lh								1

$^{*}p$-value $< 0.05/28$

$^{+}0.05/28 < p$-value $< 0.05/8$

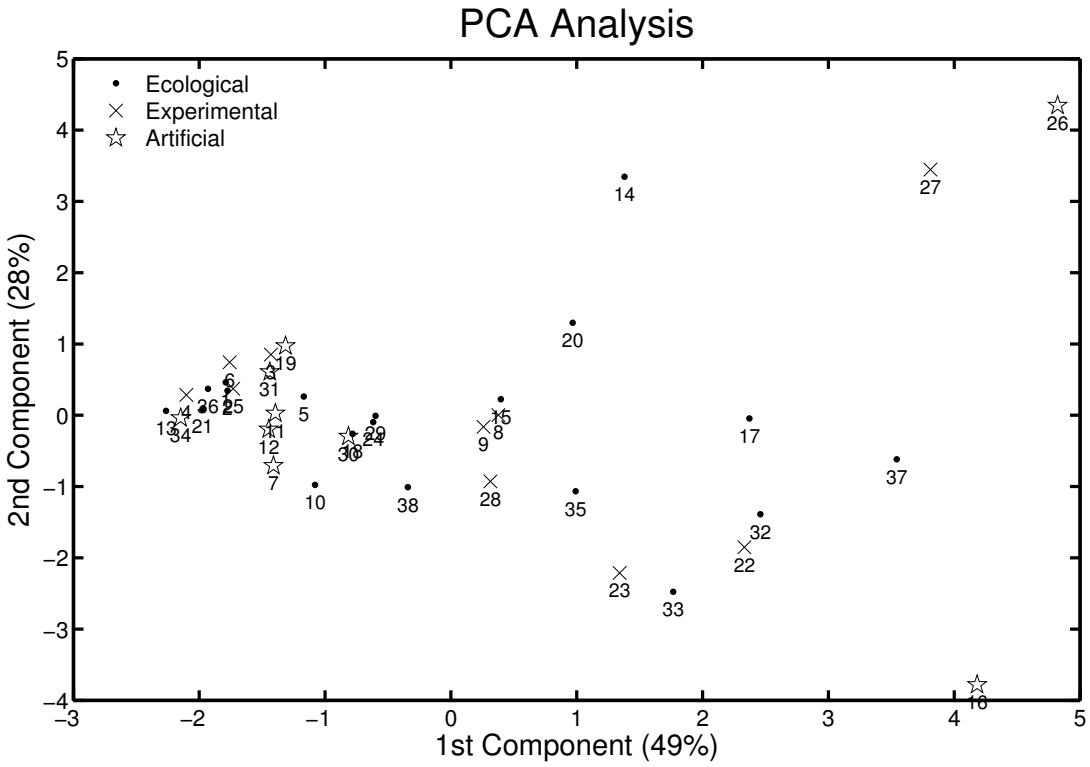

Figure 20: PCA Analysis in the global properties of the collected studies. Only the two main components are showed. There is no distinction between the three different type of studies.

Table 12: Isolation bias

	Modularity		*Nestedness*	
Study	***Original***	***Recalculated***	***Original***	***Recalculated***
Krylov 2006	$^{+}0.123$	$^{+}0.136$	$^{*}0.901$	$^{*}0.839$
Kudva 1999	$^{+}0$	$^{+}0$	0.63	0.63
McLaughlin 2008 - Matrix minus TSB control	$^{+}0.191$	$^{+}0.191$	$^{*}0.978$	$^{*}0.951$
McLaughlin 2008 - Matrix minus TSB minus isolation host	$^{+}0.191$	0.313	$^{*}0.978$	$^{*}1$
Middleboe 2009	$^{+}0.084$	$^{+}0.079$	$^{*}0.988$	$^{*}0.98$
Rybniker 2006	$^{+}0.333$	$^{+}0.274$	$^{*}0.931$	$^{*}0.908$
Stenholm 2009	$^{*}0.183$	$^{*}0.187$	$^{*}0.928$	$^{*}0.931$

*Significant modular/nested studies
$^{+}$Significant anti-modular/nested studies

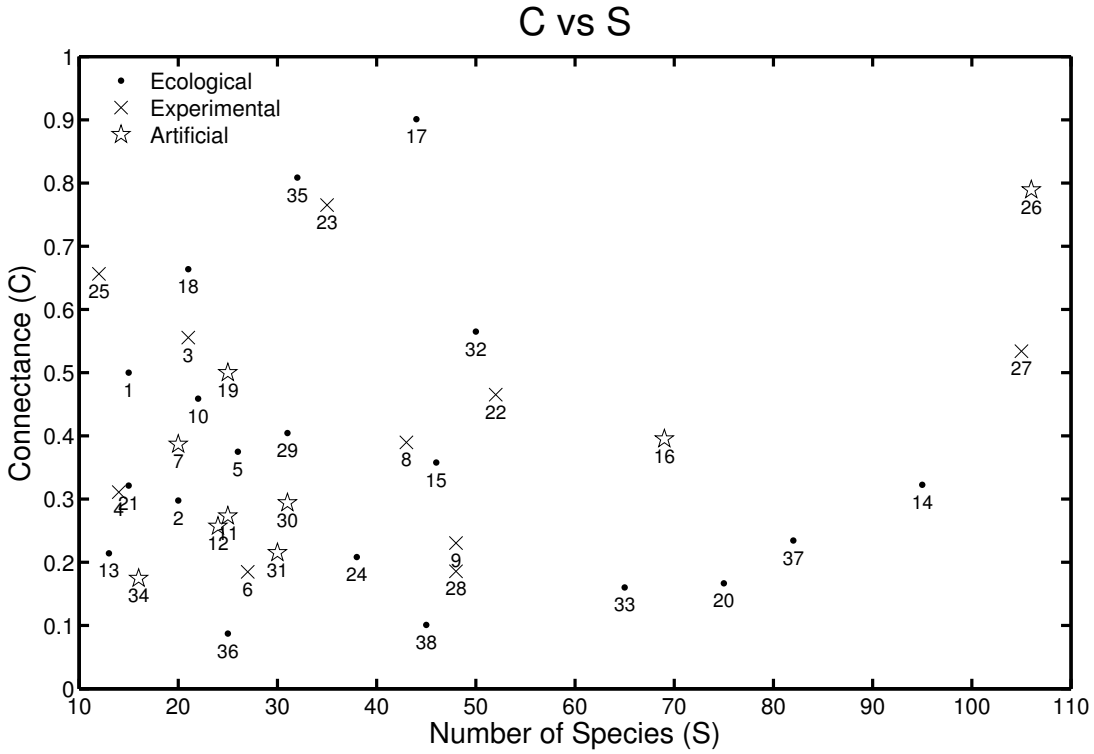

Figure 21: Correlation between connectance (C) and number of species (S). This plot shows that there is no relation between the connectance and the number of species. Numbers in both plots indicate the study id that can be consulted in the appendix

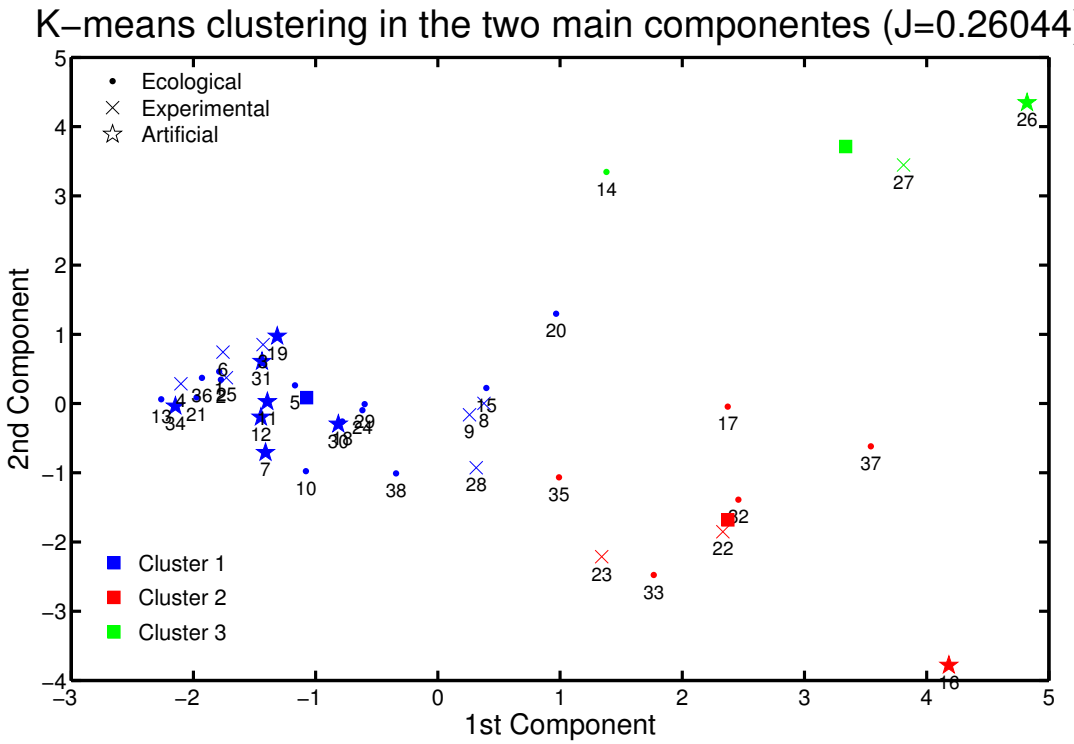

Figure 22: Output of the k–means (with $k = 3$) algorithm when applied to the two main components of the PCA–analysis output.

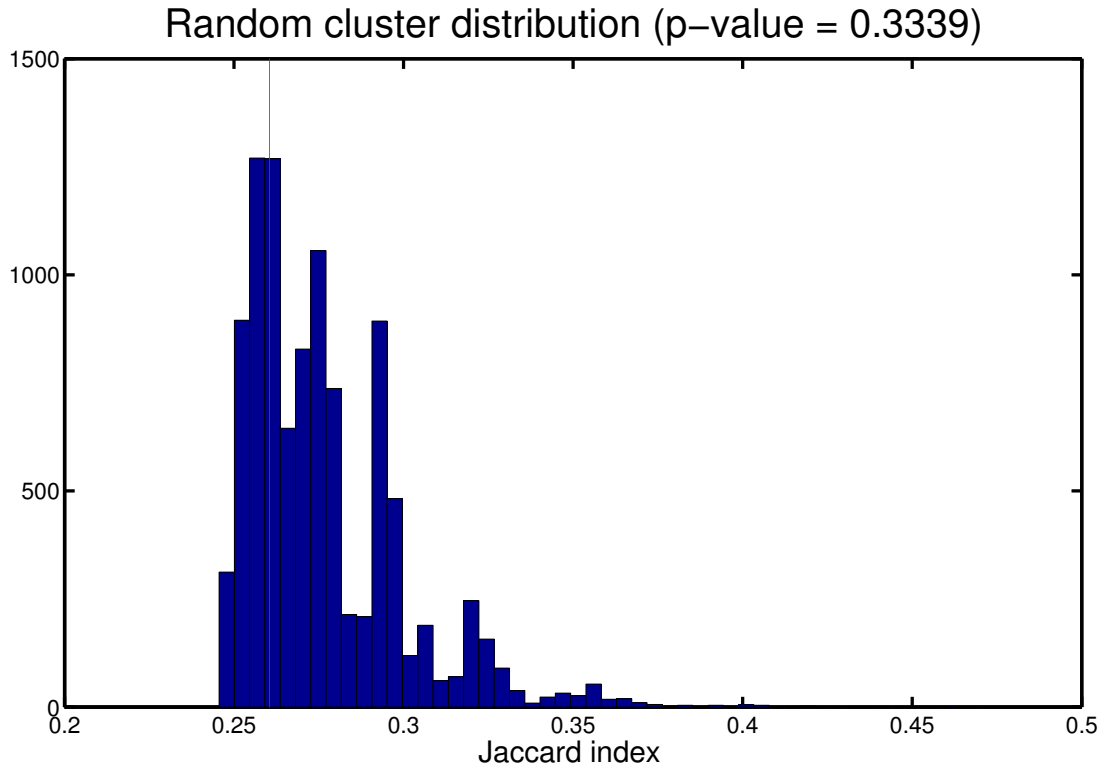

Figure 23: Distribution of clustering validity of source types (EXP, NAT and ART) based on global properties. The histogram denotes 10,000 randomization trials in which the labels of each study were relabeled while retaining the total number of each class (EXP, NAT and ART). The value on the x-axis is the Jaccard index of clustering validity (see Supplementary Materials and Methods). The red line denotes the observed clustering validity for the data set which is non-significant, $p = 0.34$.

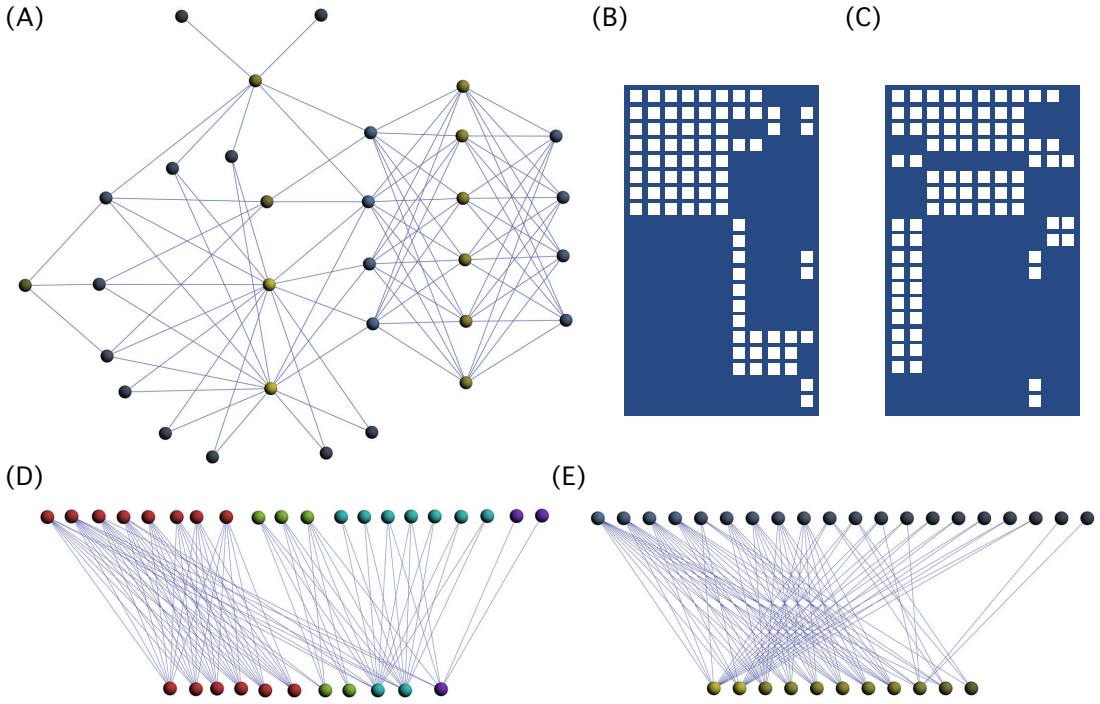

Figure 24: Matrix and network representations reveal non-random patterns in host-phage networks. **(A)** Force-directed layout of the host-phage network where yellow and blue nodes represent phages and hosts, respectively. Shading represents the number of node connections, or degree (see text). We can re-arrange the rows and columns of the adjacency matrix according to optimal network modularity **(B)** and degree of nestedness **(C)**. **(D)** Strong modularity indicates the presence of subsets of nodes with the same color (communities) having many more internal links than external links (i.e., less crossings across different modules). **(E)** Network representation evidences a high degree of nestedness overall, with a few unexpected interactions between specialist species (on the right). Notice that generalist species have more connections and they are located on the left.

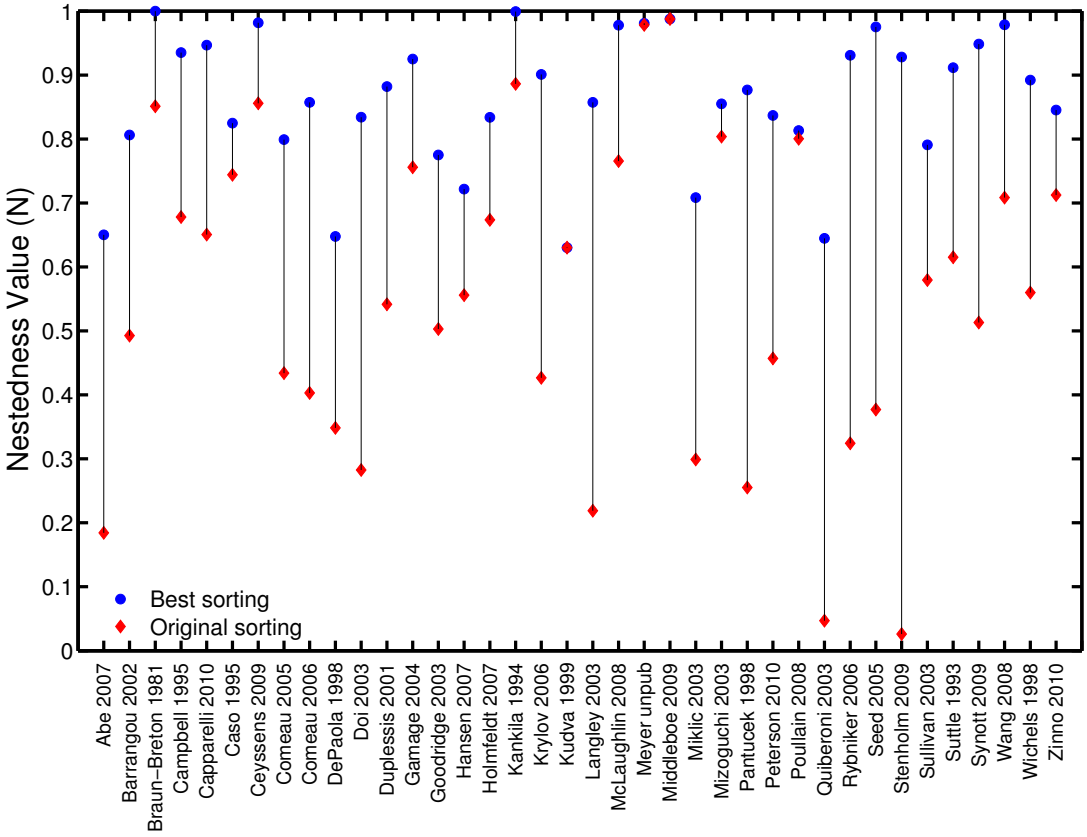

Figure 25: Nestedness value compared for the original publication format of the matrix (red diamonds) vs. the value found in this study (blue circles). X-axis lists all studies in alphabetical order. Y-axis denotes the value of nestedness. Lines connect the points for ease of comparison. Note that in all cases the current value exceeded that of the original publication.

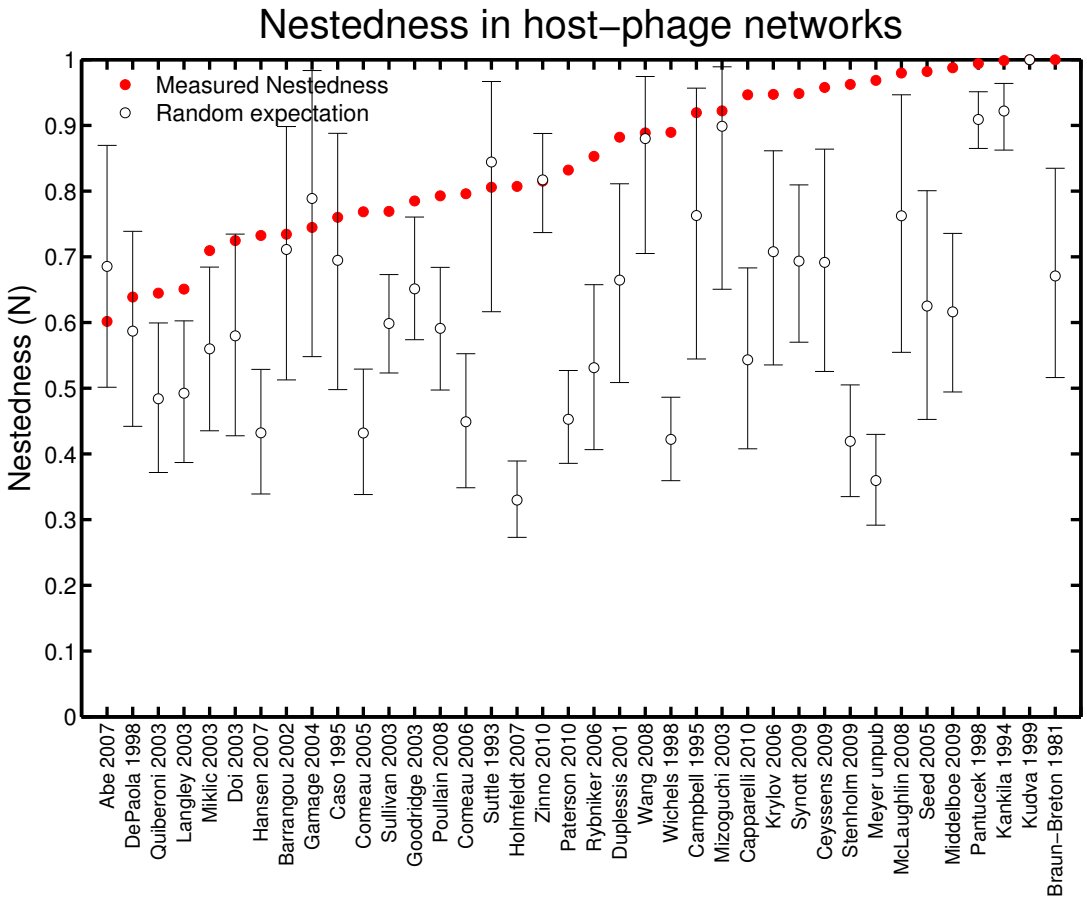

Figure 26: Statistical distribution of nestedness for random matrices compared to that of the original data. Here, empty rows/columns from all matrices were removed so that matrices only contain hosts that were infected by at least one phage and phages that infected at least one host. Error bars denote 95 % confidence intervals based on 10^5 randomizations of appropriately randomized null networks. Here 26/38 are significantly nested, where Doi et al.(22) is the only study to no longer be significant at the 0.05 level compared to the original data, yet it remains highly nested ($p = 0.067$).

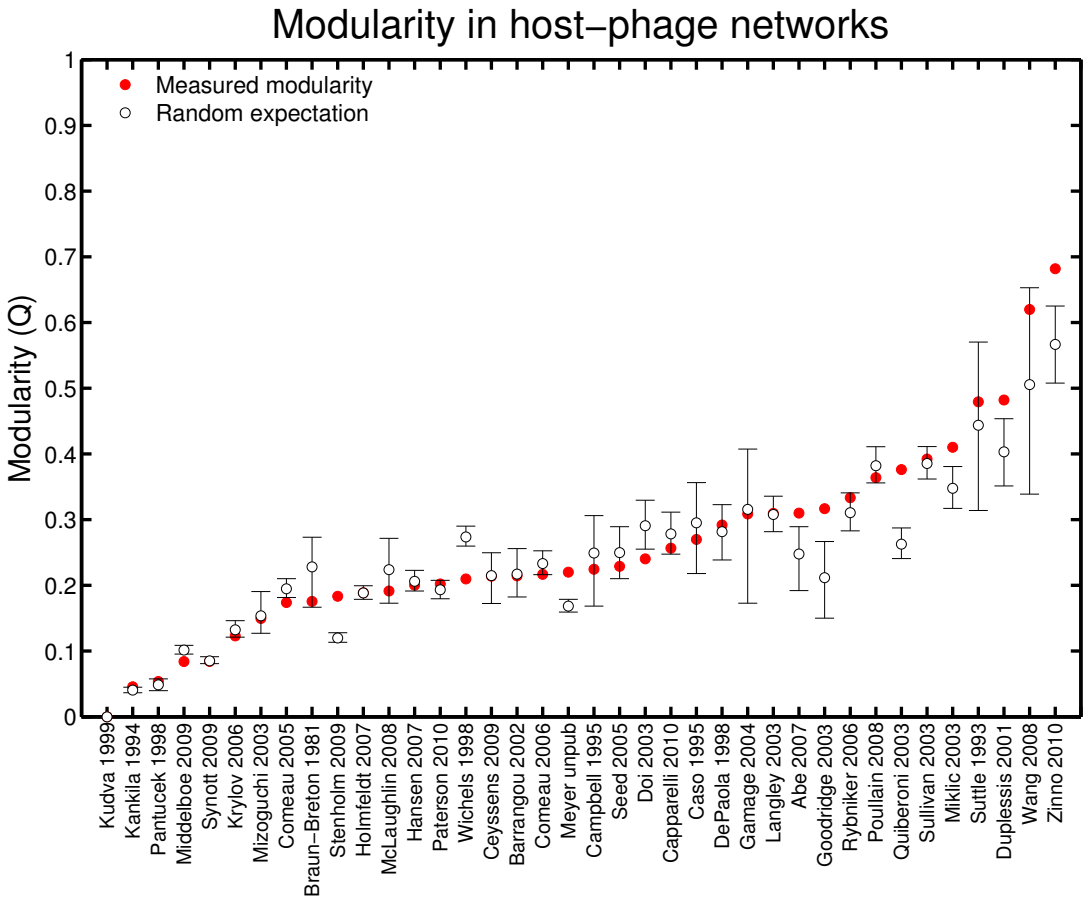

Figure 27: Statistical distribution of modularity for random matrices compared to that of the original data. Here, empty rows/columns from all matrices were removed so that matrices only contain hosts that were infected by at least one phage and phages that infected at least one host. Error bars denote 95 % confidence intervals based on 10^5 randomizations of appropriately randomized null networks. Here 9/38 are significantly modular as opposed to 6/38 which were significantly modular in the original data.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Dataset

The dataset analyzed here is a subset of the phage-bacteria cross-reaction tests reported by K. Moebus and H. Nattkemper [124]. Among all the datasets reported in this paper, we have focused in the largest collection of tests, i.e., the so-called A-series dataset. This dataset consists of $H = 733$ bacteria and $P = 258$ bacteriophages strains collected at 48 water sample stations in the Atlantic Ocean region (see Figure 28). Only 326 out of the 733 bacteria were found to be susceptible to one or more phages. From the 326 bacteria strains, 250 are unique (the infection pattern is different from each other), 38 are inter-sample doublets (bacteria that have the same infection pattern of another bacteria belonging to a different water sample or station), and 38 intra-sample doublets (doublets from the same water samples). Similarly, there are 224 unique phage strains and 4 inter-sample doublets.

The only source of information about the matrix of cross-reaction tests was the figure shown in the Moebus and Nattkemper paper (see Figure 1 in [124], Figure 29 in this document). We were unable to find other means to access this dataset and thus, we have developed a semi-automatic scanning method to recover this matrix from the printed paper to a digital format suitable for our analysis (see method below). For example, the original paper does not indicate the exact number of bacteria and phages represented in the original figure (see Figure 29). Instead, these numbers have been inferred from the original figure labels and the information given in the whole document (see below). The digitalization process includes the following steps:

1. We scanned the source image from the printed figure in [124] (see Figure 1 in

[124] and Figure 29 in this document). The quality of the image made the extraction process difficult. First, the original image is slightly rotated by an angle comprised between 0.4 and 0.6 degrees counterclockwise (depending on what side of the image is chosen as a reference). In addition, there was a tear starting at the bottom (phage station number 484) and running to the left (phage station number 462) of the image that slightly distorts the orientation at the bottom right section. Here, we have estimated the rotation angle to be 0.45 degrees, which is good compromise between the left and bottom orientations. As a consequence of the previous rotation, two bacteria records were lost.

2. We assume that matrix size is approximately equal to the number of columns and rows visible in the source image. We manually cross-checked the row and column counts and find $H = 288$ bacteria and $P = 222$ phages. Further validation comes from a computer program that counts the number of mouse clicks performed by a human over each bacteria/phage label in the "source" (scanned) image. The observed number of bacteria is consistent with the caption of the source figure that reports 288 bacteria strains (250 unique + 38 inter-sample doubles). The case for phages is more ambiguous because the original figure only labels 217 phages out of the 222 (readable) columns. Here, we have only retained labeled and readable phages to yield $H = 286$ bacteria and $P = 215$ phages.

3. We performed a binary thresholding of the source matrix to automatically detect positive interactions of phages with hosts by computing the density of filled pixels at every matrix cell. We delimited the matrix cells by overlaying a grid in the source figure, and the interactions were detected by specifying a threshold of filled pixels inside each cell. This automatic process makes no distinction between matrix cells that denote clear lysis or turbid spots.

4. We manually curated the binary thresholded image to identify and correct any false negatives (undetected interactions) and false positives (empty cells marked as interactions). In addition, empty columns were removed. The output is the curated MN (Moebus and Nattkemper) matrix used for our study (see Figure 9 of Chapter 3, and Figure 29).

B.2 Bipartite Modularity

A host-phage interaction matrix can be described as a bipartite network $G = (U, V, E)$ having two disjoint sets of nodes (phages and hosts) and a set of edges ([60]). Here, $H = \|U\|$ is the number of hosts and $P = \|V\|$ is the number of phages and there is an edge $\{u_i, v_j\} \in E$ when phage $v_j \in V$ infects host $u_i \in U$. Notice that interactions between nodes of the same type are excluded. Alternatively, the *adjacency matrix* $A = [A_{ij}]$ indicates whether the $j-$th phage can infect the $i-$th host ($A_{ij} = 1$) or not ($A_{ij} = 0$). Notice that this matrix corresponds to the binary thresholded image obtained in the previous section. A number of useful network measures can be obtained from the adjacency matrix alone. The degree $k_i = \sum_j A_{ij}$ of the $i-$th host is the number of interactions with phages (i.e. how many phages can infect the $i-$th host). The degree $d_j = \sum_i A_{ij}$ of the $j-$th phage is the number of interactions with hosts (i.e. how many hosts can be infected by the $j-$th phage). See Figure 30 for a plot of the cumulative degree frequency of the MN matrix.

An important collection of network measures involves the quantification of interaction patterns in subsets of more than two network nodes. For example, a visual inspection of the infection matrix shown in Figure 11 of Chapter 3 suggests that there are modules of hosts and phages exchanging many more "ones" between them (a higher density of internal links) than with the rest of types (nodes). Following [13], we assess the quality of a given partition in c (disjoint) modules with the *bipartite*

modularity:

$$Q = \frac{1}{m} \sum_{ij} \left(A_{ij} - P_{ij} \right) \delta(g_i, g_j) \tag{21}$$

where A_{ij} is the adjacency matrix, $m = \sum_{ij} A_{ij}$ is the total number of links, $P_{ij} = k_i d_j / m$ is the probability to connect nodes i and j, the node i has been assigned to the module g_i, and $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ when x and y are different. Intuitively, high values of Q will correspond to highly modular partitions of the bipartite network. In this case, node i and j are classified in the same module $g_i = g_j$ (and thus $\delta(g_i, g_j) = 1$) because the probability to have a link between nodes in the same module is significant (e.g., the difference $A_{ij} - P_{ij}$ is a large, positive value).

For convenience, we use the matrix form of the modularity Equation (21). Here, we replace the function g_i by the $H \times c$ index matrix $\mathbf{R} = [\mathbf{r}_1 | \mathbf{r}_2 | ... | \mathbf{r}_c]$ and the $P \times c$ index matrix $\mathbf{T} = [\mathbf{t}_1 | \mathbf{t}_2 | ... | \mathbf{t}_c]$, for hosts and phages, respectively [13]. Notice that nodes cannot be classified into more than one module. Vectors $\mathbf{r}_i$ and $\mathbf{t}_i$ consist of a single one (corresponding to the chosen module) will all the other entries being zero. For example, $r_{ik} = 1$ if the i-th host belongs to the k-th module and $r_{ij} = 0$ for every other $j \neq k$. Now, we can rewrite the modularity as follows (see Equation (22) in [13]):

$$Q = \frac{1}{m} \operatorname{Tr} \mathbf{R}^T \tilde{\mathbf{B}} \mathbf{T} \tag{22}$$

where $\tilde{\mathbf{B}} = \mathbf{A} - \mathbf{P}$ is the *bipartite modularity matrix*. The goal of the modularity algorithm is to find the optimal assignment of nodes to modules (i.e., the index vectors $\mathbf{R}$ and $\mathbf{T}$) in a way that Equation 22 becomes maximized. However, finding the optimal modularity is a NP-complete problem. In this context, there are a number of practical heuristics that we can use to guide modularity algorithms in the search for good solutions within computational constraints (we always check that the solutions found by the algorithms are meaningful). Next, we discuss the different heuristics explored here.

The original modularity algorithm (called BRIM for Bipartite, Recursively Induced Modules) described in [13] computes the optimal modularity by inducing the division of one set of nodes (say vector $\mathbf{T}$) from the division in the other set of nodes (say vector $\mathbf{R}$). At each step, BRIM assigns nodes of one type to modules in order to maximize the modularity. BRIM iterates this process until a local maximum is reached. However, the choice of a predefined number c of modules limits the efficacy of the algorithm. Barber extended the BRIM algorithm to search for the optimal number of modules along the modularity maximization process [13]. This method, which is called *"adaptive BRIM"*, assumes that there is a smooth relationship between the number of modules c and the modularity $Q(c)$. For continuous and smooth landscapes, a simple bisection method ensures that we will find the optimal value of c corresponding to maximum Q. Starting at $c = 1$ (and modularity $Q(1) = 0$ because all nodes belong to the same module) the adaptive BRIM searches for optimal c by repeatedly doubling the number of modules while modularity increases, $Q(2c) > Q(c)$. At some point, the search crosses a maximum in the modularity landscape, i.e., $Q(2c) < Q(c)$, and we interpolate the number of modules c^* to some intermediate value in the current interval $(c, 2c)$. This heuristic gives very good modularity values for the case of small matrices. For example, we have used the adaptive heuristic in the analysis of the 15 largest modules identified in the MN matrix.

A shortcoming of adaptive BRIM is that its performance degrades for large networks [108]. We propose a recursive algorithm based in [128] to find the optimal number of modules in the full cross-infection matrix. Following [128], we perform repeated divisions of the network until a local maximum of modularity is reached. The algorithm steps are: (i) find all the isolated network components and place them into separated modules, (ii) subdivide each module into $c = 2$ sub-modules using the standard BRIM algorithm and (iii) repeat the subdivision process until there is no improvement in the overall network modularity. The stop condition evaluates if the

modularity change ΔQ corresponding to the subdivision event in (ii) is significant or not. That is, $\Delta Q > 0$ means there is still room for further subdivisions. Newman suggests that is not correct to naively remove all edges falling between the subparts and apply the full modularity algorithm to each subpart in isolation [128]. We compute $\Delta Q > 0$ as the difference between the modularity value computed after and before the splitting event:

$$\Delta Q = \frac{1}{m}\left[\mathrm{Tr}\, R^{(g)^T} \hat{B}^{(g)} T^{(g)} - \mathrm{Tr}\, \hat{B}^{(g)}\right] \tag{23}$$

where $\hat{B}^{(g)}$ is the $h_g \times p_g$ bipartite modularity matrix of the h_g hosts and p_g pages within the module $g \subseteq G$, and $R^{(g)}$ and $T^{(g)}$ are the index vectors describing the splitting of the subgraph g in two sub-modules. Notice that we can restrict our computation to the subgraph g and thus, the index vectors are subsets of the full index vectors (see Equation 22). This is, to the best of our knowledge, the first time that the Newman's division algorithm has been applied to bipartite networks.

B.3 Multi-scale nested analysis

The MN matrix is significantly nested according to initial analysis using both the temperature calculator and NODF. This result is surprising giving the apparent lack of nestedness in visual inspection. However, prior work has noted that standard nestedness measures can signal spurious nested patterns when the network is comprised of nested modules [60]. In this context, Almeida-Neto and co-workers argue that we need specific models for distinct non-nested patterns because there is not an unique, working definition for the opposite of nestedness ("anti-nestedness") [6]. Here, we propose two new approaches (one for each nestedness measurement) to discard any interference of modular organization in the assessment of "true" nestedness.

We start by computing the modular organization of the full network G with our division algorithm (see Section B.2). The modules will constrain the space of possible matrix re-arrangements explored by the temperature calculator when searching for

the maximum nestedness (minimum temperature). In particular, our proposal for a constrained temperature calculator (i) permutes full modules (or matrix blocks), (ii) permutes rows and columns within a module, (iii) cannot perform any other permutation different from (i) and (ii). Still, the space of possible combinations can be quite large. We developed a heuristic algorithm that obtains good results with simple and deterministic sorting. First, we sort the rows and columns within any module in decreasing degree order (notice that rows and columns are sorted independently). Second, we rank modules according to the (sub-)matrix size and fill. The host (rows) ranking μ_g for the module $g \subset G$ is:

$$\mu_g = \frac{\sum_{i \in g} k_i}{h_g \times P} \tag{24}$$

where h_g is the number of hosts in the module g, k_i is the degree of the $i-$th host and P is the number of phages in the full network. Notice that this score can be seen as the connectance of a network composed of all phages presented in the entire network but only the hosts that belongs to module g. Similarly, there is a phage (columns) ranking ν_g for the module g:

$$\nu_g = \frac{\sum_{j \in g} d_j}{p_g \times H} \tag{25}$$

where p_g is the number of phages in the module g, d_j is the degree of the $j-th$ phage and H is the number of hosts in the full network.

In order to validate this measure of constrained nestedness, we have designed a theoretical experiment with synthetic networks having $2 \leq c \leq 50$ perfectly nested modules without interactions between them. Model networks have the same size as the MN network ($H = 286$, $P = 215$). Notice that $\mu_g = \mu$ and $\nu_g = \nu$ for all modules (blocks) because they have exactly the same size and fill. We place modules along the main diagonal to achieve optimal nestedness (see Figure 32). Every other arrangement (for example with off-diagonal blocks) yields sub-optimal nestedness values.

Our experiment confirms the initial hypothesis, i.e., unconstrained nestedness is

higher than constrained nestedness (see Figure 33). This suggests how high unconstrained nestedness of the MN matrix can be a consequence of its nested modular organization. As expected, we achieve maximum nestedness when the matrix is perfectly nested, e.g., there is only $c = 1$ module (see Figure 32 left). At $c = 2$ we have a sudden drop in (both constrained and unconstrained) nestedness because there are interactions below the isocline and absence of interactions above the isocline (see Figure 32 center). For small values of modularity ($c < 8$), the two null models have significantly lower values of constrained nestedness than the MN matrix. In general, nestedness increases with the number of modules ($c > 20$, see Figure 33) because temperature is directly related to the matrix filling (see Figure 32 right).

B.4 Geographical analysis

Both nestedness and modularity are topological, aspatial characteristics of bipartite networks. Here, we investigate the relationship between these network patterns and their spatial context. The MN matrix describes observed infections between host and phages sampled from a set of nearly equally-spaced, numbered stations in the Atlantic ocean. Here, we will review the original hypothesis of the MN study, i.e., to what extent geographical location drives the infection process. In the presence of strong spatial modularity, we should observe significant correlations between stations numbers (a surrogate of geographical location) of nodes within the same module. Otherwise, the geographical biodiversity will be very large.

We will use two different, standard metrics to measure the degree of geographical biodiversity in a topological module. For each module, we will compute the Shannon's entropy index:

$$H_k = -\sum_{i=1}^{R} \frac{n_i}{N} \log \frac{n_i}{N} \tag{26}$$

and the Simpson's diversity index:

$$D_k = 1 - \sum_{i=1}^{R} \frac{n_i(n_i - 1)}{N(N-1)} \tag{27}$$

where N are the number of different strains inside the module, R are the number of stations inside the module, and n_i are the number of strains from the $i-$th station. Low values in both indices indicate low geographical diversity within modules. Using a combination of two diversity indexes will provide additional support for our conclusions.

In order to test the NM hypothesis, we compare the observed diversity indexes (H_1, D_1), (H_2, D_2) ... (H_{15}, D_{15}) for the largest 15 modules found by the BRIM algorithm in the NM matrix (see above) with their expectations coming from an ensemble of 10^6 randomized matrices. We generate each sample by randomly permuting the row and column labels of the NM matrix. Once the random matrix is obtained, we will compare the diversity indexes of each observed module (H_k, D_k) with the pair of indices $(\tilde{H}_k, \tilde{D}_k)$ of random modules having the same size. Figure 34 indicates that, overall, the largest 15 modules display low geographical diversity, i.e., the observed value is lower than expected (considering a one-tailed p-value of 0.05 for statistical significance). This observation appears to be equally valid for hosts and phages (we have analyzed the two types of nodes separately), e.g., see Figure 34.

Table 13: Geographical data of microbial stations

Station	Latitude	Longitude	Station	Latitude	Longitude
454	47.717	-6.633	526	29.600	-57.083
456	44.750	-10.917	531	27.933	-57.733
458	43.200	-14.283	536	30.000	-58.333
460	41.350	-18.067	541	31.500	-59.667
462	39.650	-21.800	547	28.833	-59.633
464	38.000	-24.633	554	26.517	-60.233
465	37.817	-29.050	559	28.500	-61.000
469	37.967	-33.283	564	30.500	-61.000
471	37.333	-37.350	565	32.333	-64.633
472	36.550	-42.383	568	33.050	-59.983
474	35.717	-47.083	570	34.017	-55.317
476	34.867	-51.517	572	36.050	-42.467
478	34.017	-55.317	576	36.433	-39.067
480	33.217	-59.333	581	37.050	-34.350
484	32.567	-62.950	588	37.767	-26.367
489	31.967	-65.183	590	37.333	-22.033
492	30.667	-62.750	593	36.850	-17.417
497	28.783	-60.350	596	36.500	-13.000
501	27.117	-58.550	598	36.117	-8.717
504	26.100	-58.583	600	36.333	-7.467
508	26.417	-58.783	601	41.583	-10.333
513	29.617	-58.883	602	43.617	-9.567
518	31.200	-62.017	603	44.783	-8.833
522	31.067	-57.300	605	47.533	-6.283

Information that were extracted from the original Table 1 [123].

Table 14: Global properties of the extracted modules

Module	H	P	S	I	M	C	L_p	L_h
1	42	23	269	65	966	0.28	6.40	11.70
2	39	12	138	51	468	0.29	3.54	11.50
3	31	31	233	62	961	0.24	7.52	7.52
4	23	13	61	36	299	0.20	2.65	4.69
5	16	20	114	36	320	0.36	7.13	5.70
6	15	5	30	20	75	0.40	2.00	6.00
7	12	7	27	19	84	0.32	2.25	3.86
8	11	8	52	19	88	0.59	4.73	6.50
9	8	6	38	14	48	0.79	4.75	6.33
10	8	11	57	19	88	0.65	7.13	5.18
11	7	5	15	12	35	0.43	2.14	3.00
12	7	7	17	14	49	0.35	2.43	2.43
13	7	9	49	16	63	0.78	7.00	5.44
14	6	7	21	13	42	0.50	3.50	3.00
15	6	6	27	12	36	0.75	4.50	4.50
16	3	4	12	7	12	1.00	4.00	3.00
17	3	3	7	6	9	0.78	2.33	2.33
18	3	1	3	4	3	1.00	1.00	3.00
19	3	1	3	4	3	1.00	1.00	3.00
20	2	1	2	3	2	1.00	1.00	2.00
21	2	3	6	5	6	1.00	3.00	2.00
22	2	1	2	3	2	1.00	1.00	2.00
23	2	1	2	3	2	1.00	1.00	2.00
24	2	2	4	4	4	1.00	2.00	2.00
25	2	2	4	4	4	1.00	2.00	2.00
26	1	1	1	2	1	1.00	1.00	1.00
27	1	2	2	3	2	1.00	2.00	1.00
28	1	1	1	2	1	1.00	1.00	1.00
29	1	1	1	2	1	1.00	1.00	1.00
30	1	1	1	2	1	1.00	1.00	1.00
31	1	1	1	2	1	1.00	1.00	1.00
32	1	1	1	2	1	1.00	1.00	1.00
33	1	1	1	2	1	1.00	1.00	1.00
34	1	1	1	2	1	1.00	1.00	1.00
35	1	1	1	2	1	1.00	1.00	1.00
36	1	1	1	2	1	1.00	1.00	1.00
37	1	1	1	2	1	1.00	1.00	1.00
38	1	1	1	2	1	1.00	1.00	1.00
39	1	1	1	2	1	1.00	1.00	1.00
40	1	1	1	2	1	1.00	1.00	1.00
41	1	1	1	2	1	1.00	1.00	1.00
42	1	1	1	2	1	1.00	1.00	1.00
43	1	1	1	2	1	1.00	1.00	1.00
44	1	1	1	2	1	1.00	1.00	1.00
45	1	1	1	2	1	1.00	1.00	1.00
46	1	1	1	2	1	1.00	1.00	1.00
47	1	1	1	2	1	1.00	1.00	1.00
48	1	1	1	2	1	1.00	1.00	1.00
49	1	2	2	3	2	1.00	2.00	1.00
Average	5.84	4.39	24.88	10.22	75.41	0.83	2.29	2.75
Median	2.00	1.00	2.00	3.00	2.00	1.00	1.00	2.00

H : Number of hosts
P : Number of phages
$S = H + P$: Number of species
I : Number of interactions
$M = HP$: Size
$C = I/M$: Connectance or fill
$L_p = I/P$: Mean phage degree (Average number of susceptible hosts by phage)
$L_h = I/H$: Mean host degree (Average number of virulent viruses by host)

Table 15: Geographical biodiversity indexes

Module	Phages		Hosts	
	Simpson	Shannon	Simpson	Shannon
1	0.953 ($p = 0.086$)	2.487 ($p = 0.040$)	0.970 ($p = 0.272$)	3.048 ($p = 0.221$)
2	0.939 ($p = 0.065$)	2.095 ($p = 0.081$)	0.964 ($p = 0.093$)	2.908 ($p = 0.048$)
3	0.897 ($p = 0.000$)	2.179 ($p = 0.000$)	0.920 ($p = 0.000$)	2.551 ($p = 0.001$)
4	0.808 ($p = 0.000$)	1.479 ($p = 0.000$)	0.909 ($p = 0.000$)	2.198 ($p = 0.000$)
5	0.816 ($p = 0.000$)	1.817 ($p = 0.000$)	0.825 ($p = 0.000$)	1.689 ($p = 0.000$)
6	1.000 ($p = 0.280$)	1.609 ($p = 0.280$)	0.962 ($p = 0.158$)	2.396 ($p = 0.227$)
7	0.714 ($p = 0.000$)	1.004 ($p = 0.000$)	0.833 ($p = 0.000$)	1.517 ($p = 0.000$)
8	0.857 ($p = 0.004$)	1.494 ($p = 0.010$)	0.909 ($p = 0.012$)	1.846 ($p = 0.011$)
9	0.333 ($p = 0.000$)	0.451 ($p = 0.000$)	1.000 ($p = 0.552$)	2.079 ($p = 0.552$)
10	0.909 ($p = 0.020$)	1.768 ($p = 0.005$)	0.893 ($p = 0.013$)	1.667 ($p = 0.027$)
11	0.900 ($p = 0.025$)	1.332 ($p = 0.025$)	0.857 ($p = 0.005$)	1.475 ($p = 0.007$)
12	0.952 ($p = 0.111$)	1.748 ($p = 0.111$)	1.000 ($p = 0.453$)	1.946 ($p = 0.453$)
13	0.889 ($p = 0.010$)	1.677 ($p = 0.013$)	0.857 ($p = 0.006$)	1.475 ($p = 0.008$)
14	0.571 ($p = 0.000$)	0.683 ($p = 0.000$)	0.533 ($p = 0.000$)	0.637 ($p = 0.000$)
15	0.600 ($p = 0.000$)	0.868 ($p = 0.000$)	0.733 ($p = 0.001$)	1.011 ($p = 0.001$)

Small values means low geographical biodiversity. $p < 0.05$ means the module is statistically no geographically diverse. p-values were calculated as the ratio of random permutations index values that are smaller than the real index. See Equation 4 in Chapter 3 for a mathematical description of these indexes.

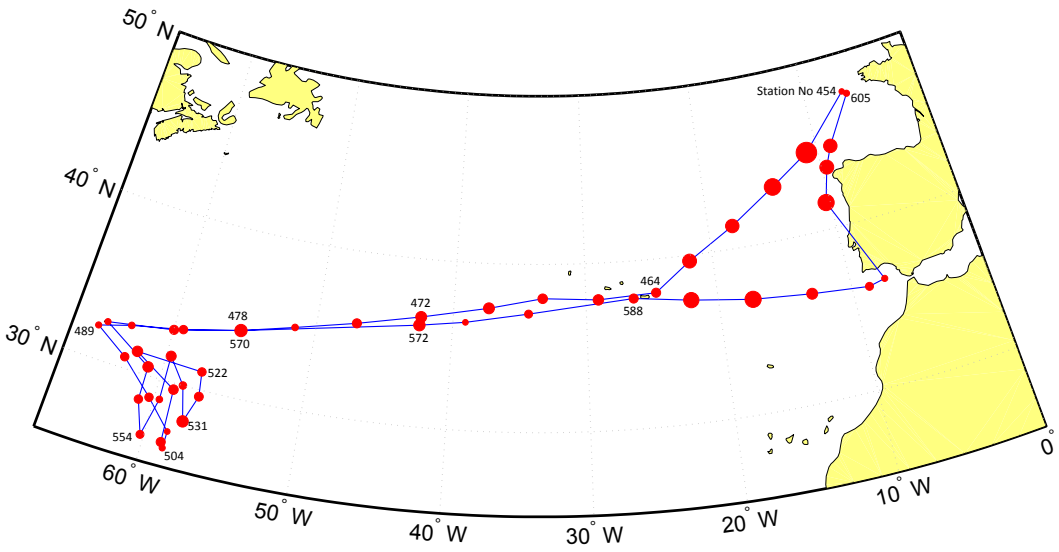

Figure 28: Originally appeared as Figure 1 on [123] with the label *Track of RV "Friedrich Heincke" in the Atlantic Ocean during cruise no. 160 and microbial stations.* Here, each circle represents the geographic location of each station. The radius of the circles corresponds linearly to the number of strains that were extracted in the corresponding station. Some number stations are indicated in order to clarify the direction of the route. Increasing station number indicate the order of visit.

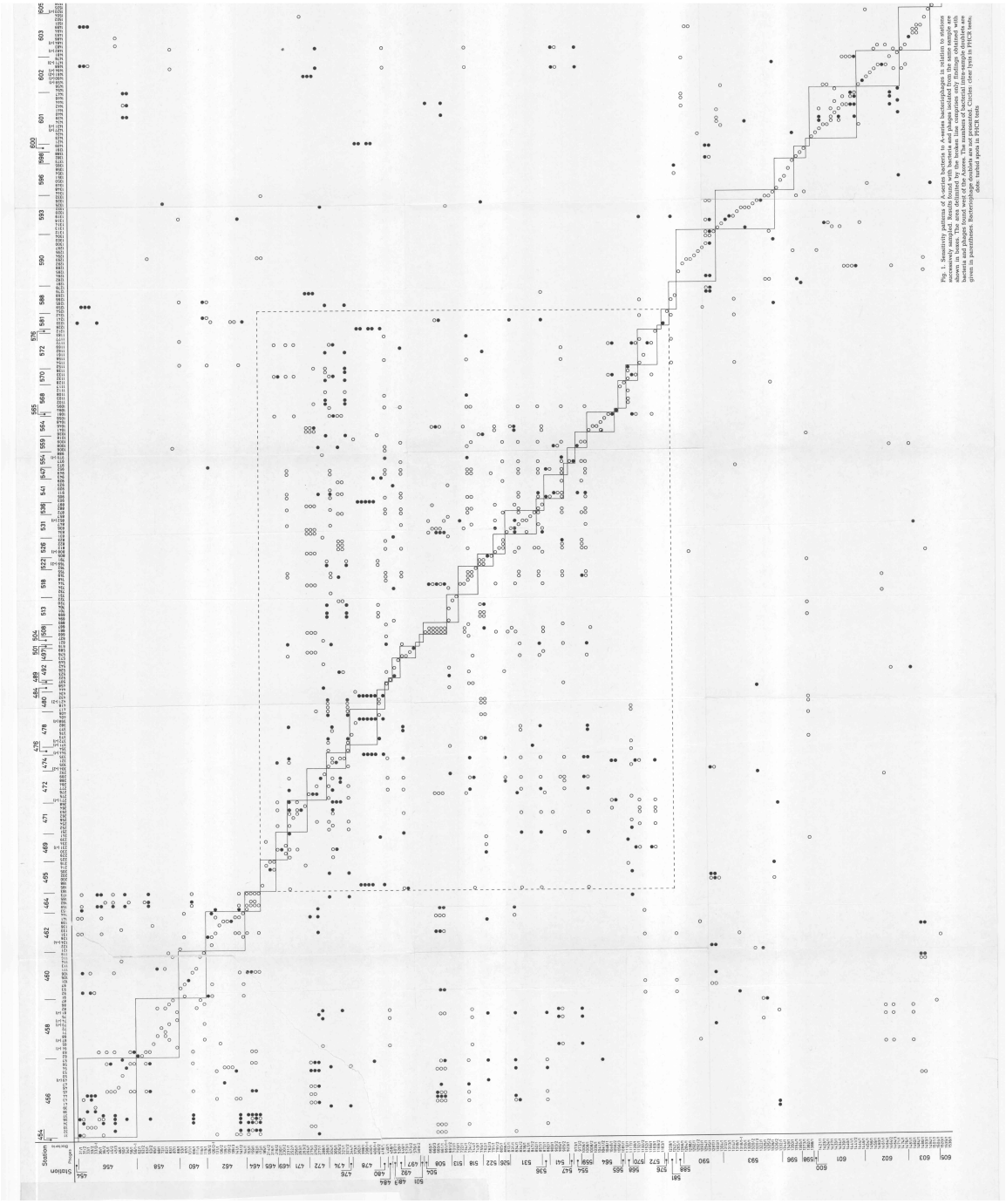

Figure 29: Moebus & Nattkemper [124] cross-reaction test in the Atlantic Ocean region. This matrix is subdivided in different stations, where each square delimits the infections inside strains of the same station. The original label reads: *"Fig 1. Sensitivity patterns of A-seres bacteria to A-series bacteriophages in relation to stations successively sampled. Results found with bacteria and phages isolated from the same sample are shown in boxes. The area delimited by the broken line comprises only findings obtained with bacteria and phages found west of the Azores. The numbers of bacteria intra-sample doublets are given in parentheses. Bacteriophage doublets are not presented. Circles: clear lysis in PHCR tests; dots: turbid spots in PHCR tests.".*

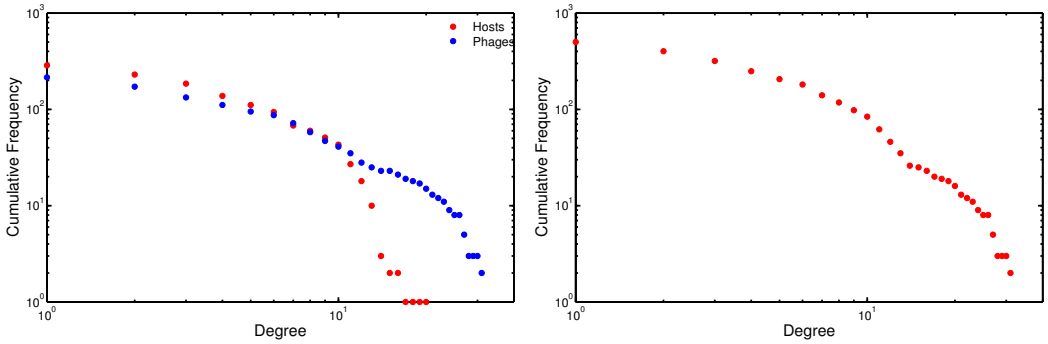

Figure 30: Cumulative degree frequency of the MN matrix. **a)** Cumulative frequency of the MN matrix with distinction between host and phage nodes. **b)** Cumulative frequency of the MN matrix without distinction between host and phage nodes. Both phages and hosts have a wide range of degree values, in which small degree values are more likely to occur than large degree values.

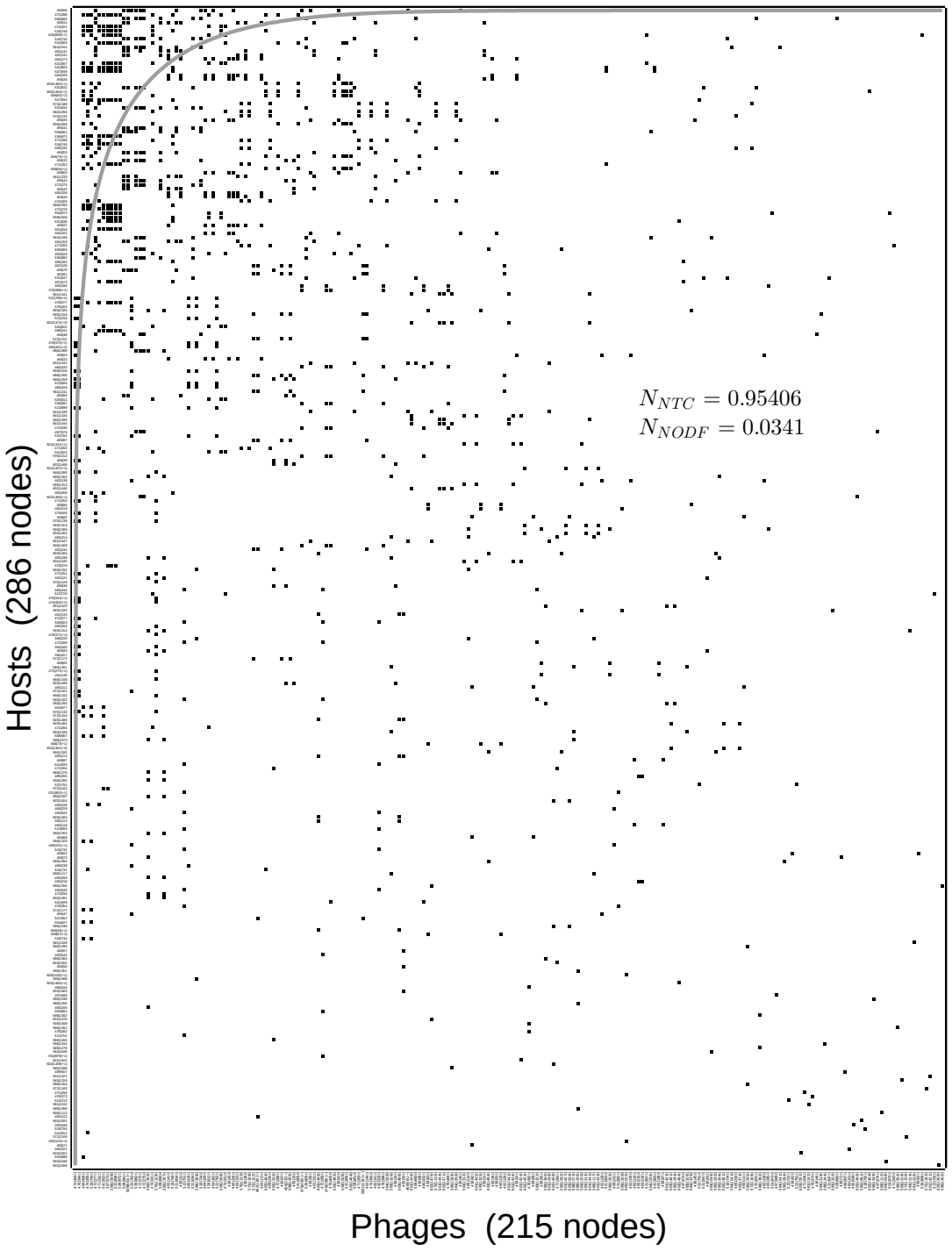

Figure 31: Arrangement of the cross-infection matrix produced with the NTC algorithm. While the nestedness value $N_{NTC} = 0.95$ has a p-value $< 10^{-5}$ in both null models, the nestedness value $N_{NODF} = 0.0341$ has a p-value $< 10^{-5}$ only in the Bernoulli random null model (see text).

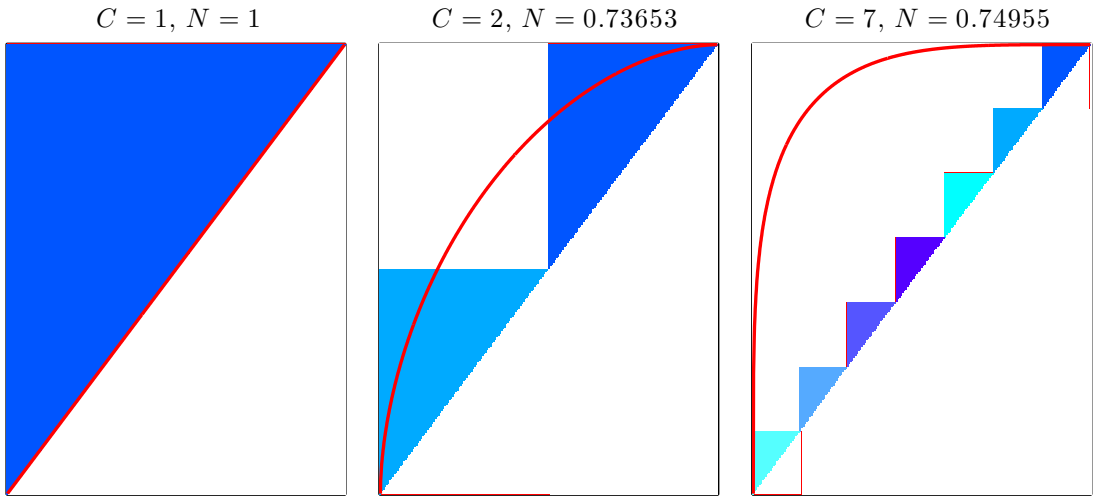

Figure 32: From left to right, correlation between nestedness and modularity in synthetic networks with $c = 1, 2, 7$ perfectly nested modules. Bold red line represents the isocline of perfect nestedness (see material and methods in Chapter 3). Blocks with red outlines indicate modules.

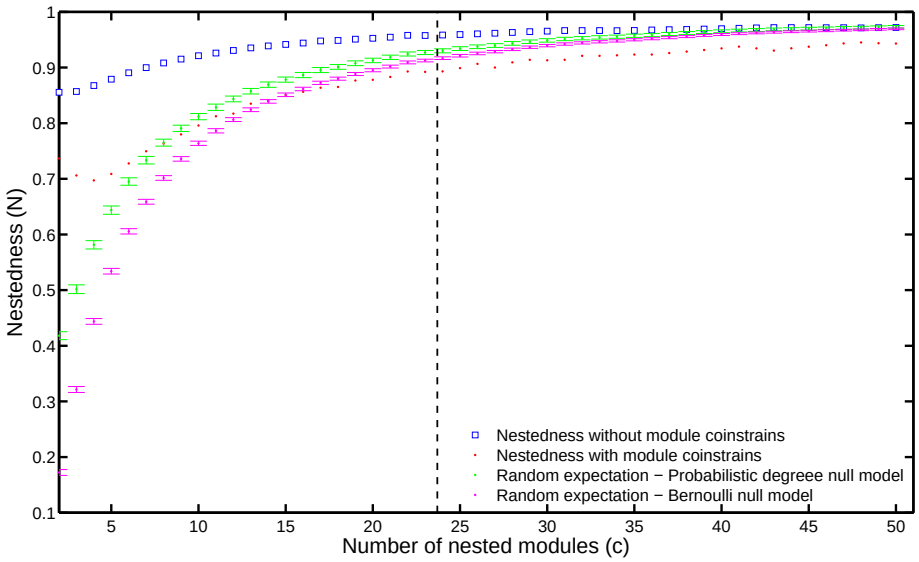

Figure 33: Comparison of constrained vs unconstrained temperature. We analyze synthetic networks with perfect nestedness with varying number of modules $2 \leq c \leq 50$ (see text). The vertical line indicate where the fill of the MN matrix coincides with that of the synthetic networks. Notice that for the corresponding fill, the nestedness of the two random expectations are larger than the value of nestedness with module constraints.

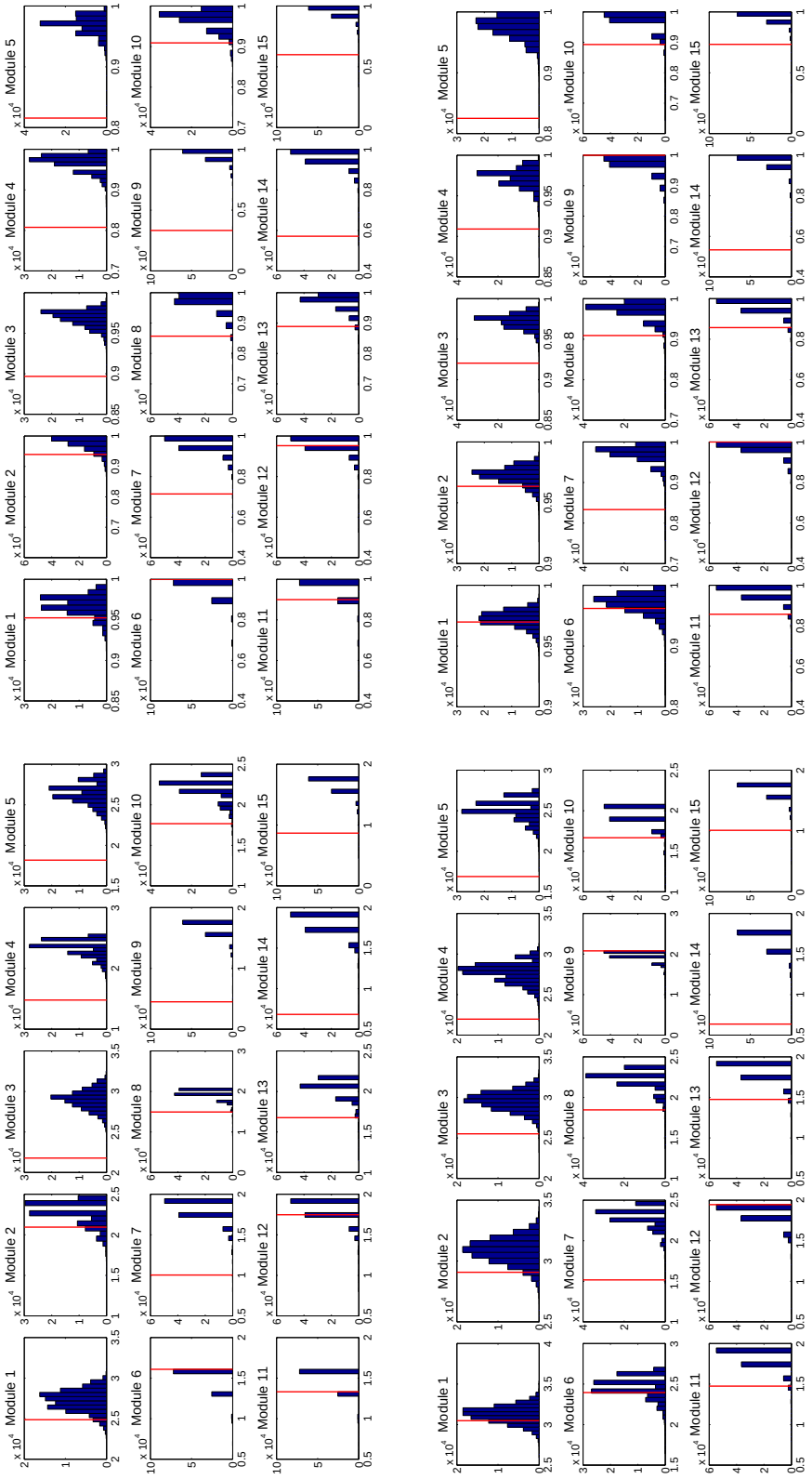

Figure 34: Distribution of geographical diversity for the 15 biggest modules. The index represent the module index. The red lines represent the real geographical diversity value of those modules. **a)** Simpson's index distribution for phages. **b)** Simpson's index distribution for hosts. **c)** Shannon's index distribution for phages. **d)** Shannon's index distribution for hosts.

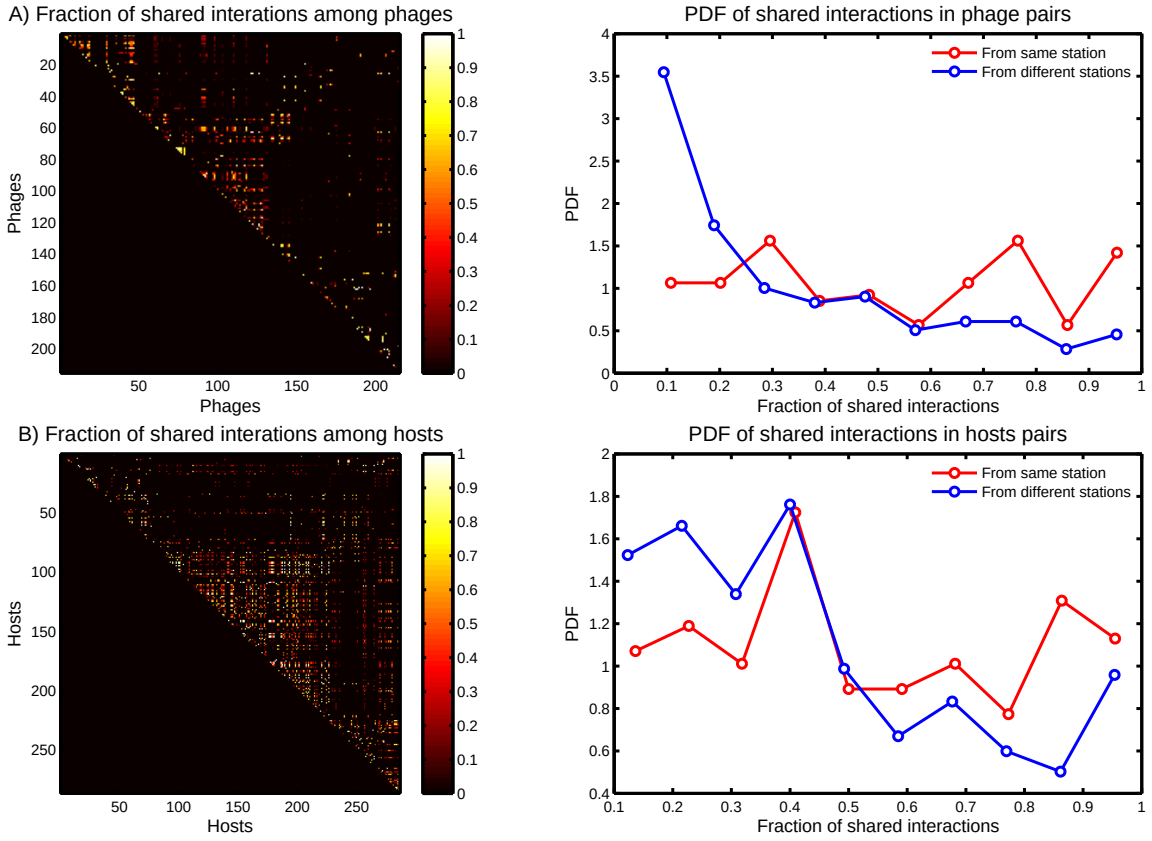

Figure 35: Fraction of shared interactions across pair of nodes. The top shows phage species and the bottom shows host species. The left shows the fraction of shared interactions across every pair of nodes. The right shows the probability density function of shared interaction between pair of nodes given that the pairs shared at least one interaction.

REFERENCES

[1] Abe, M., Izumoji, Y., and Tanji, Y., "Phenotypic transformation including host-range transition through superinfection of t-even phages," *FEMS Microbiology Letters*, vol. 269, pp. 145–52, 2007.

[2] Abedon, S. T., *Bacteriophage ecology: population growth, evolution and impact of bacterial viruses*. Cambridge, UK: Cambridge University Press.

[3] Adams, M. H. and others, "Bacteriophages," *Bacteriophages*, 1959.

[4] Agrawal, A. and Lively, C. M., "Infection genetics: gene-for-gene versus matching-alleles models and all points in between," *Evolutionary Ecology Research*, vol. 4, pp. 79–90, 2002.

[5] Allesina, S., Alonso, D., and Pascual, M., "A general model for food web structure," *Science*, vol. 320, pp. 658–661, 2008.

[6] Almeida-Neto, M., Guimarães, P. R., and Lewinsohn, T. M., "On nestedness analyses: rethinking matrix temperature and anti-nestedness," *Oikos*, vol. 116, pp. 716–722, 2007.

[7] Almeida-Neto, M., Guimaraes, P., Guimarães, P. R., Loyola, R. D., and Ulrich, W., "A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement," *Oikos*, vol. 117, no. 8, pp. 1227–1239, 2008.

[8] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[9] Andersson, A. F. and Banfield, J. F., "Virus population dynamics and acquired virus resistance in natural microbial communities," *Science*, vol. 320, no. 5879, pp. 1047–1050, 2008.

[10] Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A., and Rohwer, F., "The marine viromes of four oceanic regions," *PLoS Biology*, vol. 4, pp. 2121–2131, 2006.

[11] Atmar, W. and Patterson, B. D., "The measure of order and disorder in the distribution of species in fragmented habitat," *Oecologia*, vol. 96, pp. 373–382, 1993.

[12] Barabási, A.-L. and Albert, R., "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[13] Barber, M., "Modularity and community detection in bipartite networks," *Physical Review E*, vol. 76, p. 066102, 2007.

[14] Barrangou, R., Yoon, S. S., F., Fleming, H. P., and Klaenhammer, T. R., "Characterization of six *Leuconostoc fallax* bacteriophages isolated from an industrial sauerkraut fermentation," *Applied and Environmental Microbiology*, vol. 68, pp. 5452–5452, 2002.

[15] Bascompte, J., "Disentangling the web of life," *Science*, vol. 325, pp. 416–419, 2009.

[16] Bascompte, J. and Jordano, P., *The structure of plant-animal mutualistic networks*, pp. 143–159. Oxford: Oxford University Press, 2006.

[17] Bascompte, J. and Jordano, P., "Plant-animal mutualistic networks: the architecture of biodiversity," *Annu. Rev. Ecol. Evol. Syst.*, vol. 38, pp. 567–593, 2007.

[18] Bascompte, J., Jordano, P., Melián, C. J., and Olesen, J. M., "The nested assembly of plant-animal mutualistic networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 9383–9387, 2003.

[19] Bastian, M., Heymann, S., and Jacomy, M., "Gephi: an open source software for exploring and manipulating networks.," in *ICWSM*, 2009.

[20] Bastolla, U., Fortuna, M. A., Pascual-Garcia, A., Ferrera, A., Luque, B., and Bascompte, J., "The architecture of mutualistic networks minimizes competition and increases biodiversity," *Nature*, vol. 458, pp. 1018–1091, 2009.

[21] Beckett, S. J. and Williams, H. T., "Coevolutionary diversification creates nested-modular structure in phage–bacteria interaction networks," *Interface focus*, vol. 3, no. 6, p. 20130033, 2013.

[22] Bohannan, B. J. M. and Lenski, R. E., "Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage," *Ecology Letters*, vol. 3, pp. 362–377, 2000.

[23] Boyd, E. F., Davis, B. M., and Hochhut, B., "Bacteriophage-bacteriophage interactions in the evolution of pathogenic bacteria," *Trends in Microbiology*, vol. 9, pp. 137–144, 2001.

[24] Braun-Breton, C. and Hofnung, M., "In vivo and in vitro functional alterations of the bacteriophage lambda receptor in *lamB* missense mutants of *Escherichia coli* k-12," *Journal of Bacteriology*, vol. 148, pp. 812–845, 1981.

[25] Breitbart, M., "Marine viruses: Truth or dare," in *ANNUAL REVIEW OF MARINE SCIENCE, VOL 4* (Carlson, C. and Giovannoni, S., eds.), vol. 4 of *Annual Review of Marine Science*, pp. 425–448, ANNUAL REVIEWS, 2012.

[26] Breitbart, M., Miyake, J. H., and Rohwer, F., "Global distribution of nearly identical phage-encoded dna sequences," *FEMS microbiology letters*, vol. 236, no. 2, pp. 249–256, 2004.

[27] Brussaard, C. P., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K., Paul, J. H., and others, "Global-scale processes with a nanoscale drive: the role of marine viruses," *The ISME journal*, vol. 2, no. 6, pp. 575–578, 2008.

[28] Buckling, A. and Rainey, P. B., "Antagonistic coevolution between a bacterium and a bacteriophage," *Proceedings of the Royal Society of London Series B-Biologial Sciences*, vol. 269, pp. 931–936, 2002.

[29] Campbell, J. I. A., Albrechtsen, M., and Sø rensen, J., "Large *Pseudomonas* phages isolated from barley rhizosphere," *FEMS Microbiology Ecology*, vol. 18, pp. 63–74, 2006.

[30] Canard, E. F., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D., and Gravel, D., "Empirical evaluation of neutral interactions in host-parasite networks.," *The American Naturalist*, vol. 183, no. 4, pp. pp. 468–479, 2014.

[31] Capparelli, R., Nocerino, N., Iannaccone, M., Ercolini, D., Parlato, M., Chiara, M., and Iannelli, D., "Bacteriophage therapy of *Salmonella enterica*: a fresh appraisal of bacteriophage therapy," *The Journal of Infectious Diseases*, vol. 201, pp. 52–61, 2010.

[32] Caso, J. L., G., C., Herrero, M., Montilla, A., Rodriguez, A., and Suarez, J. E., "Isolation and characterization of temperate and virulent bacteriophages of *Lactobacillus plantarum*," *Journal of Dairy Science*, vol. 78, pp. 741–750, 1995.

[33] Ceyssens, P.-J., Noben, J.-P., Ackermann, H.-W., Verhaegen, J., Daniel, Pirnay, J.-P., Merabishvili, M., Vaneechoutte, M., Chibeu, A., Volckaert, G., and Lavigne, R., "Survey of *Pseudomonas aeruginosa* and its phages: *de novo* peptide sequencing as a novel tool to assess the diversity of worldwide collected viruses," *Environmental Microbiology*, vol. 11, pp. 1303–1313, 2009.

[34] Chartrand, G., "Introductory graph theory. 1985."

[35] Childs, L. M., Held, N. L., Young, M. J., Whitaker, R. J., and Weitz, J. S., "multiscale model of crispr-induced coevolutionary dynamics: diversification at the interface of lamarck and darwin," *Evolution*, vol. 66, no. 7, pp. 2015–2029, 2012.

[36] Chiura, H. X., "Generalized gene transfer by virus-like particles from marine bacteria," *Aquatic Microbial Ecology*, vol. 13, pp. 75–83, 1997.

[37] Cohen, J. E., Briand, F., and Newman, C. M., *Community food webs: data and theory*. Berlin: Springer-Verlag, 1990.

[38] Cohen, J. E. and Newman, C. M., "A stochastic theory of community food webs i. models and aggregated data," *Proc. R. Soc. Lond. B*, vol. 224, pp. 421–448, 1985.

[39] Cohen, J. E., Newman, C. M., and Briand, F., "A stochastic theory of community food webs ii. individual webs," *Proc. R. Soc. Lond. B*, vol. 224, pp. 4449–4461, 1985.

[40] Cohen, J. E., *Food webs and niche space*. No. 11, Princeton University Press, 1978.

[41] Comeau, A. M., Buenaventura, E., and Suttle, C. A., "A persistent, productive, and seasonally dynamic vibriophage population within pacific oysters (*Crassostrea gigas*)," *Applied and Environmental Microbiology*, vol. 71, pp. 5324–5324, 2005.

[42] Comeau, A. M., Chan, A. M., and Suttle, C. A., "Genetic richness of vibriophages isolated in a coastal environment," *Environmental Microbiology*, vol. 8, pp. 1164–1164, 2006.

[43] Cox-Foster, D. L. L., Conlan, S., Holmes, E. C. C., Palacios, G., Evans, J. D. D., Moran, N. A. A., Quan, P.-L. L., Briese, T., Hornig, M., Geiser, D. M. M., Martinson, V., Vanengelsdorp, D., Kalkstein, A. L. L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S. K. K., Simons, J. F. F., Egholm, M., Pettis, J. S. S., and Lipkin, W. I. I., "A metagenomic survey of microbes in honey bee colony collapse disorder.," *Science*, vol. 318, pp. 283–287, 2007.

[44] Csardi, G. and Nepusz, T., "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, 2006.

[45] Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S., and Kim, J. F., "Tracing ancestors and relatives of escherichia coli b, and the derivation of b strains rel606 and bl21(de3)," *Journal of Molecular Biology*, vol. 394, no. 4, pp. 634 – 643, 2009.

[46] Deng, L., Gregory, A., Yilmaz, S., Poulos, B. T., Hugenholtz, P., and Sullivan, M. B., "Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging," *mBio*, vol. 3, no. 6, pp. e00373–12, 2012.

[47] DePaola, A., Motes, M. L., Chan, A. M., and Suttle, C. A., "Phages infecting vibrio vulnificus are abundant and diverse in oysters (*Crassostrea virginica*) collected from the gulf of mexico," *Applied and Environmental Microbiology*, vol. 64, pp. 346–351, 1998.

[48] Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L., Chau, B., and others, "Biodiversity and biogeography of phages in modern stromatolites and thrombolites," *Nature*, vol. 452, no. 7185, pp. 340–343, 2008.

[49] Doi, K., Zhang, Y., Nishizaki, Y., Umeda, A., Ohmomo, S., and Ogata, S., "A comparative study and phage typing of silage-making *Lactobacillus* bacteriophages," *Journal of Bioscience and Bioengineering*, vol. 95, pp. 518–525, 2003.

[50] Dormann, C. F., Frueund, J., Bluethgen, N., and Gruber, B., "Indices, graphs and null models: analyzing bipartite ecological networks.," *The Open Ecology Journal*, vol. 2, pp. 7–24, 2009.

[51] Dormann, C. F. and Strauss, R., "A method for detecting modules in quantitative bipartite networks," *Methods in Ecology and Evolution*, vol. 5, no. 1, pp. 90–98, 2014.

[52] Dunne, J. A., "The network structure of food webs," in *Ecological Networks: Linking Structure to Dynamics in Food Webs*, pp. 27–86, Oxford University Press, 2006.

[53] Dunne, J. A., Williams, R. J., and Martinez, N. D., "Food-web structure and network theory: the role of connectance and size," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12917–12922, 2002.

[54] Duplessis, M. and Moineau, S., "Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages," *Molecular Microbiology*, vol. 41, pp. 325–336, 2001.

[55] Edwards, R. A. and Rohwer, F., "Viral metagenomics," *Nat. Rev. Microb.*, vol. 3, pp. 504–510, 2005.

[56] Erdös, P. and Rényi, A., "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[57] Faruque, S. M., Islam, M. J., Ahmad, Q. S., Faruque, A. S., Sack, D. A., Nair, G. B., and Mekalanos, J. J., "Self-limiting nature of seasonal

cholera epidemics: Role of host-mediated amplification of phage," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 6119–6124, 2005.

[58] Faruque, S. M., Naser, I. B., Islam, M. J., Faruque, A. S. G., Ghosh, A. N., Nair, G. B., Sack, D. A., and Mekalanos, J. J., "Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages," *Proceedings of the National Academy of Sciences, USA*, vol. 102, pp. 1702–1707, 2005.

[59] Fischer, J. and Lindenmayer, D. B., "Treating the nestedness temperature calculator as a "black box" can lead to false conclusions," *Oikos*, vol. 99, no. 1, pp. 193–199, 2002.

[60] Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., and Weitz, J. S., "Statistical structure of host–phage interactions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 28, pp. E288–E297, 2011.

[61] Flores, C. O., Poisot, T., Valverde, S., and Weitz, J. S., "Bimat: a matlab (r) package to facilitate the analysis and visualization of bipartite networks," *arXiv preprint arXiv:1406.6732*, 2014.

[62] Flores, C. O., Valverde, S., and Weitz, J. S., "Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages," *The ISME journal*, vol. 7, no. 3, pp. 520–532, 2012.

[63] Forde, S. E., Beardmore, R. E., Gudelj, I., Arkin, S. S., Thompson, J. N., and Hurst, L. D., "Understanding the limits to generalizability of experimental evolutionary models," *Nature*, vol. 455, pp. 220–223, 2008.

[64] Fortuna, M. A., Stouffer, D. B., Olesen, J. M., Jordano, P., Mouillot, D., Krasnov, B. R., Poulin, R., and Bascompte, J., "Nestedness versus modularity in ecological networks: two sides of the same coin?," *Journal of Animal Ecology*, vol. 79, no. 4, pp. 811–817, 2010.

[65] Fortunato, S., "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[66] Fortunato, S. and Barthélemy, M., "Resolution limit in community detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 36–41, 2007.

[67] Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., and Hanage, W. P., "The bacterial species challenge: making sense of genetic and ecological diversity," *Science*, vol. 323, pp. 741–746, 2009.

[68] Fuhrman, J. A. and Noble, R. T., "Viruses and protists cause similar bacterial mortality in coastal seawater," *Limnology and Oceanography*, vol. 40, pp. 1236–1242, 1995.

[69] Galeano, J., Pastor, J. M., and Iriondo, J. M., "Weighted-interaction nestedness estimator (wine): A new estimator to calculate over frequency matrices," *Environmental Modelling & Software*, vol. 24, no. 11, pp. 1342–1346, 2009.

[70] Gamage, S. D., Patton, A. K., Hanson, J. F., and Weiss, A. A., "Diversity and host range of shiga toxin-encoding phage," *Infection and Immunity*, vol. 72, pp. 7131–7131, 2004.

[71] Garey, M. and Johnson, D., "Crossing number is np-complete," *SIAM Journal on Algebraic Discrete Methods*, vol. 4, no. 3, pp. 312–316, 1983.

[72] Genini, J., Morellato, L. P. C., Guimarães, P. R., and Olesen, J. M., "Cheaters in mutualism networks," *Biology Letters*, vol. 6, pp. 494–497, 2010.

[73] Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E., "Metagenomic analysis of the human distal gut microbiome," *Science*, vol. 312, pp. 1355–1359, 2006.

[74] Gobler, C. J., Hutchins, D. A., Fisher, N. S., Cosper, E. M., and Sañudo Wilhelmy, S., "Release and bioavailability of c, n, p, se, and fe following viral lysis of a marine chrysophyte," *Limnology and Oceanography*, vol. 42, pp. 1492–1504, 1997.

[75] Gómez, P. and Buckling, A., "Bacteria-phage antagonistic coevolution in soil," *Science*, vol. 332, pp. 106–109, 2011.

[76] Goodridge, L. and Abedon, S. T., "Bacteriophage biocontrol and bioprocessing: application of phage therapy to industry," *SIM News*, vol. 53, pp. 254–262, 2003.

[77] Goodridge, L., Gallaccio, A., and Griffiths, M. W., "Morphological, host range, and genetic characterization of two coliphages," *Applied and Environmental Microbiology*, vol. 69, pp. 5364–5364, 2003.

[78] Gudelj, I., Weitz, J. S., Ferenci, T., Claire Horner-Devine, M., Marx, C. J., Meyer, J. R., and Forde, S. E., "An integrative approach to understanding microbial diversity: from intracellular mechanisms to community structure," *Ecology Letters*, vol. 13, pp. 1073–1084, 2010.

[79] Guimaraes Jr, P. R. and Guimarães, P., "Improving the analyses of nestedness for large sets of matrices," *Environmental Modelling & Software*, vol. 21, no. 10, pp. 1512–1513, 2006.

[80] Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N., "Modularity from fluctuations in random graphs and complex networks," *Phys. Rev. E*, vol. 70, p. 025101, Aug 2004.

[81] Haerter, J. O., Mitarai, N., and Sneppen, K., "Phage and bacteria support mutual diversity in a narrowing staircase of coexistence," *The ISME journal*, 2014.

[82] Hagberg, A., Swart, P., and S Chult, D., "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Laboratory (LANL), 2008.

[83] Hansen, V. M., Rosenquist, H., Baggesen, D. L., Brown, S., and Christensen, B. B., "Characterization of *Campylobacter* phages including analysis of host range by selected *Campylobacter* penner serotypes," *BMC Microbiology*, vol. 7, pp. 90–90, 2007.

[84] Harfoot, M. B., Newbold, T., Tittensor, D. P., Emmott, S., Hutton, J., Lyutsarev, V., Smith, M. J., Scharlemann, J. P., and Purves, D. W., "Emergent global patterns of ecosystem structure and function from a mechanistic general ecosystem model," *PLoS biology*, vol. 12, no. 4, p. e1001841, 2014.

[85] Held, N. L. and Whitaker, R. J., "Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes," *Environmental microbiology*, vol. 11, no. 2, pp. 457–466, 2009.

[86] Holmfeldt, K., Middelboe, M., Nybroe, O., and Riemann, L., "Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their *Flavobacterium* hosts," *Applied and Environmental Microbiology*, vol. 73, pp. 6730–6739, 2007.

[87] Horvath, P. and Barrangou, R., "Crispr/cas, the immune system of bacteria and archaea," *Science*, vol. 327, pp. 167–170, 2010.

[88] Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L., "Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing," *PLoS Genet*, vol. 4, pp. e1000255+, 2008.

[89] Hyman, P. and Abedon, S. T., "Bacteriophage host range and bacterial resistance," *Advances in Applied Microbiology*, vol. 70, pp. 217–248, 2010.

[90] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N., "Revealing modular organization in the yeast transcriptional network," *Nature genetics*, vol. 31, no. 4, pp. 370–377, 2002.

[91] Jaccard, P., "The distribution of the flora in the alpine zone. 1," *New Phytologist*, vol. 11, pp. 37–50, 1912.

[92] Jensen, E. C., Schrader, H. S., Rieland, B., Thompson, T. L., Lee, K. W., Nickerson, K. W., and Kokjohn, T. A., "Prevalence of broad-host-range lytic bacteriophages of sphaerotilus natans, escherichia coli, andpseudomonas aeruginosa," *Applied and Environmental Microbiology*, vol. 64, no. 2, pp. 575–580, 1998.

[93] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L., "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[94] Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., Weinbauer, M. G., Luo, T., Chen, F., and others, "Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean," *Nature Reviews Microbiology*, vol. 8, no. 8, pp. 593–599, 2010.

[95] Joppa, L. N., Bascompte, J., Montoya, J. M., Sole, R. V., Sanderson, J., and Pimm, S. L., "Reciprocal specialization in ecological networks," *Ecology letters*, vol. 12, no. 9, pp. 961–969, 2009.

[96] Jover, L. F., Cortez, M. H., and Weitz, J. S., "Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks," *Journal of theoretical biology*, vol. 332, pp. 65–77, 2013.

[97] Kankila, J. and Lindstrom, K., "Host range, morphology and dna restriction patterns of bacteriophage isolates infecting *Rhizobium leguminosarum* bv. *trifolii*," *Soil Biology and Biochemistry*, vol. 26, pp. 429–437, 1994.

[98] Karginov, F. V. and Hannon, G. J., "The crispr system: small rna-guided defense in bacteria and archaea," *Molecular Cell*, vol. 37, pp. 7–19, 2010.

[99] Kernighan, B. W. and Lin, S., "An efficient heuristic procedure for partitioning graphs," *Bell system technical journal*, vol. 49, no. 2, pp. 291–307, 1970.

[100] Koskella, B., Thompson, J. N., Preston, G. M., and Buckling, A., "Local biotic environment shapes the spatial scale of bacteriophage adaptation to bacteria," *The American Naturalist*, vol. 177, no. 4, pp. 440–451, 2011.

[101] Krylov, V. N., Miller, S., Rachel, R., Biebl, M., Pleteneva, E. a., Schuetz, M., Krylov, S. V., and Shaburova, O. V., "Ambivalent bacteriophages of different species active on *Escherichia coli* k12 and *Salmonella* sp. strains," *Russian Journal of Genetics*, vol. 42, pp. 106–114, 2006.

[102] Kudva, I. T., Jelacic, S., Tarr, P. I., Youderian, P., and Hovde, C. J., "Biocontrol of *Escherichia coli* o157 with o157-specific bacteriophages," *Applied and Environmental Microbiology*, vol. 65, pp. 3767–3773, 1999.

[103] Labrie, S. J., Samson, J. E., and Moineau, S., "Bacteriophage resistance mechanisms," *Nature Reviews Microbiology*, vol. 8, pp. 317–327, 2010.

[104] Langley, R., "Lysogeny and bacteriophage host range within the *Burkholderia cepacia* complex," *Journal of Medical Microbiology*, vol. 52, pp. 483–490, 2003.

[105] Lenski, R. E. and Levin, B. R., "Constraints on the coevolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities," *American Naturalist*, vol. 125, pp. 585–602, 1985.

[106] Levin, B. R. and Bull, J. J., "Population and evolutionary dynamics of phage therapy," *Nature Reviews Microbiology*, vol. 2, pp. 166–173, 2004.

[107] Levin, S. A., "The problem of pattern and scale in ecology: the robert h. macarthur award lecture," *Ecology*, vol. 73, no. 6, pp. 1943–1967, 1992.

[108] Liu, X. and Murata, T., "Community detection in large-scale bipartite networks," in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, vol. 1, pp. 50–57, IET, 2009.

[109] MacQueen, J. and others, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, California, USA, 1967.

[110] Marston, M. F., Pierciey, F. J., Shepard, A., Gearin, G., Qi, J., Yandava, C., Schuster, S. C., Henn, M. R., and Martiny, J. B., "Rapid diversification of coevolving marine *Synechococcus* and a virus," *Proceedings of the National Academy of Sciences*, vol. 109, no. 12, pp. 4544–4549, 2012.

[111] Martinez, N. D., Hawkins, B. A., Dawah, H. A., and Feifarek, B. P., "Effects of sampling effort on characterization of food-web structure," *Ecology*, vol. 80, no. 3, pp. 1044–1055, 1999.

[112] Martiny, J. B. H., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., and others, "Microbial biogeography: putting microorganisms on the map," *Nature Reviews Microbiology*, vol. 4, no. 2, pp. 102–112, 2006.

[113] Maynard, N. D., Birch, E. W., Sanghvi, J. C., Chen, L., Gutschow, M. V., and Covert, M. W., "A forward-genetic screen and dynamic analysis of lambda phage host-dependencies reveals an extensive interaction network and a new anti-viral strategy," *PLoS Genetics*, vol. 6, p. e1001017, 2010.

[114] McLaughlin, M. R. and King, R. A., "Characterization of *Salmonella* bacteriophages isolated from swine lagoon effluent," *Current Microbiology*, vol. 56, pp. 208–213, 2008.

[115] Memmott, J., "The structure of a plant-pollinator food web," *Ecology Letters*, vol. 2, no. 5, pp. 276–280, 1999.

[116] Mestres, J., Gregori-Puigjane, E., Valverde, S., and Sole, R. V., "Data completeness–the achilles heel of drug-target networks," *Nature Biotechnology*, vol. 26, pp. 983–984, 2008.

[117] Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., and Lenski, R. E., "Repeatability and contingency in the evolution of a key innovation in phage lambda," *Science*, vol. 335, no. 6067, pp. 428–432, 2012.

[118] Middelboe, M., Holmfeldt, K., Riemann, L., Nybroe, O., and Haaber, J., "Bacteriophages drive strain diversification in a marine *Flavobacterium*: implications for phage resistance and physiological properties," *Environmental Microbiology*, vol. 11, pp. 1971–1982, 2009.

[119] Middelboe, M. and Lyck, P. G., "Regeneration of dissolved organic matter by viral lysis in marine microbial communities," *Aquatic Microbial Ecology*, vol. 27, no. 2, pp. 187–194, 2002.

[120] Miklič, A. and Rogelj, I., "Characterization of lactococcal bacteriophages isolated from slovenian dairies," *International Journal of Food Science and Technology*, vol. 38, pp. 305–311, 2003.

[121] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[122] Mizoguchi, K., Morita, M., Fischer, C. R., Yoichi, M., Tanji, Y., and Unno, H., "Coevolution of bacteriophage pp01 and *Escherichia coli* o157: H7 in continuous culture," *Applied and Environmental Microbiology*, vol. 69, pp. 170–176, 2003.

[123] Moebus, K., "A method for the detection of bacteriophages from ocean water," *Helgoländer Meeresuntersuchungen*, vol. 34, no. 1, pp. 1–14, 1980.

[124] Moebus, K. and Nattkemper, H., "Bacteriophage sensitivity patterns among bacteria isolated from marine waters," *Helgoländer Meeresuntersuchungen*, vol. 34, no. 3, pp. 375–385, 1981.

[125] Newman, M. E. J. and Girvan, M., "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[126] Newman, M., *Networks: an introduction*. OUP Oxford, 2010.

[127] Newman, M. E., "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[128] Newman, M. E., "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[129] Olesen, J. M., Dupont, Y. L., O'Gorman, E., Ings, T. C., Layer, K., Melian, C. J., Trojelsgaard, K., Pichler, D. E., Rasmussen, C., and Woodward, G., "From broadstone to zackenberg: space, time and hierarchies in ecological networks," *Advances in Ecological Research*, vol. 42, pp. 1–69, 2010.

[130] Olesen, J. M., Bascompte, J., Dupont, Y. L., and Jordano, P., "The modularity of pollination networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19891–19896, 2007.

[131] Pantůček, R., Rosypalová, A., Doškař, J., Kailerová, J., Růžičková, V., Borecká, P., Snopková, Š., Horváth, R., Götz, F., and Rosypal, S., "The polyvalent staphylococcal phage *ϕ812*: its host-range mutants and related phages," *Virology*, vol. 246, pp. 241–252, 1998.

[132] Pascual, M. and Dunne, J. A., "Ecological networks: Linking structure to dynamics in food webs," 2005.

[133] Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., and Brockhurst, M. A., "Antagonistic coevolution accelerates molecular evolution," *Nature*, vol. 464, pp. 275–278, 2010.

[134] Pimm, S. L., Lawton, J. H., and Cohen, J. E., "Food web patterns and their consequences," *Nature*, vol. 350, pp. 669–674, 1991.

[135] Poisot, T., Canard, E., Mouquet, N., and Hochberg, M. E., "A comparative study of ecological specialization estimators," *Methods in Ecology and Evolution*, vol. 3, no. 3, pp. 537–544, 2012.

[136] Poisot, T., "An a posteriori measure of network modularity," *F1000Research*, vol. 2, 2013.

[137] Poisot, T., Lepennetier, G., Martinez, E., Ramsayer, J., and Hochberg, M. E., "Resource availability affects the structure of a natural bacteria–bacteriophage community," *Biology letters*, vol. 7, no. 2, pp. 201–204, 2011.

[138] Poisot, T., Lounnas, M., and Hochberg, M. E., "The structure of natural microbial enemy-victim networks," *Ecological Processes*, vol. 2, no. 1, pp. 1–9, 2013.

[139] Poullain, V., Gandon, S., Brockhurst, M. A., Buckling, A., and Hochberg, M. E., "The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage," *Evolution*, vol. 62, pp. 1–11, 2008.

[140] Quiberoni, A., Auad, L., Binetti, A. G., Suárez, V. B., Reinheimer, J. A., and Raya, R. R., "Comparative analysis of *Streptococcus thermophilus* bacteriophages isolated from a yogurt industrial plant," *Food Microbiology*, vol. 20, pp. 461–469, 2003.

[141] Quince, C., Curtis, T. P., and Sloan, W. T., "The rational exploration of microbial diversity," *The ISME journal*, vol. 2, no. 10, pp. 997–1006, 2008.

[142] Rives, A. W. and Galitski, T., "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–1133, 2003.

[143] Rodríguez-Gironés, M. A. and Santamaría, L., "A new algorithm to calculate the nestedness temperature of presence-absence matrices," *Journal of Biogeography*, vol. 33, pp. 924–935, 2006.

[144] Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Beltran Rodriguez-Brito, L. P., Thingstad, T. F., and Forest Rohwer, A. M., "Explaining microbial population genomics through phage predation," *Nature Reviews Microbiology*, vol. 7, no. 11, pp. 828–836, 2009.

[145] Rohwer, F. and Thurber, R. V., "Viruses manipulate the marine environment," *Nature*, vol. 459, pp. 207–212, 2009.

[146] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., and others, "The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific," *PLoS biology*, vol. 5, no. 3, p. e77, 2007.

[147] Rybniker, J., Kramme, S., and Small, P. L., "Host range of 14 mycobacteriophages in *Mycobacterium ulcerans* and seven other mycobacteria including *Mycobacterium tuberculosis*–application for identification and susceptibility testing," *Journal of Medical Microbiology*, vol. 55, pp. 37–42, 2006.

[148] Sambrook, J., Fritsch, E. F., Maniatis, T., and others, *Molecular cloning*, vol. 2. Cold spring harbor laboratory press New York, 1989.

[149] Sano, E., Carlson, S., Wegley, L., and Rohwer, F., "Movement of viruses between biomes," *Appl. Env. Microb.*, vol. 70, pp. 5842–5846, 2004.

[150] Sasaki, A., "Host-parasite coevolution in a multilocus gene-for-gene system," *Proceedings of the Royal Society B: Biological Sciences*, vol. 267, pp. 2183–2188, 2000.

[151] Sawardecker, E., Amundsen, C., Sales-Pardo, M., and Amaral, L., "Comparison of methods for the detection of node group membership in bipartite networks," *The European Physical Journal B*, vol. 72, no. 4, pp. 671–677, 2009.

[152] Seed, K. D. and Dennis, J. J., "Isolation and characterization of bacteriophages of the *Burkholderia cepacia* complex," *FEMS Microbiology Letters*, vol. 251, pp. 273–280, 2005.

[153] SHANNON, C., "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[154] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[155] Silander, O. K., Weinrich, D. M., Wright, K. M., O'Keefe, K. J., Rang, C. U., Turner, P. E., and Chao, L., "Widespread genetic exchange among terrestrial bacteriophages," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 19009–19014, 2005.

[156] Sillankorva, S., Neubauer, P., and Azeredo, J., "Isolation and characterization of a t7-like lytic phage for pseudomonas fluorescens," *BMC Biotechnology*, vol. 8, no. 1, p. 80, 2008.

[157] Simpson, E. H., "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, 1949.

[158] S.J., B., C.A., B., and H.T.P., W., "Falcon: nestedness statistics for bipartite networks. figshare.."

[159] Spanakis, E. and Horne, M. T., "Co-adaptation of *Escherichia coli* and coliphage λvir in continuous culture," *Journal of General Microbiology*, vol. 133, pp. 353–360, 1987.

[160] Staniczenko, P. P., Kopp, J. C., and Allesina, S., "The ghost of nestedness in ecological networks," *Nature communications*, vol. 4, p. 1391, 2013.

[161] Stenholm, A. R. n., Dalsgaard, I., and Middelboe, M., "Isolation and characterization of bacteriophages infecting the fish pathogen *Flavobacterium psychrophilum*," *Applied and Environmental Microbiology*, vol. 74, pp. 4070–4078, 2008.

[162] Stouffer, D. B. and Bascompte, J., "Compartmentalization increases food-web persistence," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3648–3652, 2011.

[163] Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W., "Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*," *Nature*, vol. 424, pp. 1047–1051, 2003.

[164] Suttle, C. A. and Chan, A. M., "Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics," *Marine Ecology Progress Series*, vol. 92, pp. 99–109, 1993.

[165] Suttle, C. A., "Viruses in the sea," *Nature*, vol. 437, no. 7057, pp. 356–361, 2005.

[166] Suttle, C. A., "Marine viruses – major players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, no. 10, pp. 801–812, 2007.

[167] Suttle, C. A. and Chan, A. M., "Dynamics and distribution of cyanophages and their effect on marine *Synechococcus* spp.," *Applied and Environmental Microbiology*, vol. 60, no. 9, pp. 3167–3174, 1994.

[168] Synnott, A. J., Kuang, Y., Kurimoto, M., Yamamichi, K., Iwano, H., and Tanji, Y., "Isolation from sewage influent and characterization of novel *Staphylococcus aureus* bacteriophages with wide host ranges and potent lytic capabilities," *Applied and Environmental Microbiology*, vol. 75, pp. 4483–4490, 2009.

[169] Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R., and Phillips, R., "Probing individual environmental bacteria for viruses by using microfluidic digital pcr," *Science*, vol. 333, no. 6038, pp. 58–62, 2011.

[170] Thingstad, T. F., "Elements of a theory for the mechanisms controlling abundance, diversity and biogeochemical role of lytic bacterial viruses in aquatic systems," *Limnology and Oceanography*, vol. 45, pp. 1320–1328, 2000.

[171] Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A., and Giske, J., "A theoretical analysis of how strain-specific viruses can control microbial species diversity," *Proceedings of the National Academy of Sciences*, vol. 111, no. 21, pp. 7813–7818, 2014.

[172] Thingstad, T. and Lignell, R., "Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand," *Aquatic Microbial Ecology*, vol. 13, pp. 19–27, 1997.

[173] Thompson, J. N., *The geographic mosaic of coevolution*. University of Chicago Press, 2005.

[174] Thompson, J. N., "Specific hypotheses on the geographic mosaic of coevolution," *the american naturalist*, vol. 153, no. S5, pp. S1–S14, 1999.

[175] Torsvik, V., Ovreas, L., and Thingstad, T. F., "Prokaryotic diversity – magnitude, dynamics, and controlling factors," *Science*, vol. 296, pp. 1064–1066, 2002.

[176] Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., and Rubin, E. M., "Comparative metagenomics of microbial communities," *Science*, vol. 308, pp. 554–557, 2005.

[177] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F., "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, pp. 37–43, 2004.

[178] Ulrich, W., Almeida-Neto, M., and Gotelli, N. J., "A consumer's guide to nestedness analysis," *Oikos*, vol. 118, no. 1, pp. 3–17, 2009.

[179] Ulrich, W. and Gotelli, N. J., "Null model analysis of species nestedness patterns," *Ecology*, vol. 88, pp. 1824–1831, 2007.

[180] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. a., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O., "Environmental genome shotgun sequencing of the sargasso sea," *Science*, vol. 304, pp. 66–74, 2004.

[181] Vos, M., Birkett, P. J., Birch, E., Griffiths, R. I., and Buckling, A., "Local adaptation of bacteriophages to their bacterial hosts in soil," *Science*, vol. 325, no. 5942, pp. 833–833, 2009.

[182] Wang, K. and Chen, F., "Prevalence of highly host-specific cyanophages in the estuarine environment," *Environmental Microbiology*, vol. 10, pp. 300–312, 2008.

[183] Waser, N. M. and Ollerton, J., *Plant-pollinator interactions: from specialization to generalization*. Chicago: University of Chicago.

[184] Watts, D. J. and Strogatz, S. H., "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[185] Weinbauer, M. G., "Ecology of prokaryotic viruses," *FEMS microbiology reviews*, vol. 28, no. 2, pp. 127–181, 2004.

[186] Weitz, J. S., Hartman, H., and Levin, S. A., "Coevolutionary arms races between bacteria and bacteriophage," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 9535–9540, 2005.

[187] Weitz, J. S., Poisot, T., Meyer, J. R., Flores, C. O., Valverde, S., Sullivan, M. B., and Hochberg, M. E., "Phage–bacteria infection networks," *Trends in microbiology*, vol. 21, no. 2, pp. 82–91, 2013.

[188] Weitz, J. S. and Wilhelm, S. W., "Ocean viruses and their effects on microbial communities and biogeochemical cycles," *F1000 biology reports*, vol. 4, p. 17, 2012.

[189] Wichels, A., Biel, S. S., Gelderblom, H. R., Brinkhoff, T., Muyzer, G., and Schütt, C., "Bacteriophage diversity in the north sea," *Applied and Environmental Microbiology*, vol. 64, pp. 4128–4133, 1998.

[190] Wilhelm, S. W. and Suttle, C. A., "Viruses and nutrient cycles in the sea," *BioScience*, vol. 49, no. 10, pp. pp. 781–788, 1999.

[191] Williams, R. J. and Martinez, N. D., "Simple rules yield complex food webs," *Nature*, vol. 404, pp. 180–183, 2000.

[192] Williams, R. J., Anandanadesan, A., and Purves, D., "The probabilistic niche model reveals the niche structure and role of body size in a complex food web," *PloS one*, vol. 5, no. 8, p. e12092, 2010.

[193] Winter, C., Bouvier, T., Weinbauer, M. G., and Thingstad, T. F., "Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited," *Microbiology and Molecular Biology Reviews*, vol. 74, no. 1, pp. 42–57, 2010.

[194] Wommack, K. E. and Colwell, R. R., "Virioplankton: viruses in aquatic ecosystems," *Microbiology and Molecular Biology Reviews*, vol. 64, pp. 69–114, 2000.

[195] Yooseph, S., Sutton, G., Rusch, D. B. B., Halpern, A. L. L., Williamson, S. J. J., Remington, K., Eisen, J. A. A., Heidelberg, K. B. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S. S., Li, H., Mashiyama, S. T. T., Joachimiak, M. P. P., van Belle, C., Chandonia, J.-M. M., Soergel, D. A. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J. J., Bafna, V., Friedman, R., Brenner, S. E. E., Godzik, A., Eisenberg, D., Dixon, J. E. E., Taylor, S. S. S., Strausberg, R. L. L., Frazier, M., and Venter, J. C. C., "The sorcerer ii global ocean sampling expedition: Expanding the universe of protein families.," *PLoS Biol*, vol. 5, no. 3, pp. 432–466, 2007.

[196] Yu, M. X., Slater, M. R., and Ackermann, H.-W., "Isolation and characterization of thermus bacteriophages," *Archives of Virology*, vol. 151, no. 4, pp. 663–679, 2006.

[197] Zinno, P., Janzen, T., Bennedsen, M., Ercolini, D., and Mauriello, G., "Characterization of *Streptococcus thermophilus* lytic bacteriophages from

mozzarella cheese plants," *International Journal of Food Microbiology*, vol. 138, pp. 137–144, 2010.

VITA

Cesar Flores was born in Tlacolula de Matamoros, Oaxaca, Mexico. After completing his high schoolwork at Instituto Carlos Gracida in Oaxaca, Cesar entered the Monterrey Institute of Technology in Nuevo Leon, Mexico in the year of 2001. He graduated with a B.Sc. in Physics Engineering from this Institute and a double degree in Management of Information Systems from the EPF (Ecole d'Ingenieurs) in Sceaux, France in 2006. He immediately continue his studies with a M.Sc in Intelligent systems at the Monterrey Institute of Technology, graduating in 2008. After working with Leonardo Garrido during his master, he attended Georgia Institute of Technology from 2009 to 2014 as a PhD student, working under the advice of Prof. Joshua S. Weitz in Complex Ecological Networks. Among the achievements during his PhD are the realization of an internship at Microsoft Research Cambridge under Dr. Drew Purves, and the participation at the Complex System Summer School 2013 organized by the Santa Fe Institute.