

Seminar Paper

Community detection with consensus clustering in marine microbial association networks: Atlantic vs Pacific (GRUMP)

Department of Statistics
Ludwig-Maximilians-Universität München

Maximilian Frei

Munich, February 28th, 2026



Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Daniele Pugno

Abstract

We apply community detection and consensus clustering to marine microbial association networks from the GRUMP database, comparing Atlantic and Pacific basins. The theoretical framework follows Fortunato and Hric (2016): definitions of community, null models, the stochastic block model, validation and detectability, and algorithms (Louvain, Infomap, spectral, consensus). We describe the data and network construction, report modularity and partitions under several methods, and interpret modules using taxonomy and environmental variables. The goal is to compare microbial (bacteria, archaea, eukaryotes) ASV networks across basins and to assess the stability and ecological interpretability of detected communities.

Contents

1	Introduction	1
1.1	Background on Community Detection	1
1.2	Overview of Fortunato & Hric (2016)	1
1.3	Marine Microbial Networks and the GRUMP Database	1
1.4	Research Questions	2
1.5	Structure of this Paper	2
2	Methods	3
2.1	Graphs, null models, and the definition of community	3
2.2	The Stochastic Block Model	4
2.3	Validation, detectability, and significance testing	5
2.4	Community detection algorithms and method selection	8
2.4.1	Modularity maximisation and the Louvain algorithm	8
2.4.2	Flow-based methods: Infomap and the map equation	9
2.4.3	Spectral methods and the non-backtracking matrix	10
2.4.4	Other popular algorithms	11
2.4.5	Consensus clustering	11
2.5	Method selection	13
3	Results	14
3.1	Network structure and descriptive statistics	14
3.1.1	Data source and network description	14
3.1.2	Basic statistics: Atlantic and Pacific	14
3.2	Community structure and modularity	15
3.2.1	Detection results and modularity by algorithm	15
3.2.2	Algorithm comparison and stability	17
3.3	Ecological interpretation and basin comparison	17
3.3.1	Environmental and taxonomic associations	18
3.3.2	Basin-specific patterns	18
4	Conclusion	21
4.1	Summary of Findings	21
4.2	Methodological Insights	21
4.3	Ecological Implications	22
4.4	Limitations	22
4.5	Future Directions	22
A	Appendix	V
A.1	Atlantic network: individual panels	V
A.2	Pacific network: individual panels	V
B	Electronic appendix	XVII

1 Introduction

1.1 Background on Community Detection

Detecting communities in networks is one of the most studied and most difficult problems in network science (Fortunato and Hric, 2016). A central challenge is that there is no unique, rigorous definition of what a “community” is or what it means for a node to belong to one. The notion is inherently context-dependent and method-dependent, and different methods encode different notions of community (Fortunato and Hric, 2016). In this sense the problem is ill-posed. No algorithm is universally optimal, and results must be interpreted in context (Fortunato and Hric, 2016).

In formal terms, the problem is to partition a network into groups of densely connected nodes, with only sparse connections between groups (Blondel et al., 2008). Precise formulations of this optimisation problem are known to be computationally intractable, so in practice one relies on heuristics that find reasonably good partitions in a reasonably fast way (Blondel et al., 2008). Blondel et al. (2008) distinguish three broad types of algorithms: *divisive* methods identify and remove inter-community links (e.g. edge-betweenness), *agglomerative* methods merge similar nodes or communities recursively, and *optimisation* methods maximise an objective function such as modularity. The search for fast, scalable algorithms has attracted much interest with the increasing availability of large network data sets (Blondel et al., 2008).

1.2 Overview of Fortunato & Hric (2016)

Fortunato and Hric (2016) published a landmark user guide to community detection in *Physics Reports* in 2016. Their work presents a critical analysis of the problem intended for practitioners but accessible to readers with basic notions of network science. It is not an exhaustive survey but focuses on general aspects in the light of recent findings and gives advice on the usage of popular algorithm classes. The review surveys the main classes of methods and the underlying theory (definitions of community, validation and benchmarks, detectability, and a critical discussion of modularity, flow-based and spectral approaches) and gives practical recommendations for choosing and evaluating algorithms. We adopt this user-guide perspective: we do not propose a new method but apply and compare existing ones, with an emphasis on validation and consensus clustering. Broader surveys of network clustering are cited in Fortunato and Hric (2016).

1.3 Marine Microbial Networks and the GRUMP Database

Networks are widely used to represent associations among microbial taxa in marine systems, with nodes as taxa (e.g. amplicon sequencing variants, ASVs) and edges as co-occurrence or inferred interactions. Such networks capture which taxa tend to co-occur across samples and can reveal structure that is invisible from taxonomy or single-sample diversity alone. Detecting communities in these networks can suggest functionally or environmentally coherent groups and support comparative questions across regions or time. The Global Repeating Upper-ocean Microbial Patterns (GRUMP) database (McNichol et al., 2025) provides a large, standardized resource of plankton metabarcoding data

from ocean cruises, covering bacteria, archaea, and eukaryotes. Data are processed in a consistent way, so that association networks built from GRUMP are comparable across ocean basins. This makes GRUMP well suited for asking whether microbial co-occurrence structure and its modular organisation differ between regions such as the Atlantic and the Pacific.

1.4 Research Questions

The goal of this paper is to compare microbial ASV networks (bacteria, archaea, and eukaryotes) across ocean basins (e.g. Atlantic vs Pacific), estimate modularity under different community-detection methods, and enrich the analysis with biological and environmental features. We ask, first, whether the two basins differ in the strength and organisation of modular structure (as measured by modularity and the number and size of communities). Second, we compare several detection algorithms and use consensus clustering to obtain a stable partition for interpretation, so that results do not depend on a single run of one method. Third, we interpret the detected modules using taxonomy (e.g. phylum, class) and node-level environmental summaries (temperature, salinity, depth, oxygen), without treating agreement with metadata as a validation criterion; the aim is to attach ecological meaning to the structural partition and to compare how modules align with environmental or taxonomic gradients in each basin.

1.5 Structure of this Paper

The rest of the paper is organised as follows. The **Introduction** has outlined the motivation and research questions. The **Methods** section follows the logic of Fortunato and Hric (2016): it presents the theoretical framework for community detection—definitions of community, null models, and algorithms—with the formulas needed for the subsequent analysis, so that the choices made for the GRUMP networks are well grounded. The **Results** section first describes the data and networks (GRUMP-derived Atlantic and Pacific ASV association networks) and their basic statistics. It then reports community structure and modularity under different detection algorithms, including algorithm comparison and consensus clustering, and finally discusses ecological interpretation and basin comparison. The **Conclusion** summarises the findings, methodological lessons, and limitations.

2 Methods

This section lays the groundwork for the microbial network analysis by reviewing the theory of community detection in networks. We follow the framework of Fortunato and Hric (2016) and related literature, including key definitions (graphs, communities, null models), generative models such as the stochastic block model, validation and significance testing (as theory; we do not perform it in the GRUMP analysis), and the main algorithm families (modularity maximisation, flow-based and spectral methods, consensus clustering). The presentation includes the necessary formulas and prepares the choice of methods and validation used later on the GRUMP networks.

2.1 Graphs, null models, and the definition of community

We consider undirected graphs (networks) as the main object of study. A graph $G = (V, E)$ consists of a set of nodes V (e.g. taxa or ASVs) and a set of edges E between them. Here $n = |V|$ is the number of nodes and $m = |E|$ the number of edges. The graph can be represented by an $n \times n$ adjacency matrix \mathbf{A} , with $A_{ij} = 1$ if there is an edge between nodes i and j (and $A_{ij} = 0$ otherwise). For weighted graphs, A_{ij} takes the edge weight. The degree of node i is $k_i = \sum_j A_{ij}$. Many real-world networks, including microbial association networks, are sparse ($m \ll n^2$) and exhibit heterogeneous degree distributions and local clustering, so that the arrangement of edges is far from random (Fortunato and Hric, 2016).

A *community* (or *module*) is informally a subset of nodes that are more densely connected among themselves than to the rest of the network. There is no single, universally agreed mathematical definition. Different algorithms implicitly or explicitly optimise different notions (Fortunato and Hric, 2016). Early, *classic* definitions relied on edge counting (Fortunato and Hric, 2016). For a candidate subgraph C , the *internal degree* of a node $i \in C$ is the number of edges from i to other nodes in C , and the *external degree* is the number of edges from i to the rest of the network. The ratio of internal to total degree is sometimes called *embeddedness*. A natural idea is that a community should have more internal than external edges. Fortunato and Hric (2016) summarise two influential formalisations: a *strong* community (or LS-set) is a subgraph in which *every* node has internal degree greater than its external degree. A *weak* community requires only that the *total* internal degree of the subgraph exceeds its total external degree (Fortunato and Hric, 2016). These definitions assume non-overlapping, well-separated modules and are *extensive* (they scale with community size), so they can perform poorly in large, heterogeneous networks. Other formalisations use cut-based measures: for example, *conductance* of a set C is the ratio of edges leaving C to the total volume of C . If conductance is sufficiently low for clusters of a given size, one has a well-separated community at that scale (Fortunato and Hric, 2016).

Real networks often exhibit overlapping communities, hierarchical structure, and fuzzy boundaries, so edge-counting definitions are insufficient and motivate a *modern*, probabilistic view (Fortunato and Hric, 2016). In that view, a community is a set of nodes among which the *probability* of connection is higher than between groups (e.g. $p_{\text{in}} > p_{\text{out}}$). Communities are thus statistical regularities rather than purely geometric ones. This

probabilistic view naturally leads to generative models such as the stochastic block model (see next subsection).

Community detection is the task of partitioning the node set into (possibly overlapping) communities. Figure 1 illustrates the traditional view of communities as dense subgraphs.

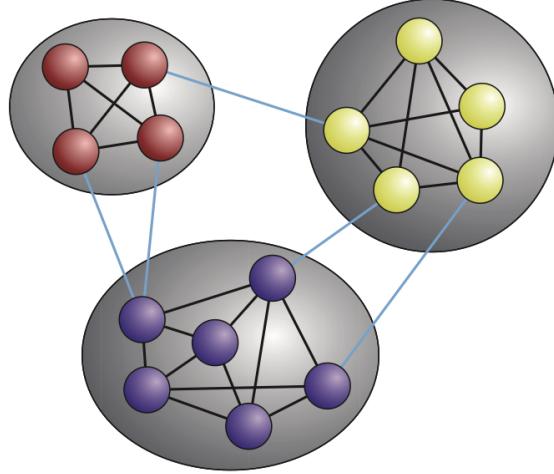


Figure 1: Traditional view of community structure: dense subgraphs with more internal than external edges (from Fortunato & Hric, 2016).

A *null model* is a random graph model that preserves certain features of the observed graph while randomising the rest. Comparing an observed quantity (such as the density of links inside a putative community, or the value of a quality function like modularity) to its distribution under the null model allows one to ask whether the observed structure is statistically meaningful rather than a chance fluctuation. The *configuration model* is a standard null model: it generates random graphs with a prescribed degree sequence (e.g. that of the observed graph), typically by randomising edge endpoints. Preserving the degree sequence matters because many structural properties (including modularity) are defined relative to this baseline. It avoids attributing structure to effects that arise purely from heterogeneous degrees (Fortunato and Hric, 2016). Modularity-based methods, introduced by Newman and Girvan (2004) and developed further by Newman (2006), compare the observed pattern of edges within communities to the expectation under the configuration model. The modularity formula and algorithm details are given later in this Methods section (subsection on modularity maximisation and the Louvain algorithm).

2.2 The Stochastic Block Model

The stochastic block model (SBM) gives a precise *generative* definition of community structure: one specifies a probabilistic process that produces the graph, rather than optimising a quality function such as modularity (Fortunato and Hric, 2016, Holland et al., 1983). The model has roots in the social networks literature (Holland et al., 1983). We follow the formulation in Fortunato and Hric (2016). Nodes are assigned to one of q blocks (groups). Let $g_i \in \{1, \dots, q\}$ denote the block of node i . Edges are then drawn independently: the probability of an edge between nodes i and j depends only on their

block labels, $P(A_{ij} = 1) = \omega_{g_i, g_j}$, where $\boldsymbol{\omega}$ is a $q \times q$ matrix of connection probabilities. Typically, diagonal entries ω_{rr} are larger than off-diagonal ones, so that nodes in the same block are more likely to be connected (planted community structure). The likelihood of an observed graph under the SBM is a function of the block assignment and $\boldsymbol{\omega}$. Inference amounts to estimating these parameters, e.g. by maximum likelihood or expectation–maximisation (EM), and the number of blocks q can be chosen via model selection criteria such as the Bayesian information criterion (BIC) or minimum description length (Fortunato and Hric, 2016). The SBM thus differs from quality-function methods (e.g. modularity maximisation): it defines communities as blocks in a generative model, provides a likelihood and a clear notion of fit, and allows comparison of different partition sizes. Figure 2 shows an adjacency matrix from an SBM with blocks visible after reordering nodes by group.

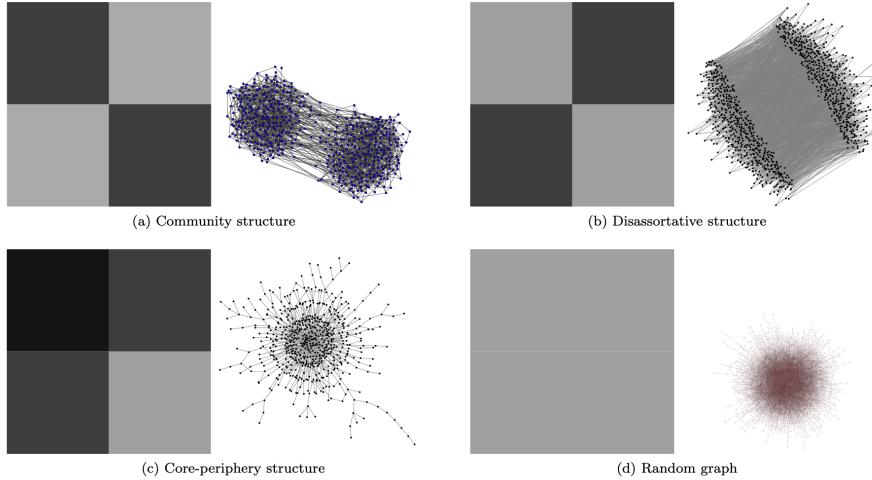


Figure 2: Adjacency matrix of a graph generated from a stochastic block model. Nodes are ordered by block membership so that blocks appear as dense blocks (from Fortunato & Hric, 2016).

2.3 Validation, detectability, and significance testing

Evaluating community detection is difficult when ground truth is unknown. We summarise how benchmarks, partition similarity measures, structure versus metadata, significance testing, and two fundamental limits—resolution and detectability—shape the interpretation of results. This motivates our use of multiple algorithms and consensus clustering in the GRUMP analysis.

Benchmark networks with a *planted* partition allow comparison of algorithm output to ground truth. The *Girvan–Newman* (GN) benchmark uses graphs with equal degree and equal community size (Figure 3). It is simple but unrealistic, because real networks typically have heterogeneous degrees and community sizes (Fortunato and Hric, 2016). The *Lancichinetti–Fortunato–Radicchi* (LFR) benchmark (Lancichinetti et al., 2008) generates graphs with power-law degree and power-law community size distributions (Figure 4), pro-

viding a more realistic test. Benchmarks evaluate *algorithms*, not real data. They do not validate that a given network has community structure.

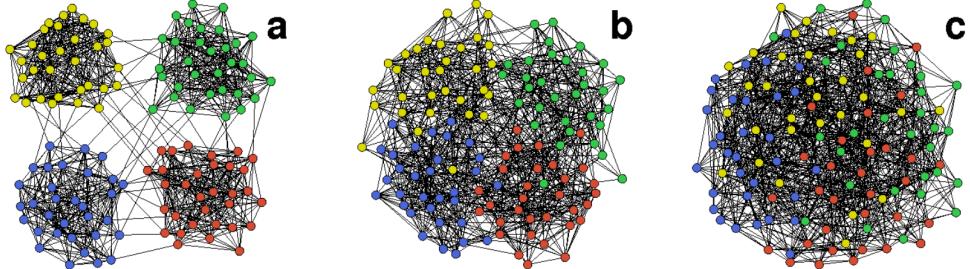


Figure 3: Girvan–Newman benchmark: graphs with known community structure, equal degree and size (from Fortunato & Hric, 2016).

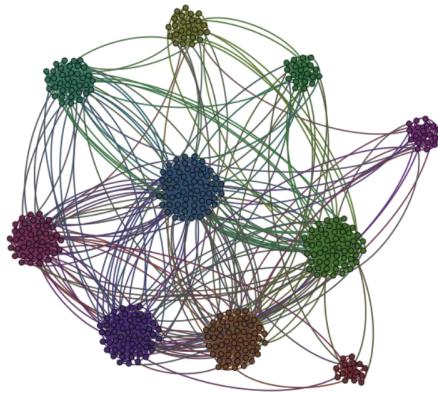


Figure 4: LFR benchmark: heterogeneous degrees and community sizes (Lancichinetti et al., 2008; from Fortunato & Hric, 2016).

To compare two partitions (e.g. algorithm output vs. planted partition, or two algorithm runs), similarity measures are needed. Similarity measures fall into pair-counting, cluster-matching, and information-theoretic classes (Fortunato and Hric, 2016). *Normalized mutual information* (NMI) is widely used:

$$\text{NMI} \in [0, 1]; \quad \text{NMI} = 1 \text{ when partitions are identical, } \text{NMI} = 0 \text{ when independent.} \quad (1)$$

NMI is sensitive to the number of clusters, so that partitions with more clusters can receive higher NMI even when not closer to the true partition (Fortunato and Hric, 2016). The *variation of information* (VI) (Meilă, 2007) is a distance between partitions:

$$\text{VI}(X, Y) = H(X|Y) + H(Y|X), \quad (2)$$

where $H(X|Y)$ is the conditional entropy. VI is a proper metric (non-negative, symmetric, satisfies the triangle inequality) and is theoretically better behaved. Lower VI means

greater agreement (Fortunato and Hric, 2016, Meilă, 2007). We report both NMI and VI where we compare partitions (e.g. across algorithms or runs).

For real networks, ground truth is usually unknown. Fortunato and Hric (2016) advise against “validating” communities by agreement with metadata (e.g. taxonomic labels): structural communities need not align with external covariates. Metadata should *interpret* the partition, not serve as a criterion of success (Figure 5). In the Results we use taxonomy and environmental variables to interpret modules, not to validate them.

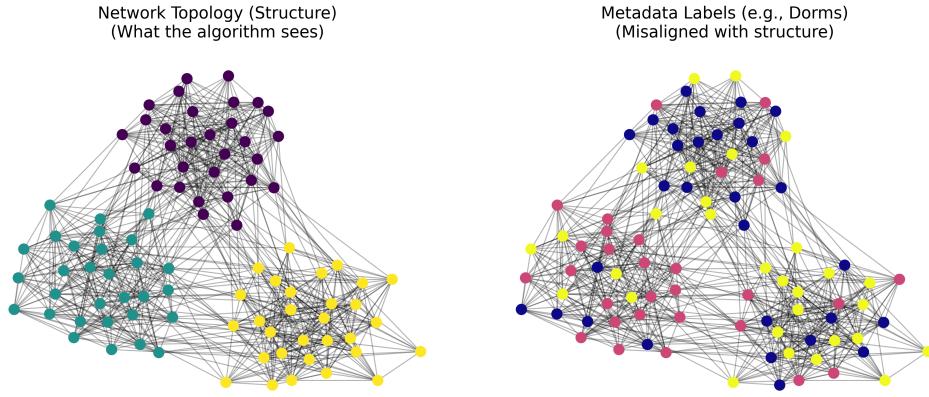


Figure 5: Structural communities (left) need not align with metadata such as taxonomic classification (right). Interpretation rather than validation is the goal.

Observed community structure (e.g. modularity Q or the number of within-community edges) should be compared to a null model. The *configuration model* (random graphs with the same degree sequence) is the standard null (Fortunato and Hric, 2016). One computes the same quantity on many random graphs from this model. If the observed value is far in the tail of the null distribution, the structure is unlikely to be due to chance. Modularity alone is insufficient—comparison to the configuration model is required to assess significance (Fortunato and Hric, 2016). We do not perform this step in the GRUMP analysis; it is left for future work (see Conclusion, Future Directions).

Two limits are important for interpretation.

(1) *Resolution limit* (Fortunato and Barthélémy, 2007): modularity-based methods (e.g. Louvain) have a preferential scale and cannot detect communities smaller than roughly $\sqrt{2m}$. Even a small, dense clique may be merged with a neighbouring group if that increases global modularity. This is an *algorithmic* bias of modularity maximisation. Since we use Louvain and modularity to compare Atlantic and Pacific networks, our results may miss smaller, biologically relevant clusters.

(2) *Detectability phase transition* (Decelle et al., 2011): in sparse networks, if the “signal” (edges within communities) is too weak relative to “noise” (edges between communities), no algorithm can recover the partition better than random—the structure is information-theoretically undetectable. This is a *fundamental* limit of the network, not of a particular algorithm. Marine microbial networks are often sparse. Referencing Decelle et al. (2011) allows discussion of whether weak or absent modular structure reflects biology or a non-detectable regime. Awareness of both limits justifies comparing multiple algorithms and

interpreting results cautiously (Fortunato and Hric, 2016, Fortunato and Barthélemy, 2007, Decelle et al., 2011).

2.4 Community detection algorithms and method selection

2.4.1 Modularity maximisation and the Louvain algorithm

Modularity was introduced by Newman and Girvan (2004) as a quality function to evaluate partitions: it measures how much the density of edges inside communities exceeds the expectation under a null model, typically the configuration model (random graphs with the same degree sequence). For a partition into communities,

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \quad (3)$$

where c_i is the community of node i and $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise (Fortunato and Hric, 2016, Newman and Girvan, 2004). Up to a multiplicative constant, Q is the number of edges within groups minus the expected number in an equivalent network with edges placed at random (Newman, 2006). Thus Q is high when links are concentrated inside communities relative to the configuration-model baseline. Modularity takes values in $[-1, 1]$. For weighted networks it is often interpreted as 0 for random structure and 1 for strong community structure. Values above about 0.3–0.4 are often taken as evidence of modular structure, though the *maximal* modularity attainable depends on the network and its size, so Q should not be compared directly across different networks (Fortunato and Barthélemy, 2007, Fortunato and Hric, 2016).

Exact modularity optimisation is NP-hard, so approximation algorithms are necessary for large networks (Blondel et al., 2008). The *Louvain* algorithm (Blondel et al., 2008) is a fast, scalable heuristic with two phases:

1. *Local movement*: repeatedly move each node to the neighbouring community that yields the largest increase in Q , until no improvement is possible.
2. *Aggregation*: replace each community by a single super-node, with edge weights between super-nodes given by the sum of weights between the corresponding communities, then repeat phase 1 on the aggregated network.

The process is hierarchical and runs in time $O(n \log n)$ in practice, making it one of the most widely used methods. Before Louvain, the fastest competitive approximation for large networks was the agglomerative method of Clauset et al. (2004). Because the order of node updates and ties can lead to different partitions, Louvain is non-deterministic—this motivates repeated runs and consensus clustering (see below).

Two important caveats affect interpretation.

(1) *Resolution limit* (see above): it is therefore a priori impossible to tell whether a module found by Louvain is a single community or a cluster of smaller modules that were merged. This introduces caveats when interpreting Louvain output (e.g. biologically distinct microbial groups may be merged).

(2) *Modularity trap*: the modularity landscape has many local optima, so different runs or initialisations may yield different partitions with similar Q , and there is no unique “best” partition from modularity alone (Fortunato and Hric, 2016). Louvain can also produce relatively few, large “super-communities” even on networks with no strong underlying structure (Blondel et al., 2008). Testing whether observed modularity is significant requires comparison to the distribution of Q under the configuration model (Figure 6).

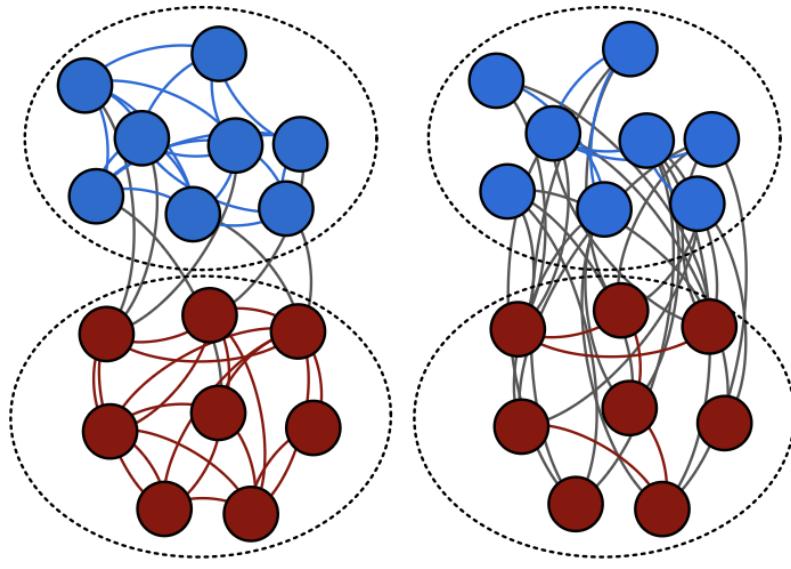


Figure 6: Modularity of observed network compared to distribution under a null model - grey edges indicate connections to nodes outside of the community (from Fortunato & Hric, 2016).

2.4.2 Flow-based methods: Infomap and the map equation

Flow-based methods define communities through the dynamics of a process on the network rather than through static structural quantities (e.g. edge counts or degrees). Random walks are the most widely used: if communities have high internal edge density and are well-separated from each other, a random walker tends to remain inside each cluster for a long time before escaping (Fortunato and Hric, 2016). Good communities are thus regions where flow is “trapped” relative to the rest of the graph.

Infomap (Rosvall and Bergstrom, 2008) exploits this idea in an information-theoretic way. The goal is to describe the trajectory of a random walk on the network as efficiently as possible. A naive encoding assigns a unique code to every node. A *two-level* encoding instead assigns codes to *modules* and reuses shorter codes for nodes within each module. When the partition reflects real flow bottlenecks, the expected description length of the walk is shorter. The *map equation* quantifies this expected length. Infomap finds the partition that minimises it (Rosvall and Bergstrom, 2008, Fortunato and Hric, 2016). Figure 7 illustrates this idea: a two-level description (module names plus codes within modules) yields a shorter description of the random walk than a one-level encoding. The approach does not assume a fixed null model (unlike modularity) and can recover structure

at multiple scales. It has been extended to hierarchical partitions and to directed and weighted networks (Rosvall and Bergstrom, 2008).

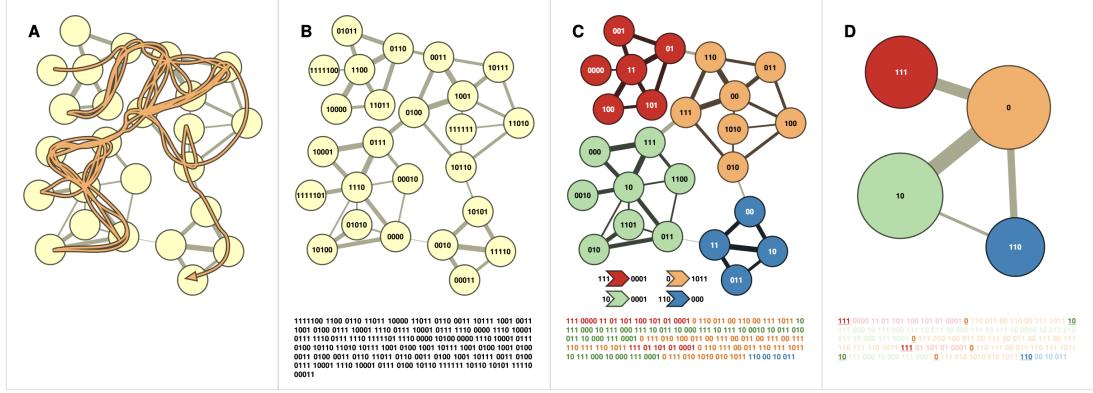


Figure 7: Detecting communities by compressing the description of information flows (Rosvall & Bergstrom, 2008). (A) A random walk on the network. (B) A one-level encoding gives a unique code to every node (e.g. Huffman). (C) A two-level encoding assigns codes to modules and reuses shorter codes for nodes within each module, yielding on average a shorter description. (D) Reporting only module names gives an optimal coarse-graining.

Infomap and its variants often return *different* partitions than structure-based methods such as modularity optimisation, because they optimise flow compression rather than density relative to the configuration model (Fortunato and Hric, 2016). The difference is particularly striking on directed graphs, where edge directions strongly constrain possible flows (Rosvall and Bergstrom, 2008). When comparing algorithms in the Results, we therefore expect flow-based (Infomap) and density-based (Louvain) partitions to differ. Both can be informative. Infomap shares resolution-type limitations (e.g. minimum detectable community scale) and depends on the flow model (unweighted vs weighted, undirected vs directed), so results should be interpreted in light of the chosen network representation (Fortunato and Hric, 2016).

2.4.3 Spectral methods and the non-backtracking matrix

Spectral methods use the eigenvalues and eigenvectors of a matrix associated with the graph to embed nodes and then cluster them. For the adjacency matrix or the Laplacian, a well-known problem in *sparse* networks is that many eigenvalues lie in a *bulk*, and the few eigenvalues that carry information about community structure can merge with this bulk, so that standard spectral clustering fails (Fortunato and Hric, 2016). Real-world networks (including microbial association networks) are often sparse, so this limitation is relevant in practice.

Krzakala et al. (2013) introduced spectral algorithms based on the *non-backtracking* matrix B : it encodes steps along directed edges that never go back immediately (from (u, v) to (v, w) only when $w \neq u$). Formally,

$$B_{(u \rightarrow v), (v \rightarrow w)} = 1 \quad \text{if } w \neq u, \quad \text{and } 0 \text{ otherwise.} \quad (4)$$

The spectrum of B is much better behaved than that of the adjacency matrix or other commonly used matrices: it maintains a strong separation between the *bulk* eigenvalues and the eigenvalues relevant to community structure even in the sparse case (Krzakala et al., 2013). For graphs generated by the stochastic block model, the resulting algorithm is optimal in the sense that it detects communities all the way down to the theoretical detectability limit (Krzakala et al., 2013). The number of real eigenvalues of B lying outside the bulk can be used to *infer* the number of communities q . The corresponding eigenvectors then support clustering of nodes into communities without requiring q to be fixed in advance (Fortunato and Hric, 2016, Krzakala et al., 2013). For further theoretical treatment of the non-backtracking spectrum, see Glover and Kempton (2020).

2.4.4 Other popular algorithms

Several other algorithms are commonly used and implemented in standard tools. *Fast Greedy* (Clauset et al., 2004) is an agglomerative method that maximises modularity by repeatedly merging the pair of communities that yields the largest increase in Q . It uses efficient (e.g. heap-based) data structures and runs in $O(n \log^2 n)$ time on sparse graphs, producing a dendrogram. It establishes greedy modularity optimisation as a scalable alternative to exact optimisation and is often faster than Louvain but generally less accurate in practice (Clauset et al., 2004, Fortunato and Hric, 2016). *Leading Eigenvector* (Newman, 2006) reformulates modularity in terms of the *modularity matrix* and uses the leading eigenvector to bipartition the graph. The process can be applied recursively to obtain a hierarchical partition. This bridges modularity maximisation with spectral partitioning and is fast for large networks (Newman, 2006). *Walktrap* (Pons and Latapy, 2005) exploits the idea that short random walks tend to stay within communities: it defines a distance between nodes from the difference in the probability of being at each node after a fixed number of steps, then performs hierarchical clustering on this distance to obtain a dendrogram. Walktrap thus bridges structural (modularity-like) and flow-based (random-walk) intuition and is often competitive with other methods (Pons and Latapy, 2005, Fortunato and Hric, 2016). All of these can be combined with consensus clustering (below) for robustness.

2.4.5 Consensus clustering

Many algorithms, especially Louvain and other stochastic heuristics, can return different partitions across runs because the modularity (or other objective) landscape has many local optima and the order of node updates or tie-breaking is not fixed (Fortunato and Hric, 2016). Single runs are therefore unstable: multiple near-optimal solutions exist, and there is no guarantee that one run captures the most meaningful or reproducible structure. *Consensus clustering* is recommended to increase robustness (Fortunato and Hric, 2016). A simple approach is to run the algorithm many times and either build a co-association matrix and re-cluster it, or choose a representative partition (e.g. the run with highest modularity). A more rigorous, iterative procedure is given by Lancichinetti and Fortunato (2012):

1. Apply the algorithm A on the original graph G n_P times, yielding n_P partitions.

2. Compute the *consensus matrix* D , where D_{ij} is the number of partitions in which vertices i and j are assigned to the same community, divided by n_P .
3. Set all entries of D below a chosen threshold τ to zero.
4. Apply A on the graph with adjacency D (i.e. the weighted graph whose edge weights are the entries of D) n_P times, yielding n_P new partitions.
5. If the n_P partitions are all equal, stop and take that partition as the consensus. Otherwise go back to step 2.

This iteration emphasises structure that is stable across runs: pairs of nodes that are rarely grouped together get weak or zero weight in D , so the consensus graph sharpens the community signal (Figure 8). For the GRUMP analysis we use this iterative procedure with Louvain as A , $n_P = 30$, and $\tau = 0.5$, and report the consensus partition (and the number of iterations until convergence) for downstream interpretation.

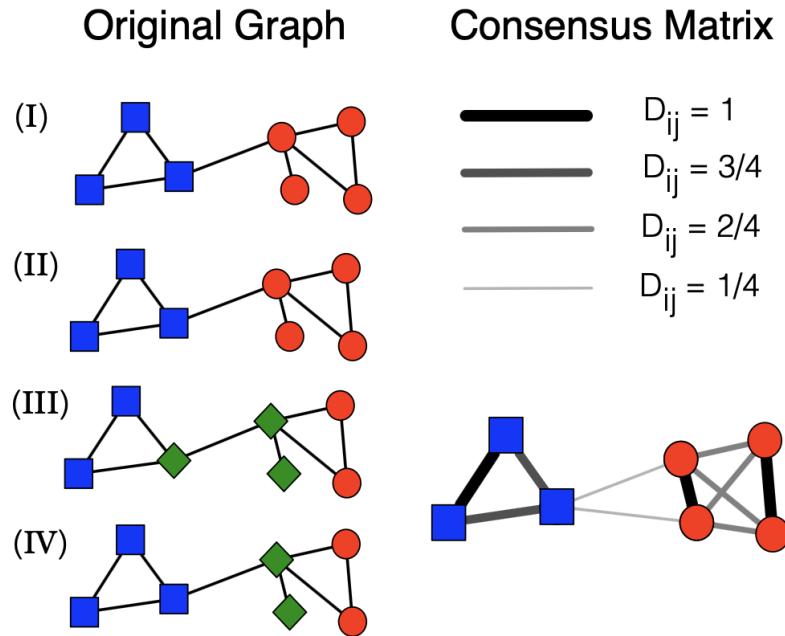


Figure 8: Consensus clustering: run algorithm repeatedly, build co-association matrix, then re-cluster to obtain stable partition (from Fortunato & Hric, 2016).

2.5 Method selection

Table 1 summarizes the methods discussed above: for each algorithm we give a short description, its main strengths, and its limitations. No single method is best for all networks. The choice depends on the type of graph (e.g. weighted, directed), the scale of structure of interest, and whether the number of communities is known or must be inferred (Fortunato and Hric, 2016).

Table 1: Community detection methods: brief overview of strengths and limitations.

Method	Description	Strengths	Limitations
Louvain	Greedy modularity maximisation. Aggregate nodes, then repeat.	Fast and scalable. Very widely used.	Resolution limit. Modularity trap (many local optima).
Infomap	Minimise description length of random walk (map equation).	No fixed null model. Multi-scale.	Resolution-type issues. Depends on flow model.
Spectral / non-backtracking	Use spectrum of (e.g.) non-backtracking matrix.	Can infer no. of communities q . Principled.	Sensitive to model assumptions. Bulk vs outliers.
Fast Greedy	Agglomerative modularity. Merge by ΔQ .	Fast. Yields dendrogram.	Resolution limit. Greedy.
Leading Eigenvector	Leading eigenvector of modularity matrix. Recursive split.	Spectral. Hierarchical.	Bisection bias. Recursive only.
Walktrap	Distance from random-walk probabilities. Then cluster.	Intuitive (walks stay in modules).	Fixed walk length. Resolution.
Consensus clustering	Run algorithm repeatedly. Co-association or representative partition.	Stability. Reduces run-to-run noise.	Computationally heavier. Needs many runs.

Fortunato and Hric (2016) stress that there is no one-size-fits-all choice: the best method depends on the type of network, the scale of structure one is interested in, and whether the number of communities is known or must be inferred. They recommend two practices. First, *significance testing*: detected communities should deviate significantly from what would be expected under a null model (e.g. the configuration model). If the observed structure is compatible with random fluctuations, it should not be interpreted as meaningful (Fortunato and Hric, 2016). Second, *comparative use of several algorithms*: because different methods optimise different objectives (e.g. modularity vs. flow compression), they can yield different partitions. Trying several and, where possible, using consensus clustering improves stability and guards against over-interpreting the output of any single run (Fortunato and Hric, 2016).

3 Results

This section presents the application of the methods described above to the GRUMP data. We first describe the data source and network construction and report basic statistics for the Atlantic and Pacific association networks. We then report community structure and modularity under several detection algorithms, compare partitions across methods, and use iterative consensus clustering to obtain a stable partition for interpretation. Finally we interpret the consensus modules using taxonomy and node-level environmental summaries and compare how structural clusters align with environmental or taxonomic gradients in each basin.

3.1 Network structure and descriptive statistics

3.1.1 Data source and network description

The data come from the Global Repeating Upper-ocean Microbial Patterns (GRUMP) database (McNichol et al., 2025), comprising roughly 1000 samples from ocean cruises between 2003 and 2020. Counts are available for bacteria, archaea, and eukaryotes as amplicon sequencing variants (ASVs), with taxonomic annotation from Silva (Quast et al., 2013) (bacteria/archaea) and PR2 (Guillou et al., 2013) (eukaryotes). Raw data were converted to sample-by-ASV count tables, 16S chloroplast sequences were removed to avoid double-counting, and a single `phyloseq` (McMurdie and Holmes, 2013) object (OTU table, taxonomy, and sample metadata) was built. For network inference, samples were restricted to the upper 200 m and split by ocean basin (Atlantic and Pacific). Within each basin, a Milici-style prevalence and abundance filter was applied: ASVs had to contribute more than 0.001% of total reads in that basin and satisfy at least one of three criteria—(i) present at relative abundance $> 1\%$ in at least one sample, (ii) present at $> 0.1\%$ in at least 2% of samples, or (iii) present in at least 5% of samples at any non-zero abundance. Filtering was applied separately in each basin, and zero-abundance taxa per basin were pruned. So that both basin-level networks are comparable, the same set of taxa was used for the analysis—the intersection of taxa retained in the Atlantic and Pacific basins. Association networks were then estimated separately for each basin using SPIEC-EASI (Kurtz et al., 2015) via the NetCoMi package (Peschel et al., 2021), with the Meinshausen–Bühlmann graphical lasso and stability selection (StARS: 5 replicates, 20 lambda values, threshold 0.01). The resulting adjacency matrices can in principle be signed. In the analysed export all edge weights are positive (co-occurrence strength), as negative associations were not retained in the exported networks. Edgelists and node-level metadata (e.g. taxonomy, environmental summaries) were exported for each basin. The analyses reported here use these provided edgelists and node features.

3.1.2 Basic statistics: Atlantic and Pacific

Both basin-level networks share the same node set (the intersection of taxa retained in each basin), so that basin comparison is not confounded by different taxon lists. Each network has $n = 3175$ nodes (ASVs) and approximately 25 000 edges (Atlantic: 25 173, Pacific: 25 383). Density is low (0.005 in both basins), as expected for sparse microbial association

networks. Table 2 summarises basic statistics and Table 3 gives connectivity metrics (clustering coefficient, average path length and diameter in number of steps, assortativity, and connected components). Average path length and diameter were computed with edge weights ignored (unweighted, i.e. number of hops) so that values are comparable across basins and interpretable as steps. The Atlantic network has two connected components (99.6% of nodes in the largest). The Pacific network is connected. Clustering is higher in the Atlantic (0.097) than in the Pacific (0.062). Both networks show slight negative degree assortativity. Mean edge weight is similar in both basins (≈ 0.31).

Table 2: Basic network statistics: Atlantic and Pacific GRUMP association networks ($n = 3175$ nodes each).

Metric	Atlantic	Pacific
Nodes	3175	3175
Edges	25 173	25 383
Density	0.0050	0.0050
Avg. degree	15.86	15.99
Max. degree	220	460
Mean edge weight	0.3126	0.3116
Median edge weight	0.3026	0.3026

Table 3: Connectivity metrics: Atlantic and Pacific. Path length and diameter are in number of steps (unweighted).

Metric	Atlantic	Pacific
Clustering coefficient	0.0972	0.0624
Average path length	3.88	3.74
Diameter	16	11
Assortativity (degree)	-0.1002	-0.0943
Connected components	2	1
Largest component size	3162	3175
% in largest component	99.6	100

3.2 Community structure and modularity

3.2.1 Detection results and modularity by algorithm

We apply five algorithms to the GRUMP Atlantic and Pacific networks: Louvain (Blondel et al., 2008), Infomap (map equation), Fast Greedy, Leading Eigenvector, and Walktrap (all via `igraph`), and we use consensus clustering over multiple Louvain runs to obtain a stable partition for downstream analysis. We compare partitions across algorithms using normalized mutual information (NMI) and variation of information (VI), and we use taxonomy and environmental variables to *interpret* the detected modules rather than to

validate them. Table 4 gives the number of communities and modularity Q per algorithm and basin. Figure 9 shows the same quantities as grouped bar charts for both basins. The number of communities varies strongly across methods: from 5–6 (Fast Greedy, Leading Eigenvector in the Pacific) to 49–57 (Walktrap, Infomap). Modularity-based methods (Louvain, Fast Greedy, Leading Eigenvector, Walktrap) typically return fewer, larger modules and higher Q . Infomap, which optimises the map equation rather than modularity, yields more, smaller communities. For every algorithm, the Atlantic network attains higher modularity than the Pacific (Figure 9), suggesting a somewhat clearer modular structure in the Atlantic. This comparison is descriptive rather than inferential, as modularity was not tested against a null model and should not be interpreted as a formal significance result. Because there is no ground truth, we do not choose a single “best” algorithm. The partition similarity measures below quantify agreement across methods, and we then use consensus clustering to obtain a robust partition.

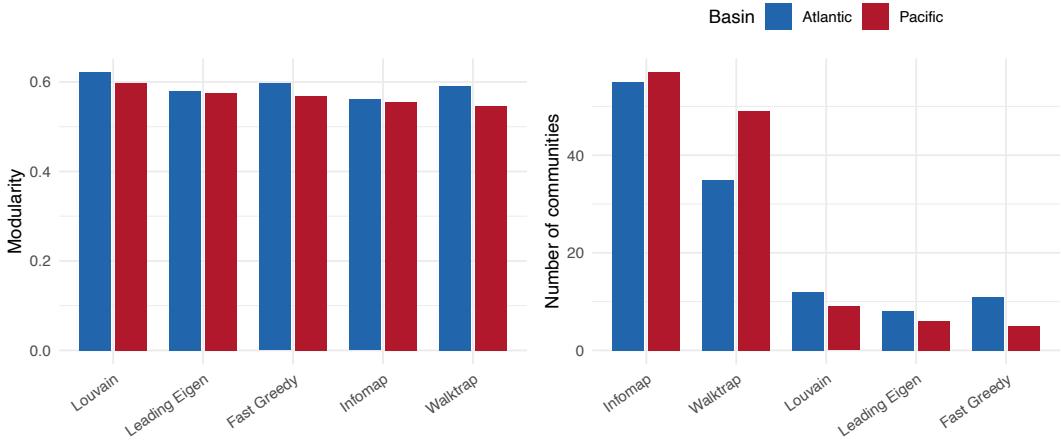


Figure 9: Community detection by algorithm and basin. Left: modularity Q . Right: number of communities K . Atlantic (blue) attains higher modularity than Pacific (red) for every algorithm. Infomap and Walktrap return many more communities than the modularity-based methods.

Table 4: Community detection: number of communities and modularity Q by algorithm and basin (single run per algorithm).

Algorithm	K (Atl.)	K (Pac.)	Q (Atl.)	Q (Pac.)
Louvain	12	9	0.622	0.598
Infomap	55	57	0.561	0.555
Fast Greedy	11	5	0.597	0.569
Leading Eigenvector	8	6	0.580	0.575
Walktrap	35	49	0.591	0.546

3.2.2 Algorithm comparison and stability

We quantify agreement between the five algorithm partitions using NMI (1 = identical, 0 = independent) and VI (0 = identical, larger = more different) (Meilă, 2007). Figure 10 shows the pairwise NMI matrix for Atlantic and Pacific. Pairwise NMI values lie in a moderate range (approximately 0.51–0.67): no two partitions are identical, but all share some structure. Modularity-based methods (Louvain, Fast Greedy, Walktrap) agree more with each other (higher NMI, lower VI), as seen in the block structure of the NMI matrices. Infomap and Leading Eigenvector disagree more with the others, consistent with their different objectives. The pattern is similar in both basins. We use the iterative consensus procedure described in the Methods (Lancichinetti and Fortunato, 2012): build the consensus matrix D from $n_P = 30$ Louvain runs, set entries below $\tau = 0.5$ to zero, re-apply Louvain on the graph with adjacency D , and repeat until all n_P partitions agree. The resulting consensus partition is used in the following section for taxonomic and environmental interpretation (Blondel et al., 2008).

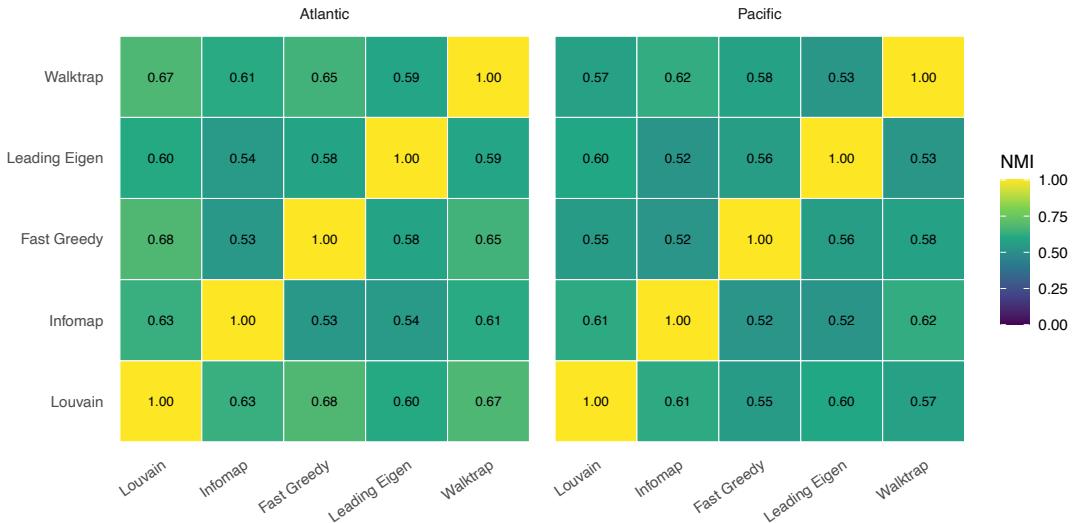


Figure 10: Normalized mutual information (NMI) between algorithm partitions, by basin. Colour indicates NMI (1 = identical partitions, 0 = independent). Diagonal entries are 1. Off-diagonals show moderate agreement (roughly 0.51–0.67), with stronger agreement among modularity-based methods (Louvain, Fast Greedy, Walktrap) and weaker agreement involving Infomap and Leading Eigenvector.

3.3 Ecological interpretation and basin comparison

We interpret the consensus communities using taxonomy (Phylum, Class) and node-level environmental summaries (mean temperature, salinity, depth across samples where each ASV was observed). This is *interpretation*, not validation: structural modules need not align with taxonomic or environmental groupings (Fortunato and Hric, 2016).

3.3.1 Environmental and taxonomic associations

We do not compute NMI between the structural partition and a taxonomic grouping. The goal is to attach ecological meaning to the modules, not to judge success by alignment. To visualise the co-occurrence network and how node attributes align with structure, we plot the same network layout once per basin, with nodes coloured in turn by (a) consensus module, (b) temperature, (c) absolute latitude, (d) depth, (e) salinity, and (f) Class (Figure 11, Figure 12). Node size reflects degree. The shared layout allows direct comparison: if modules match environmental or taxonomic gradients, the corresponding panels will show coherent colouring within spatial clusters. Otherwise the structural partition is largely independent of those covariates. Larger, single-panel versions of each colouring are provided in the Appendix (Sections A.1 and A.2).

3.3.2 Basin-specific patterns

The iterative consensus procedure converged in two iterations in both basins. The Atlantic network yields 11 consensus communities with modularity $Q = 0.613$ on the original graph. The Pacific yields 10 communities with $Q = 0.593$. Atlantic thus has slightly higher modularity and one more module than Pacific, consistent with the single-run algorithm comparison (Atlantic attained higher Q for every method). Mean community size is 289 (Atlantic) and 318 (Pacific), with median 176 (Atlantic) and 378 (Pacific). Pacific modules are therefore larger on average and more even in size (higher median), whereas the Atlantic has one more community and a lower median size, suggesting the presence of smaller modules. The network figures (Figures 11 and 12) allow a qualitative comparison of how module structure and environmental or taxonomic gradients align in each ocean.

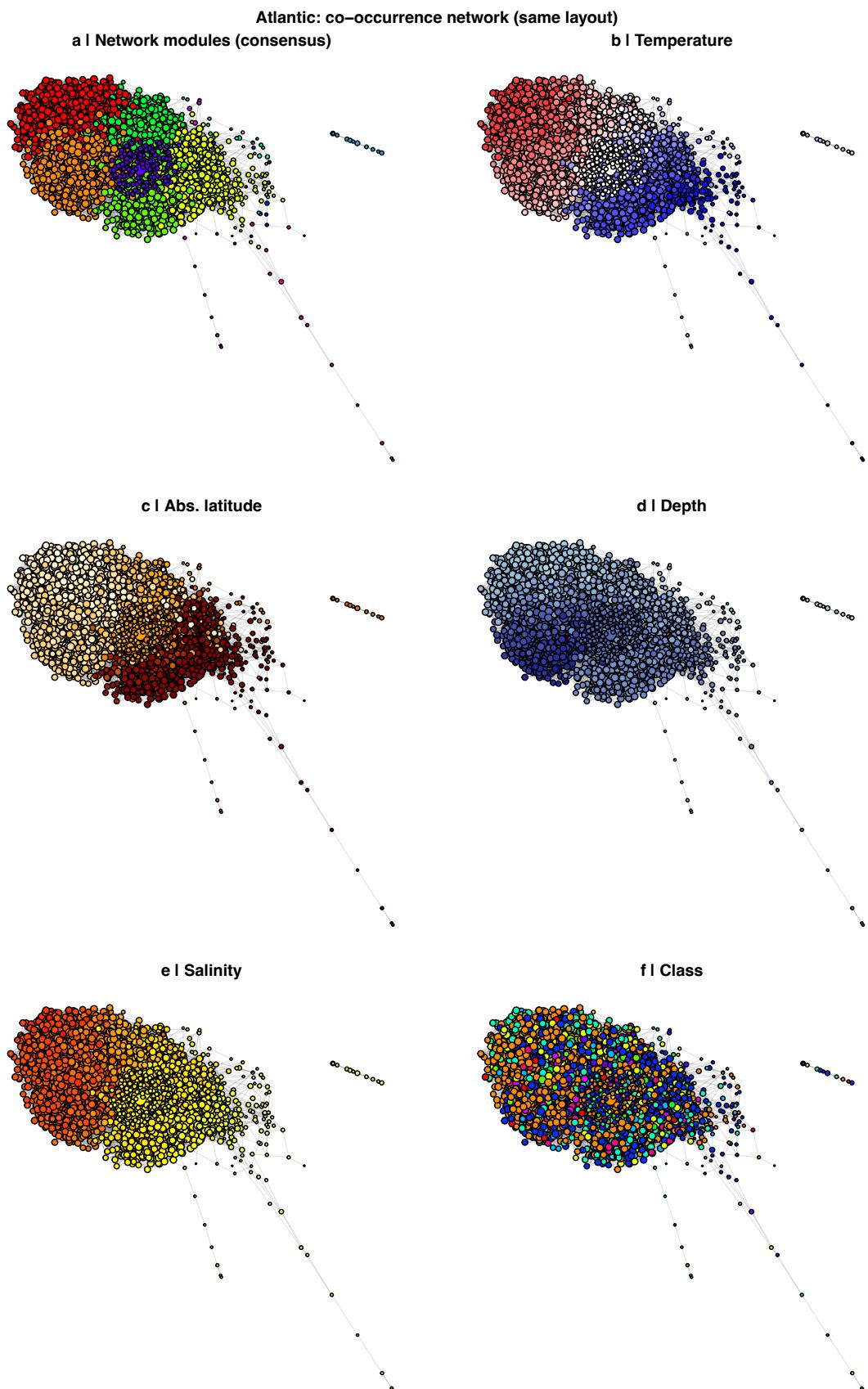


Figure 11: Atlantic co-occurrence network: same layout, six panels (3×2). (a) Consensus modules. (b)–(e) Environmental variables (temperature, absolute latitude, depth, salinity). (f) Taxonomy (Class). Node size \propto degree.

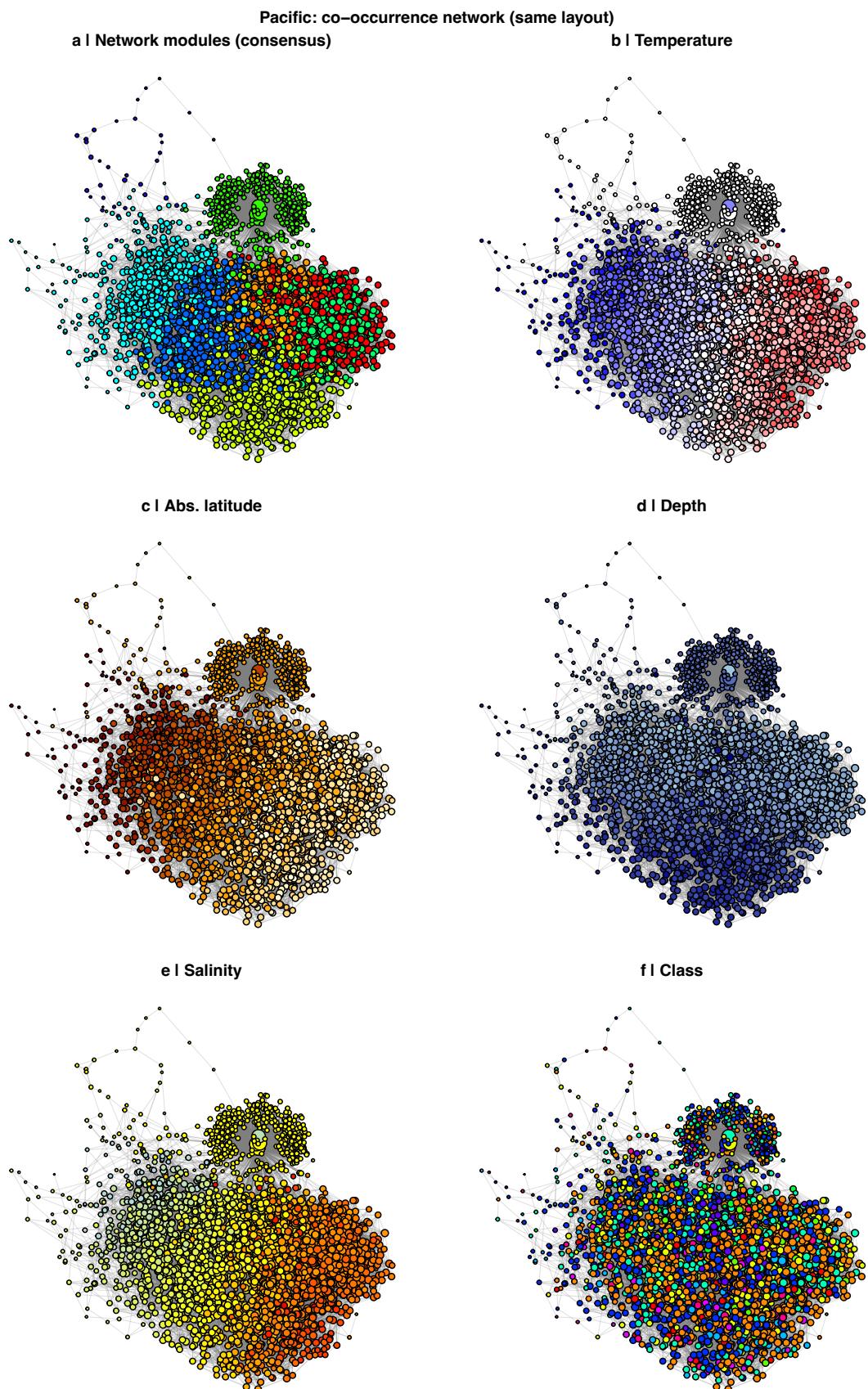


Figure 12: Pacific co-occurrence network: same layout, six panels (3×2). Panels as in Figure 11.

4 Conclusion

Community detection is a difficult field with no one-size-fits-all method. There is no unique, rigorous definition of what a community is; different algorithms optimise different objectives (e.g. modularity vs. flow compression), and results depend on the algorithm, the network, and the scale of structure of interest (Fortunato and Hric, 2016). That is why we applied several methods, compared partitions with NMI and VI, and used consensus clustering to obtain a stable partition for interpretation rather than relying on a single run. It also motivated our choice to *interpret* the detected modules with taxonomy and environment rather than to validate them by agreement with metadata. Below we summarise the findings, methodological lessons, and limitations.

4.1 Summary of Findings

We analysed microbial co-occurrence networks from the GRUMP database for the Atlantic and Pacific basins, using the same set of 3175 ASVs (nodes) and approximately 25 000 edges per basin. Both networks are sparse (density ≈ 0.005) and show similar mean degree and edge weights. The Atlantic has higher clustering (0.097 vs. 0.062), two connected components (with 99.6% of nodes in the largest), and a larger diameter than the Pacific. We ran five community detection algorithms (Louvain, Infomap, Fast Greedy, Leading Eigenvector, Walktrap) and found that the number of communities varied strongly across methods (from 5 to 57), and that the Atlantic attained higher modularity Q than the Pacific for every algorithm. Pairwise NMI between partitions was moderate (roughly 0.51–0.67), with greater agreement among modularity-based methods. We used iterative consensus clustering (Louvain, $n_P = 30$, $\tau = 0.5$) to obtain a stable partition for downstream analysis: the procedure converged in two iterations in both basins, yielding 11 consensus communities in the Atlantic ($Q = 0.61$) and 10 in the Pacific ($Q = 0.59$). Mean community size was 289 (Atlantic) and 318 (Pacific), with median 176 and 378 respectively—Pacific modules are larger on average and more even in size, while the Atlantic has one more community and at least one relatively small module. We interpreted these consensus modules using taxonomy (Phylum, Class) and node-level environmental summaries (temperature, salinity, depth, oxygen) and visualised the networks with a shared layout and multiple colourings. We did not validate partitions by alignment with metadata, but used metadata to attach ecological meaning to the structure.

4.2 Methodological Insights

Applying multiple algorithms made clear that there is no unique “best” partition when ground truth is unknown: modularity-based methods returned fewer, larger communities and higher Q , while Infomap and Walktrap often returned many more, smaller communities. Reporting NMI and VI between partitions quantified this variation and showed that modularity-based methods agree more with each other than with Infomap or Leading Eigenvector. Consensus clustering over many Louvain runs (following Lancichinetti and Fortunato (2012)) provided a single, reproducible partition for interpretation and avoided dependence on a single random run. We deliberately refrained from judging success by agreement with taxonomy or environment. Instead we used the framework of Fortunato

and Hric (2016) to *interpret* the structural modules with metadata rather than to validate them. The Methods section reviewed the role of significance testing (e.g. against the configuration model) and the resolution and detectability limits. In the analysis we reported modularity and partition similarity and used consensus as a stability device. We did not perform formal significance testing and leave null-model comparison and benchmark evaluation for future work (see Future Directions).

4.3 Ecological Implications

The consensus communities were interpreted in light of taxonomic composition (e.g. top phyla and classes per module) and environmental covariates (mean temperature, salinity, depth, oxygen per module). The seven-panel network figures (same layout, colourings by consensus module, temperature, latitude, depth, salinity, Phylum, and Class) allow a qualitative view of how structural clusters align with environmental or taxonomic gradients in each basin. We did not assert that modules “match” taxonomy or environment. Structural communities need not align with external covariates, and the goal was to attach ecological meaning (e.g. potential warm-surface vs. deep assemblages, or phylum-dominated vs. mixed modules) rather than to score the partition. Basin-specific patterns—higher modularity and one more community in the Atlantic, larger and more even-sized modules in the Pacific—suggest that the two basins differ in how clearly modular the co-occurrence structure is, consistent with the single-run algorithm comparison. These patterns can inform hypotheses about basin-level differences in microbial biogeography and the role of environmental gradients in shaping association structure.

4.4 Limitations

Several limitations affect the interpretation of our results. (1) **Resolution limit** (Fortunato and Barthélemy, 2007): modularity-based methods, including Louvain, cannot detect communities smaller than a scale of order $\sqrt{2m}$. Biologically meaningful small clusters may have been merged. (2) **Detectability**: in sparse networks, community structure can be information-theoretically undetectable (Decelle et al., 2011). We did not test whether the GRUMP networks lie above or below the detectability threshold. (3) **Sparsity**: both networks are very sparse ($\approx 25k$ edges for 3175 nodes), so modularity and partition quality should be interpreted cautiously. (4) **Data and inference**: the node set and edges are fixed by the GRUMP filtering and SPIEC-EASI (NetCoMi) pipeline. Different filters or inference methods would yield different networks and hence different community structure. (5) **Single consensus strategy**: we used Louvain-based iterative consensus only. Consensus over multiple algorithms or other stability schemes could yield different partitions. (6) We did not perform formal significance testing of observed modularity against the configuration model in the reported pipeline.

4.5 Future Directions

Natural extensions include applying the same pipeline to other ocean basins (e.g. Indian Ocean, Southern Ocean) and to temporal or seasonal slices of the data to compare community structure across space and time. Deeper integration with environmental drivers

(e.g. gradients, fronts, nutrient availability) could help explain why the Atlantic shows higher modularity and a different distribution of community sizes than the Pacific. On the methodological side, *formal significance testing* of observed modularity (or other structure) against the configuration model would assess whether the detected communities are statistically meaningful rather than due to chance; this was not carried out in the present analysis. Comparing algorithms on benchmark networks (e.g. LFR) with planted partitions would characterise their behaviour before application to real data. Exploring consensus over multiple algorithms (e.g. Louvain, Infomap, Walktrap) or alternative stability criteria could yield more robust or interpretable partitions for ecological interpretation.

A Appendix

A.1 Atlantic network: individual panels

Each panel below shows the Atlantic co-occurrence network with the same layout. Node size \propto degree. Larger, single-panel figures can be found in the repository (see Electronic appendix).

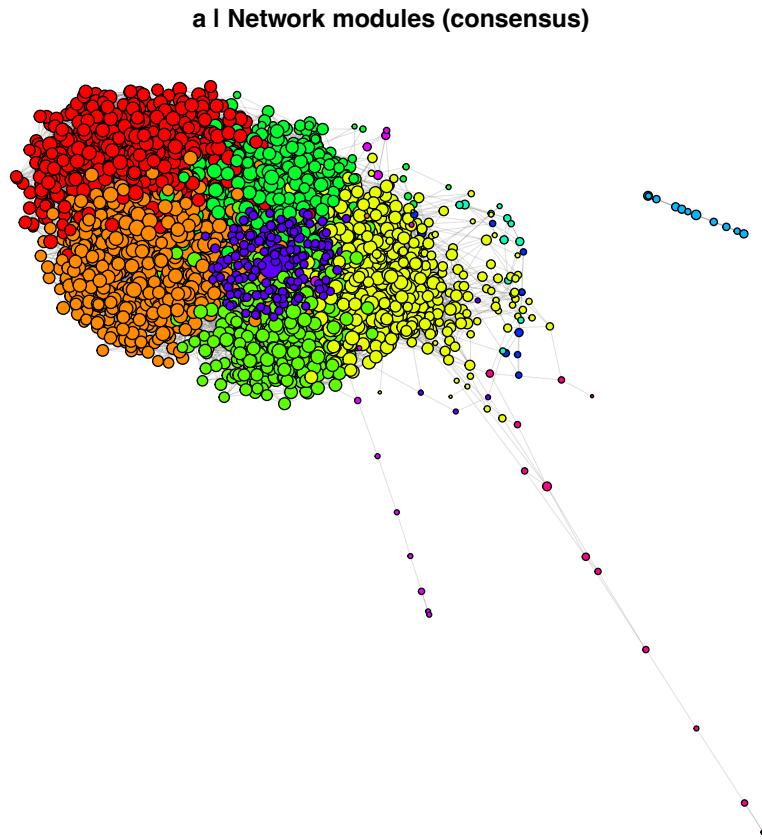


Figure 13: Atlantic: consensus modules.

A.2 Pacific network: individual panels

Same layout and colourings as for the Atlantic. See main text and Atlantic panels above for interpretation.

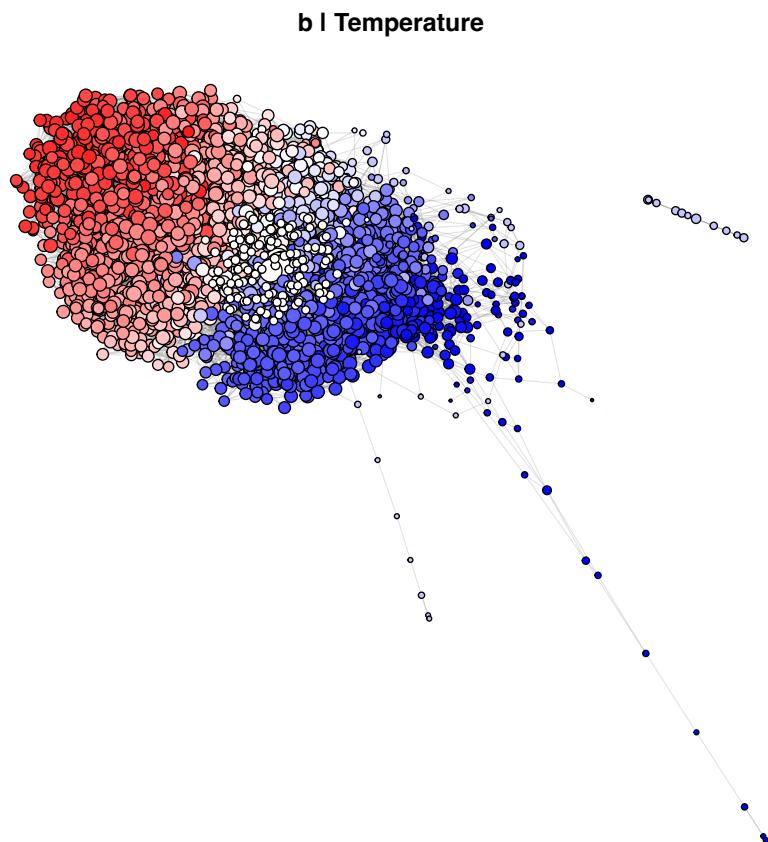


Figure 14: Atlantic: temperature.

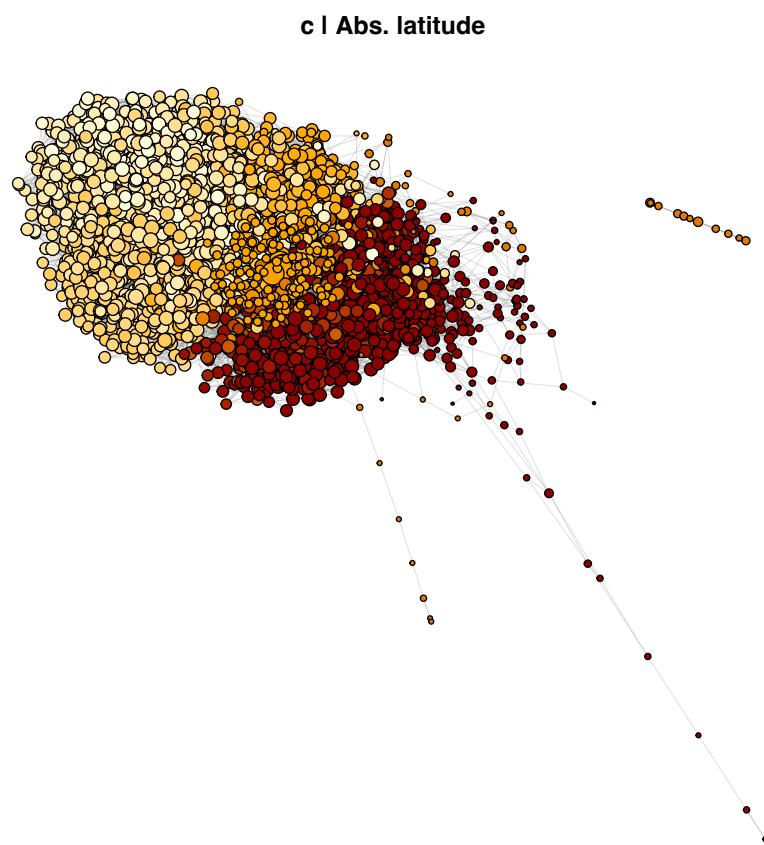


Figure 15: Atlantic: absolute latitude.

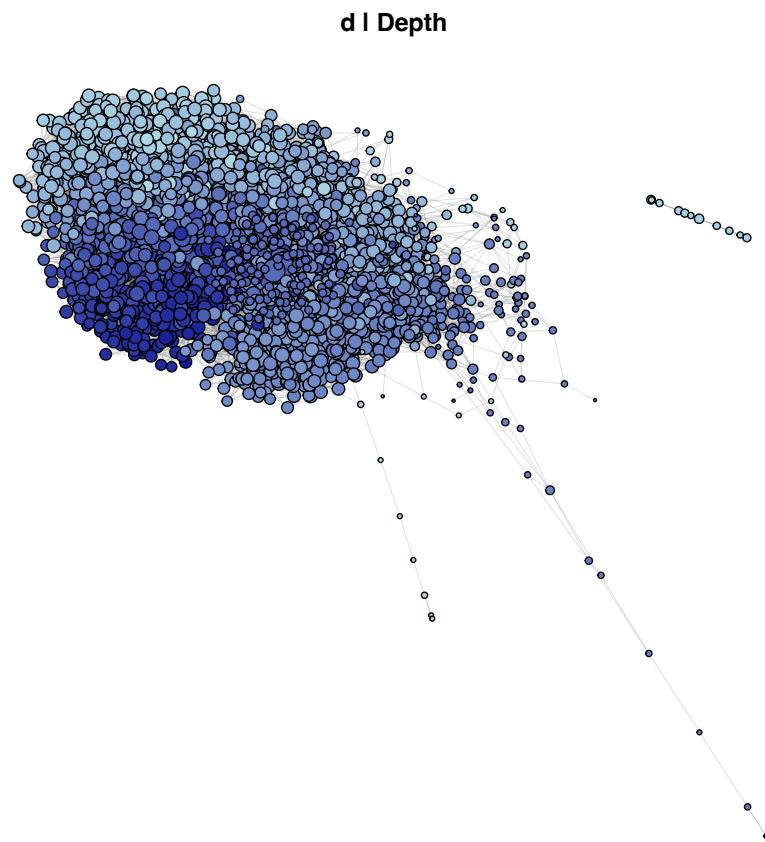


Figure 16: Atlantic: depth.

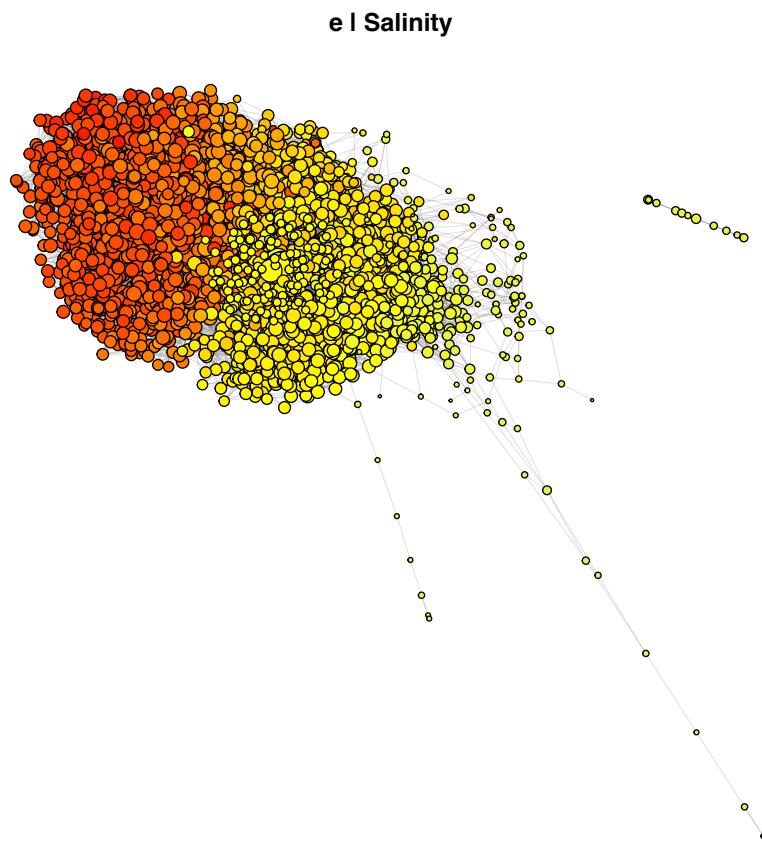


Figure 17: Atlantic: salinity.

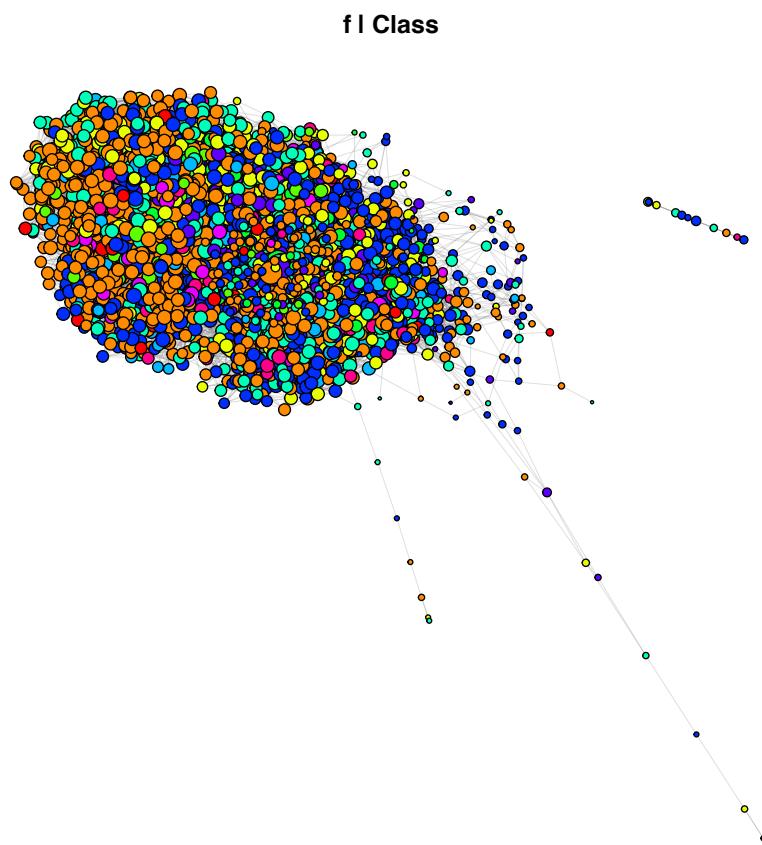


Figure 18: Atlantic: Class (taxonomy).

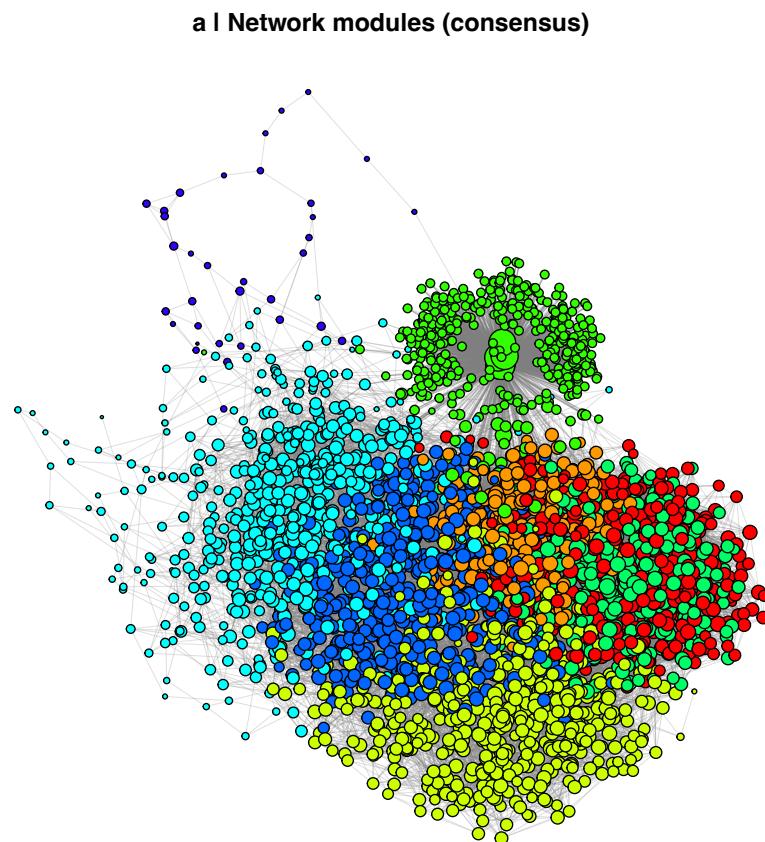


Figure 19: Pacific: consensus modules.

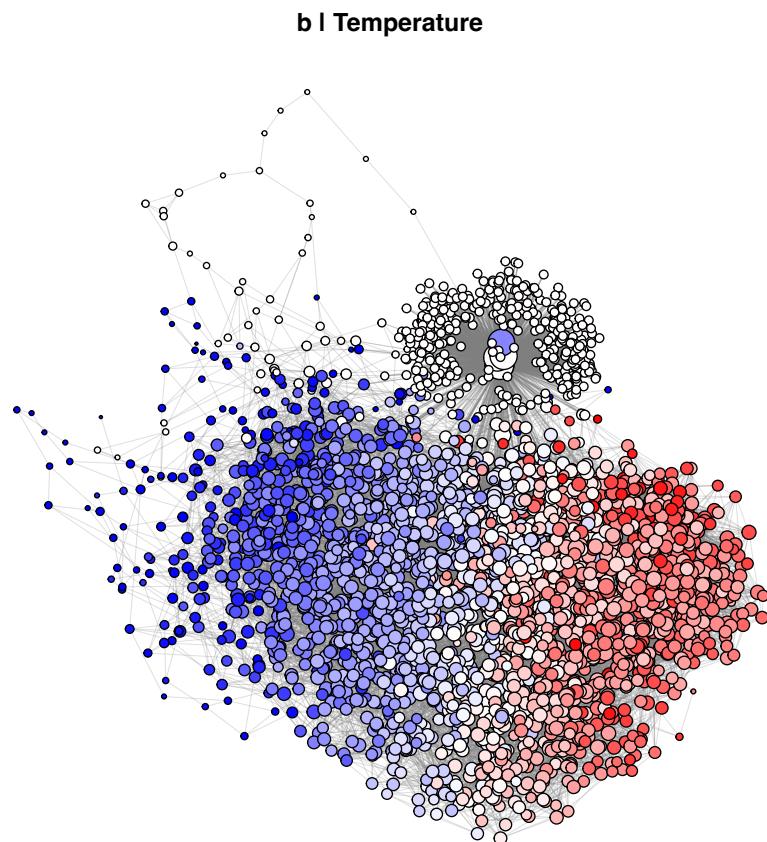


Figure 20: Pacific: temperature.

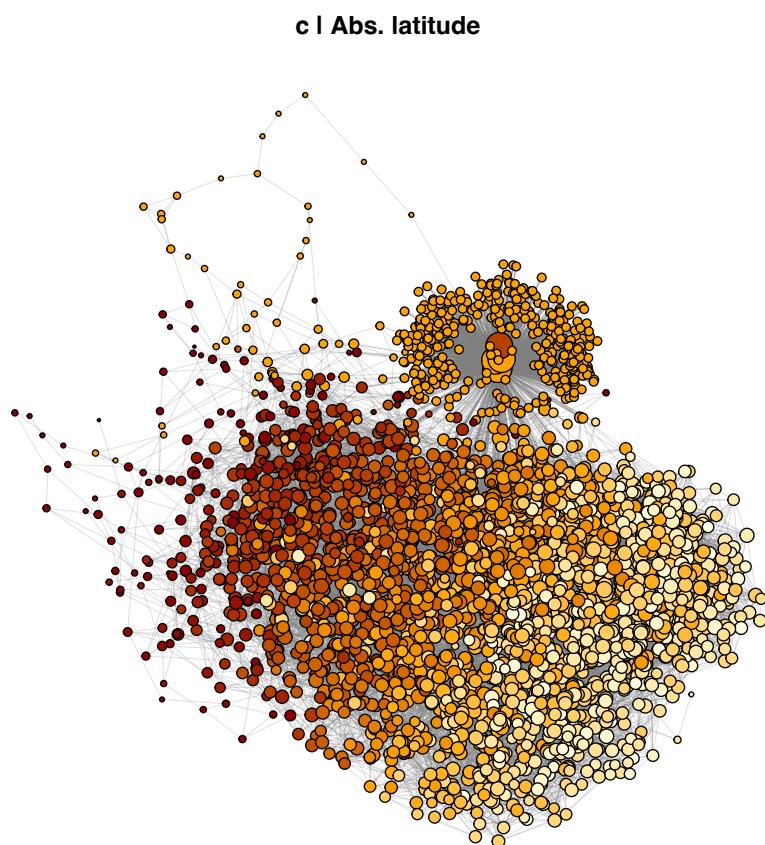


Figure 21: Pacific: absolute latitude.

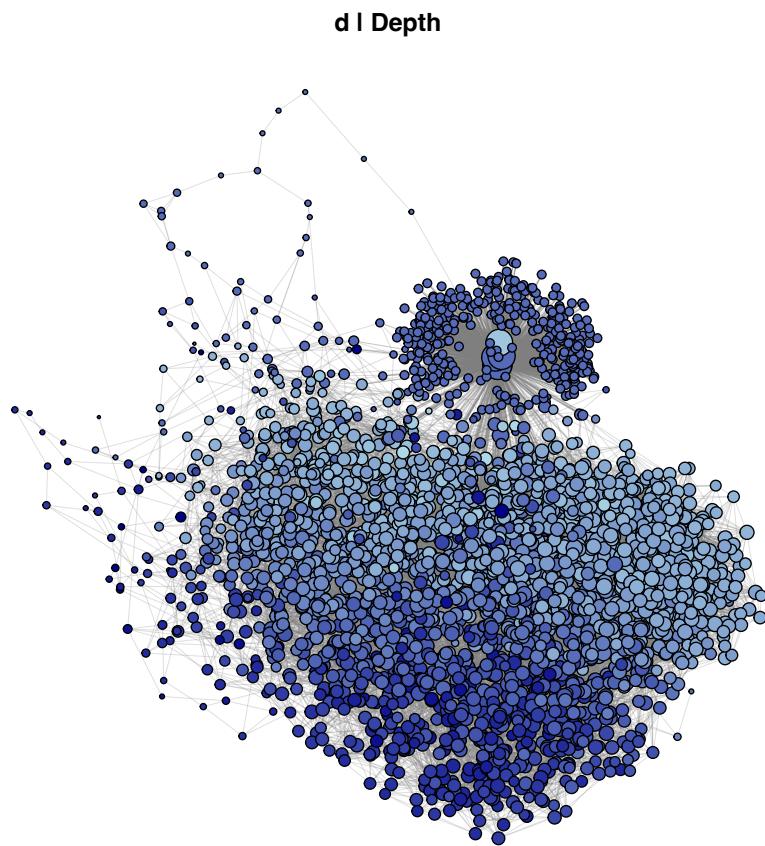


Figure 22: Pacific: depth.

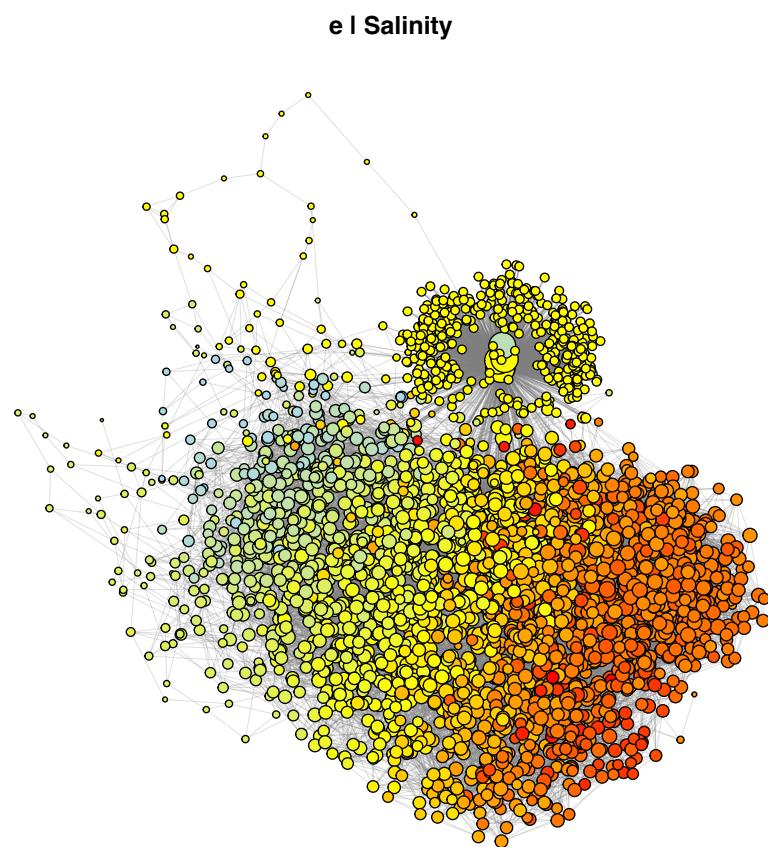


Figure 23: Pacific: salinity.

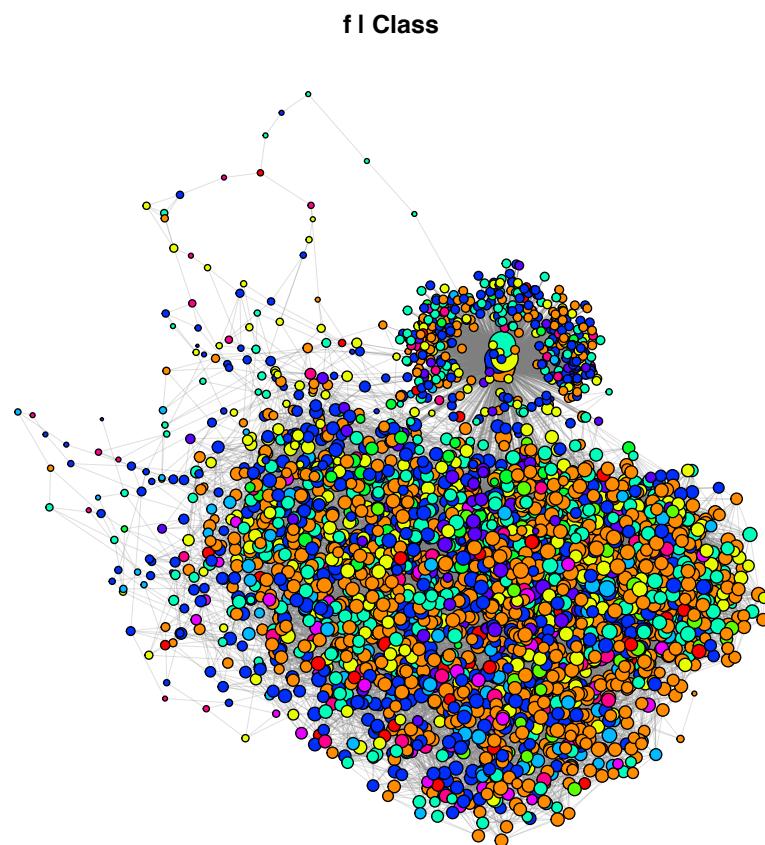


Figure 24: Pacific: Class (taxonomy).

B Electronic appendix

Data, code and figures are provided in this Github repository. <https://github.com/mafrei6/microbial-network-communities>

References

- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10): P10008.
- Clauset, A., Newman, M. E. J. and Moore, C. (2004). Finding community structure in very large networks, *Physical Review E* **70**(6): 066111.
- Decelle, A., Krzakala, F., Moore, C. and Zdeborová, L. (2011). Inference and phase transitions in the detection of modules in sparse networks, *Physical Review Letters* **107**(6): 065701.
- Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection, *Proceedings of the National Academy of Sciences* **104**(1): 36–41.
- Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide, *Physics Reports* **659**: 1–44.
- Glover, C. and Kempton, M. (2020). Spectral properties of the non-backtracking matrix of a graph.
URL: <https://arxiv.org/abs/2011.09385>
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Bur-gaud, G., de Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Ed-vardsen, B., Holzmann, M., Kooistra, W. H., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A. L., Siano, R., Stoeck, T., Vaultot, D., Zimmermann, P. and Christen, R. (2013). The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy, *Nucleic Acids Res* **41**(D1): D597–D604.
- Holland, P., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps, *Social Networks* **5**: 109–137.
- Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L. and Zhang, P. (2013). Spectral redemption in clustering sparse networks, *Proceedings of the National Academy of Sciences* **110**(52): 20935–20940.
- Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks, *PLoS Comput Biol* **11**(5): e1004226.
- Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks, *Scientific Reports* **2**: 336.
- Lancichinetti, A., Fortunato, S. and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms, *Physical Review E* **78**(4): 046110.

- McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data, *PLoS One* **8**(4): e61217.
- McNichol, J., Williams, N. L. R., Raut, Y. et al. (2025). Characterizing organisms from three domains of life with universal primers from throughout the global ocean, *Sci Data* **12**: 1078.
- Meilă, M. (2007). Comparing clusterings—an information based distance, *Journal of Multivariate Analysis* **98**(5): 873–895.
- Newman, M. E. J. (2006). Modularity and community structure in networks, *Proceedings of the National Academy of Sciences* **103**(23): 8577–8582.
- Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks, *Physical Review E* **69**(2): 026113.
- Peschel, S., Müller, C. L., von Mutius, E., Boulesteix, A.-L. and Depner, M. (2021). NetCoMi: network construction and comparison for microbiome data in R, *Brief Bioinform* **22**(4): bbaa290.
- Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks (long version), *arXiv preprint physics/0512106* .
URL: <https://arxiv.org/abs/physics/0512106>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, *Nucleic Acids Res* **41**(D1): D590–D596.
- Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure, *Proceedings of the National Academy of Sciences* **105**(4): 1118–1123.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, February 28th, 2026

A handwritten signature in black ink, appearing to read "Maximilian Frei".

Maximilian Frei