Exploring microbial community structure and resilience through visualization and analysis of microbial co-occurrence networks

by

Sarah Isa Esther Perez

B.Sc. Honours Physics, McGill University, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Masters of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

June 2015

Abstract

Cultivation-independent microbial ecology research relies on high throughput sequencing technologies and analytical methods to resolve the infinite diversity of microbial life on Earth. Microorganisms live in communities driven by genetic and metabolic processes as well as symbiotic relationships. Interconnected communities of microorganisms provide essential functions in natural and human engineered ecosystems. Modelling the community as an inter-connected system can give insight into the community's functional characteristics related to the biogeochemical processes it performs. Network science resolves associations between elements of structure to notions of function in a system and has been successfully applied to the study of microbial communities and other complex biological systems. Microbial co-occurrence networks are inferred from community composition data to resolve structural patterns related to ecological properties such as community resilience to disturbance and keystone species. However, the interpretation of global and local network properties from an ecological standpoint remains difficult due to the complexity of these systems creating a need for quantitative analytical methods and visualization techniques for co-occurrence networks.

This thesis tackles the visualization and analytical challenges of modelling microbial community structure from a network science approach. First, Hive Panel Explorer, an interactive visualization tool, is developed to permit data driven exploration of topological and data association patterns in complex systems. The effectiveness of Hive Panel Explorer is validated by resolving known and novel patterns in a model biological network, the *C. elegans* connectome. Second, network structural robustness analysis methods are applied to study microbial communities from timber harvested forest soils from a North American long term soil

productivity study. Analyzing these geographically dispersed soils revealed biogeographic patterns of diversity and enabled the discovery of conserved organizing principles shaping microbial community structure. The capacity of robustness analysis to identify key microbial community members as well as model shifts in community structure due to environmental change is demonstrated. Finally, this work provides insight into the relationship between microbes and their ecosystem, and characterizing this relationship can help us understand the organization of microbial communities, survey microbial diversity and harness its potential.

Preface

The sections in this work have not yet been published. Chapter 2 and 3 are in the process of being submitted to peer reviewed scientific journals in the coming months.

Chapter 1 Sarah Perez wrote the main text with input from Steven J. Hallam. Aria S. Hahn provided input and feedback overall in particular on the soil ecology section.

Chapter 2 Hive Panel Explorer is an explorative visualization built by Sarah E. I. Perez. The design development process was conducted by Sarah E. I. Perez with support from Aria S. Hahn who provided user-based feedback and insight. Martin Krzywinski, the developer of hive plots (off of which this tool is built), provided feedback on the figures, the structure of the text for improved readability, and the content of the text. The main text was written by Sarah E. I. Perez with editorial input from Steven J. Hallam.

Chapter 3 The design of the methodological procedure for the construction and analysis of the microbial community networks was developed by Sarah E. I. Perez with feedback from Aria S. Hahn and Steven J. Hallam. Sarah E. I. Perez wrote the Python scripts to conduct the network analysis of the Long term Soil Productivity project data. This project is part of a multi-lab effort and data collection was undertaken by numerous scientist (see associated public reports). Network computation and inference was primarily conducted by Aria S. Hahn with guidance from Steven J Hallam and Karoline Faust and assistance from Sarah E. I. Perez. Networks were built using the CoNet software on high-performance computational resources provided

by Compute Canadas Western Canadian Compute Consortium. Sarah E. I. Perez wrote the main text and figures with editorial support from Steven J. Hallam.

Throughout this dissertation the word *we* refers to Sarah E. I. Perez unless otherwise stated. None of the work encompassing this dissertation required consultation with the Univeristy of British Columbia Research Ethics Board.

Table of Contents

List of Tables

List of Figures

Glossary

BC betweenness centrality

DNA deoxyribonucleic acids, a molecule that encodes genetic information

HYPE Hive Panel Explorer

JP Jack Pine zone, an LTSP study ecozone

LCC largest connected component of a network

LTSP Long Term Soil Productivity, a study of timber harvesting in North American forest soils

MD Mediterranean zone, an LTSP study ecozone

OM organic matter removal, a type of treatment implemented in forest harvesting

OTU operational taxonomic unit, as defined by prokaryotic rRNA sequence similarity

PPI protein-protein interaction

RNA ribonucleic acids, a molecule that encodes genetic information

SBS Sub Boreal Spruce zone, an LTSP study ecozone

SSU small subunit of the ribosomal molecule

Acknowledgments

This academic journey has been quite an adventure. First, I was fortunate enough to get funding from CIHR and conduct three research rotations which gave me exposure to different fields within Bioinformatics. After joining the Hallam lab, I was surrounded by supportive colleagues, notably Aria Hahn who quickly became a collaborator and a mentor. Aria, thank you for those endless discussions and your friendship. I also want to thank my supervisor, Dr. Steven Hallam, for engaging me in all those brainstorming sessions, for guiding me in my research and for ensuring that I get a rich learning experience. I offer my gratitude to my thesis committee, Dr. Anne Condon and Dr. Martin Hirst, for being both attentive and insightful in my committee meetings and for their helpful comments and suggestions. I have thoroughly enjoyed my experience at UBC and in the Bioinformatics training program, and would do it all over again if given the choice.

Finally, I am thankful for the incredible support that my partner, my family and close friends have given me throughout this journey.

À mes grand-parents, Renée et Jo, Gilou et Arié, qui m'ont appris le pouvoir de la persévérance, chacun de leur façon.

Chapter 1

Introduction

With an estimated cell abundance approaching 10^{30} cells [167], microorganisms represent the invisible majority of life on Earth. From the mesosphere to the lithosphere, microorganisms are adapted to thrive across a wide range of habitats and environmental conditions [167]. Interconnected communities of microorganisms provide essential functions in natural and engineered ecosystems and play integral roles in global scale biogeochemical processes [113, 114]. Resolving the complexity of these communities can reveal the inner workings of the Earth system with far reaching implications for biotechnology development and conservation. By harnessing the hidden metabolic potential of microbial communities, we can develop sustainable solutions in energy and materials production [6, 140], synthetic biology [111], medical diagnosis and therapeutics [28, 38] that are more in sync with the natural world.

Despite the impact that microbial communities have on the world around them, charting microbial community metabolism is extremely challenging as less than 1% of microbial diversity has been cultured in laboratory settings [114]. Advances in sequencing technology are beginning to bridge this cultivation gap through plurality sequencing of microbial community deoxyribonucleic acids (DNA) and ribonucleic acids (RNA) directly from the environment. Applications of these techniques enables the characterization of microbial taxonomic diversity and metabolic potential. Such environmental surveys have helped discover "who is there" (e.g., through taxonomic assessment based on ribosomal RNA gene abundance) and "what

are they doing" (e.g., metabolic reconstruction through functional gene and pathway analysis). Thus, sequencing technologies enable the study of microbial communities as structured and dynamic systems.

small subunit (SSU) rRNA sequencing of environmental samples allows direct evaluation of taxonomic identity, abundance and diversity in communities [29, 141]. These measurements provide knowledge of community structure, which can be modelled to develop biomarkers for environmental factors and processes. For example, studies have sequenced environmental SSU rRNA to evaluate the environmental impact of logging on soil productivity [76, 135, 136] and oil spills on coastal ecosystems [98]. Different statistical methods, such as the cluster analysis of samples and indicator species analysis of taxonomic distributions, model the presence and role of individual members, from rare to abundant taxa, and characterize community composition in relation to environmental parameter data.

Microbial communities rely on interconnected genetic and metabolic processes to drive matter and energy transformations. In particular, metagenomic studies have provided evidence that different reactions within a metabolic pathway may be performed by and distributed across different community members [29, 73, 171]. The genetic distribution of the community can also be altered through horizontal gene transfer [40, 51]. Furthermore, co-culture experiments have demonstrated that cooperation and competition drive community member dynamics: groups of taxa engage in a variety of positive, neutral and negative interactions [46, 54, 128, 138]. Though the analysis of individual community members can provide valuable insight into specific metabolic processes, holistic understanding of the ecosystem requires an awareness of the dynamic interconnections between community members. Thus, the "whole is greater than the sum of its parts" as evaluating both community composition and interactivity can build more elaborate ecosystem models [54, 73, 171]. Accordingly, microbial communities can be modelled as a dynamic system where taxonomic, genetic and metabolic distributions are interrelated.

The structure of an interactive system can be modelled by studying its connectivity [118]. Through network abstraction, the connective structure of a system can be expressed using nodes and edges: the nodes of the network represent the members of the system and the edges represent the relationships between members. For example, modelling a microbial community as a connected system, individual

taxa become nodes and their interactive relationships become edges. Studies have built community networks by applying co-occurrence analysis to taxonomic abundances obtained via SSU rRNA sequencing data [11, 54]. To construct the microbial co-occurrence network of a community, the significant positive and negative co-occurrences are evaluated and assigned as edges. Microbial ecologists have recently adopted this network approach to study both the taxonomic distribution and the interconnected structure of microbial communities [47, 54, 89, 105, 129, 169].

Network approaches are widely used to study relationships between the structure and the function of a system in the social, biological and technological sciences. Networks and their structure elucidate functional aspects of the modelled system such as the wiring efficiency of the *C. elegans* connectome [34], the vulnerability to attack of the World Wide Web [5], extinction dynamics in foodwebs [45, 122], and missing annotations in protein-protein interactomes [102], etc. In particular, structural properties of microbial co-occurrence networks have been characterized to infer biological attributes of the community such as its resilience to disturbance [76]. Despite the power and the promise of graph theory selecting the appropriate or optimal quantitative method to accurately discern patterns within complex systems remains a challenging enterprise. For instance, biological network studies have difficulty justifying which of the many different network centrality measurements should be used to identify nodes that are "important", "central", or even "essential" to the structure of the network and have difficulty interpreting the results of their measurements in relation to system properties [62, 85]. Along the same lines, the adaptation of certain quantitative methods from foodweb studies in macroecology to microbial co-occurrence network studies remains difficult to visualize, interpret and validate [55].

The visualization of complex systems can help reveal patterns, motivate analysis and generate hypotheses [125, 147]. In order to go beyond presenting known patterns and reveal new ones, visualizations need to be designed to permit interactive exploration [115, 125, 147]. The discovery of topological features and patterns can help drive a quantitative analysis of a network and formulate hypotheses inferring the modelled system's function from its underlying structure. Current network visualization techniques are not designed for interactive exploration. Network representations, such as adjacency matrices do not provide flexibility in adapting their

layout rules to systems [115]. Other representations such as force-directed layouts, are not suitable for large networks for they are often inconsistent and difficult to interpret. On the other hand, rule-based network layouts have been developed to create consistent and coherent network visualizations [96, 115]. In particular, hive plots is a rule-based network layout whose design attempts to provide a visual query language from which to organize and study networks using system properties [96]. However, these different network layouts have been developed to illustrate specific connectivity features and more flexible visualization designs are required to maximize the exploration of patterns in the network. Adapting hive plots to develop a versatile network visualization and combining this design with interactive features could allow for the exploration and interpretation of patterns in highly dimensional and complex networks.

This thesis outlines the development and application of a visualization tool and a quantitative modelling approach to study microbial co-occurrence networks constructed from SSU rRNA sequencing data from soil environments that is extensible to other forms of data. In the following chapter, I review the state of network science and complexity, describe network visualization and quantitative ecology tools, and explore their application to the study of microbial structure and function in natural and engineered ecosystems. In Chapter 2, I describe and evaluate the development and design of Hive Panel Explorer as an interactive network visualization tool and demonstrate its effectiveness on a known and well studied biological network: The *C. elegans* connectome. In Chapter 3, I demonstrate the application of Hive Panel Explorer (HYPE) to soil microbial communities from the Long Term Soil Productivity (LTSP) study based on samples from varied locations and ecosystems across North America [76, 136]. As the most diverse environment, the soil microbiome epitomizes the complexity of microbial communities and successfully characterizing their structure and function on local and global scales using the methods outlined in this thesis will be readily adaptable to study less diverse microbial communities. Furthermore, analyzing geographically dispersed sample collections may reveal biogeographic patterns of diversity and uncover conserved organizing principles shaping microbial community structure. Thus this thesis has the potential to provide insight into the complex relationships between individual microbes, their community and their environment.

1.1 Networks and complexity

As put by science writer Dorian Sagan [143]:

> Nature no more obeys the territorial divisions of scientific academic disciplines than do continents appear from space to be coloured to reflect the national divisions of their human inhabitants. For me, the great scientific satoris, epiphanies, eurekas, and aha! moments are characterized by their ability to connect.

While scientific breakthroughs are accomplished via cross-disciplinary synthesis, some disciplines' knowledge and methodologies lend themselves better to cross-disciplinary applications. In particular, network science, a mathematical methodology, has been applied in social, biological and technological sciences to model systems using networks [118]. Network models can be used to capture the connective structure of a system and promote the study of interconnectivity in nature. Connections can be drawn within and across many levels, from fundamental particle interactions to the influence of gravitational fields. Particularly in this Digital Age, our world has become more interconnected: people and knowledge are virtually and globally "hyperlinked" through social media and online databases.

Network science harnesses the potential of this interconnected world by studying its structure. Anchored in graph theory, the field of network science has developed to resolve these dynamic interconnected structures in a variety of systems as diverse as social circles, ecosystems, and the World Wide Web [118].

Graph theory is thought to have a evolved from a few seminal papers including one entitled "Seven bridges of Konisberg" published in 1736 by Leonhard Euler where he analyzes the topology of connected bridges to find a path which crosses every bridge once [17]. The concepts of system topology first introduced in this paper quickly evolved to model relationships between objects such as social interactions, predator-prey relations, and web links between articles on Wikipedia. An ensemble of relationships is called a graph and is denoted by the letter G. The graph G is formed from nodes, the objects, connected by edges, the relationships. In mathematical terminology, we say the graph $G = (N,E)$ is composed of the set of nodes N and the set of edges E [17, 118]. The topology of a graph is its

"shape" or connective structure. Certain topologies imply that a graph will have different structural properties [17, 118]. Graph theory encompasses the topological algorithms and metrics applied to graphs.

Whereas the mathematical concept is denoted a graph, once applied to a system with specific objects, the model becomes a network[1]. The following section motivates the modelling of a system as a network, provides an overview of graph theory methods and describes topological characteristics of biological networks more specifically.

1.1.1 From an interactive system to networks

Graph theory methods have been used to study the relationship between the structure and the function of complex systems by building social, biological and technological networks [118, 119, 159]. In social networks, nodes are typically people and the edges connecting these people represent an interaction between them [162]. Similarly, in foodwebs, the edges are trophic interactions (i.e. "who eats who") that connect species, the nodes, in an ecosystem [45, 122]. Graph theory measures such as degree distributions, modularity and connectance, can help formulate associations between elements of structure to notions of function in the system and within its parts [118].

Graphs come in all shapes in sizes: they can be directed or undirected, weighted or not, connected or not [118], as illustrated in Figure 1.1. Directed graphs are used to model systems where nodes have relationships with implicit direction, such as the synaptic connections between neurons in a connectome: one neuron, a node, fires a signal to another neuron via a synapse, a *directed* edge [71]. Weighted graphs model the relationships between nodes quantitatively by assigning a *weight* to each edge. For example, the strength of a friendship between two individuals in a social network can be encoded in the weight of the edge connecting them. Finally, a connected graph is one where all nodes can be reached by following a *path*, a sequence of connected edges. A graph is defined as *not connected* if a node or groups of nodes can't be reached by following a path and the graph is then said to have more than one *connected component*.

[1] In this thesis, both terms will be used according to the context: *graphs* when discussing theory and *networks* when discussing systems modelled using nodes and edges.

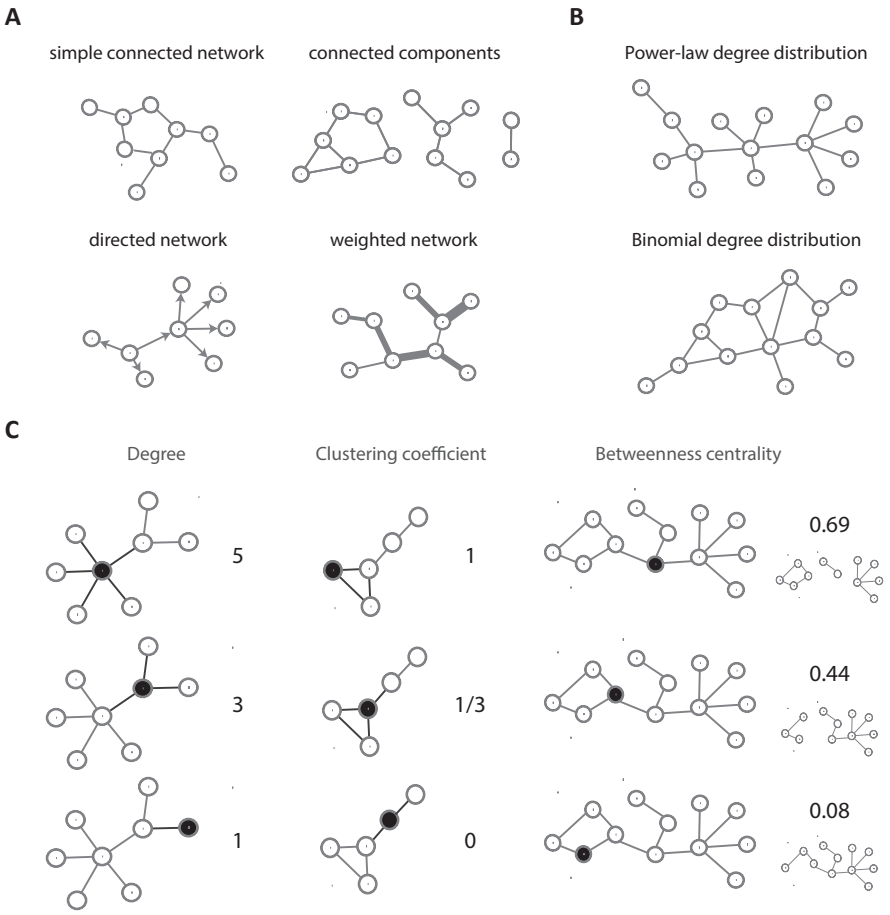

Figure 1.1: An overview of graph types and graph theory metrics. Graphs are composed of nodes and edges, here represented as circles and links between circles, respectively. A) Graphs can be simple, directed, weighted, completely connected or composed of connected components. B) Two graph topologies that differ in their degree distribution are shown: a power law degree distribution and the characteristic binomial degree distribution of randomly generated graphs. C) The topology of nodes is characterized by graph theory metrics including degree, clustering coefficient and centrality measures such as betweenness centrality. The number next to each graph corresponds to the metric value of the coloured-in node.

Structural properties of graphs are used to characterize properties of individual nodes, individual edges, paths, groups of nodes, groups of edges, connected components and the whole graph. Several review papers [17, 118–120] provide an overview of different network analyses and graph theory measures, a few of which are presented in the following section.

1.1.2 Graph theory

Many properties can be calculated to characterize a graph's topology: degree distribution, global clustering coefficient, diameter, average shortest path length, centrality measure, scale free index, modularity, etc [5, 118–120]. Whereas parameters such as the number of nodes and the number of edges provide a quantitative assessment of a graph's size, other measures such as the diameter of a graph, the longest shortest path between two nodes in the graph, evaluates the topological size of the graph: two graphs with orders of magnitude differences between their number nodes can have the same diameter. Similarly, the global interconnectivity of two graphs of different topological size or with different number of nodes and edges can be compare by measuring their connectance, the proportion of realized edges calculated from:

$$Connectance = \frac{|E|}{|N|(|N|-1)/2} \tag{1.1}$$

where $|E|$ is the number of edges and $|N|$ the number of nodes [118]. Just as statistical methods evaluate dependencies and similarities in multivariate datasets, graph theory measures assess repeated structures, connective patterns, partitions, and other complex topological patterns.

Graph theory measures take on different formulations depending on whether they are applied to directed or undirected, weighted or not weighted, connected or not connected graphs. However, we will present them as applied to unweighted undirected graphs for simplicity as other formulations are simply derivations of the ones presented here. Different measures are applied to characterize a system's structure at global and local scales by analyzing the resulting network [96]. They can be divided into two types: ones that measure properties of individual nodes and edges and ones that evaluate global properties of the graph and connected

components. An overview of these measures is illustrated in Figure 1.1.

Node degree

The degree of a node is simply the number of edges connected to it. In a social network, the node degree represents the number of social interactions of that node, which could be used for example to infer that individual's popularity [162]. This node property can be used to classify nodes by their connectivity.

$$\bar{d} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} e_{ij}}{|N|} = \frac{|E|}{|N|} \tag{1.2}$$

where $|N|$ is the number of nodes, $|E|$ is the number of edges, and the edge $e_{ij} = 1$ if the $i^{th} and j^{th}$ nodes are connected, otherwise $e_{ij} = 0$ [118]. To characterize the ensemble of node degrees of a graph we evaluate a graph's degree distribution, which we described next.

Degree distribution

The degree distribution of a graph is a global property which communicates the basic connective topology of the graph. Certain characteristic distributions imply structural properties and specific connectivity patterns in the graph. For instance a random graph, one which is constructed progressively by joining nodes by an edge with a certain probability, will have a binomial degree distribution [118, 119] most nodes will have a degree close to the mean of the distribution and few nodes will have very low or very high degree.

Another characteristic degree distribution is the power law degree distribution where node degrees approximately follow:

$$P(d) = 1/d^k \tag{1.3}$$

where d is the degree of a node, $P(d)$ is the frequency of that degree in the graph, and k is a constant that defines the scale of the power law distribution [24, 88, 118]. In this case, the frequency of a node having a certain degree is inversely proportional to that degree by the factor k. Therefore power law graphs tend to have a limited number of highly connected nodes, often called hubs. The

proportion of high and low degree nodes is dependent on the value of k and implies certain structural properties [24, 88, 118]. When k is small ($k < 2$), there are very few hubs that the rest of the graph depends on to remain connected. When k is large ($k > 3$), there are many high degree nodes and the graph structure is close to that of a random graph [24, 88, 118]. When $2 < k < 3$ there tends to be a hierarchical connectivity where the hubs are connected to medium degree nodes which are connected to low degree nodes [24, 88, 118]. From this connective structure, and the characterization of hubs from their connectivity patterns, the functional role of hubs can be interpreted. For example, the *C. elegans* connectome has a power law degree distribution and most of its hubs are neurons with a particular cell type called interneurons whose role is to connect more specialized cell types, sensory and motor neurons [157]. This example demonstrates how the connective role of each node can be established through the characterization of the degree distribution of a graph.

Triangles, cliques, and clustering coefficients

Many graph theory measures have been developed to evaluate the connective structure of nodes on a local scale. Nodes can be connected in triangles: a set of three nodes all connected to each other. The transitivity of a graph is accordingly the proportion of realized to unrealized triangles [118]. Higher order structures of fully connected nodes are called cliques: a k-clique is one where k nodes have all possible edges between them realized. Many measures stem from these types of structures such as the number of triangles, the number of k-cliques in the graph, the size of the largest clique in the graph, etc.

In order to measure the local connective behaviour on an individual node basis, the clustering coefficient of a node is calculated as follows:

$$c_i = \frac{\sum_i^N \sum_j^N e_{jk} * e_{ij} * ejk}{d_i(d_i - 1)/2} \tag{1.4}$$

where the numerator of the fraction is the number of triangles through node i with $e_{jk} * e_{ij} * ejk = 1$ if the nodes i, j and k are all connected [118]. The denominator represents the total number of possible triangles connected to node i given that it has a degree of d_i. The clustering coefficient expresses the connec-

tivity between neighbours: if all of its neighbours are connected than a node has a clustering coefficient of 1.

The clustering coefficient of an edge can also be measured by evaluating the number of overlap in neighbours of the two nodes connected by the edge in question. An edge's clustering coefficient is calculated using:

$$c_{i,j} = \frac{|N_i \cap N_j| + 1}{\min(d_i, d_j)} \tag{1.5}$$

where N_i and N_j are neighbours of nodes i and j respectively [103].

The clustering coefficient of a node and its degree are independent properties though the higher the degree of a node the more connected its neighbours need to be in order to also have a high clustering coefficient. The global clustering coefficient is calculated as the average of the nodes' clustering coefficients. Along with notions of degrees, triangles and cliques, these graph theory measures characterize the global connectivity of a graph and the local connectivity of each node.

Centrality measures

Centrality measures are used to evaluate the position of a node within the graph to determine its centrality with respect to the other nodes in the graph [62, 118, 127, 176]. There are many different centrality measures, each of which uses different metrics to evaluate the topological position [19, 62, 123] of a node as illustrated in Figure 1.2.

Each measure estimates the centrality of the position of a node over a certain range. For instance, degree centrality evaluates the position of a node locally by simply taking into account it's degree. Betweenness centrality evaluates the position of a node globally by evaluating the importance of that node relative to all paths in the graph, as expressed by the following equation:

$$bc_i = \sum_{j,k,j\neq k}^{N} \frac{p_{jk}(i)}{p_{jk}} \tag{1.6}$$

where $p_{jk}(i)$ is the number of paths between node j and k that go through i while p_{jk} is the total number of paths going through j and k [118].

The centrality of a node may reflect its importance in maintaining the overall

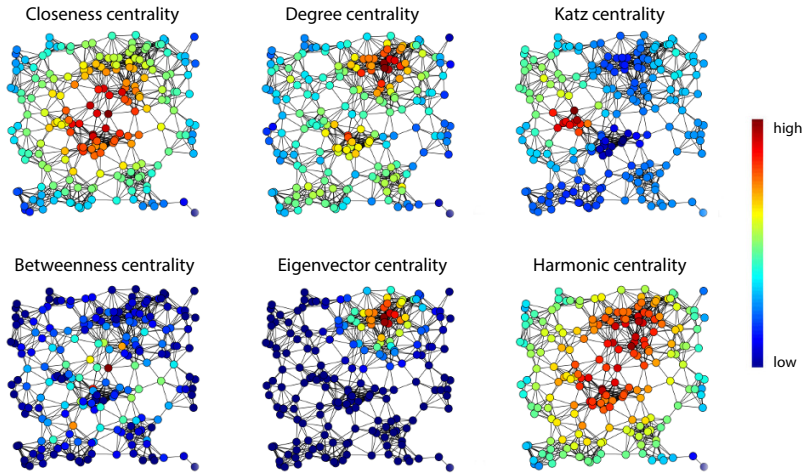

Figure 1.2: An overview of different node centrality measures. Nodes are coloured by their relative centrality (adapted from ©Tapiocozzo (2015) under CC-SA license).

structure of a network [5, 85, 118]. The centrality measure that is most appropriate to find the structurally important or essential nodes in a network depends on the range of centrality appropriate for the system under study [5, 85]. For example in a communications grid network where signals are sent between computers, it is imperative that certain computers don't go down (as in during a power shortage) such that all messages can make it to their destination: computers with high betweenness centrality tend to connect others which would be disconnected in their absence and thus if shut down the message will have no alternate route to follow [5]. Furthermore, the structure of the network can also influence the applicability of a centrality measure. For instance, in a network with a very high global clustering coefficient and thus where most nodes are connected to their neighbours, the nodes' betweenness centrality values would be somewhat evenly distributed and probably not as useful to discern low to high centrality nodes than another measure such as degree centrality. Therefore picking the appropriate centrality measure for a system depends on the structure of the network and the roles played by central nodes in the system.

Modules

The graph theory measures so far presented have focused on global and local topological properties in a graph. Modularity analysis evaluates the sub-global topology of a graph by finding structurally meaningful subgraphs (i.e.. subsets of the graph) [120]. A module is defined as a subgraph whose connectivity pattern between its members is greater than the connectivity patterns with nodes outside that subgraph [103, 120]. Modularity analysis can thus be interpreted as a form of topological clustering on the graph [120]. One famous example of the application of modularity analysis is the Karate Club network [120, 176]: subsequently to the modelling the two modules in the social network, the karate club split and its members separated to form two karate clubs following the modules predicted [173] (see Figure 1.3).

Determining the optimal partitions in the graph to find modules depends on the type of connectivity patterns evaluated on the subgraph [120, 176]. As in the case of centrality measures, what is considered an appropriate connectivity pattern depends on the context of the system being modelled. Several modularity algorithms have been developed and each assesses different connectivity patterns. The type of pattern used influences the interpretation of the modules and the method used to find the patterns determines the computational complexity of a modularity algorithm.

Most modularity algorithms rely on measuring node properties such as degree or clustering coefficient to evaluate the connectivity of subgraphs. Others measure larger structures such as triangles and cliques. In summary, modularity algorithms assess the connectivity of subgraphs by measuring one or a combination of the following [103, 120, 176]

1. cliques or clusters of cliques
2. minimum edge cut of a graph
3. node density in a subgraph
4. high betweenness centrality nodes between subgraphs
5. total degree within a subgraph

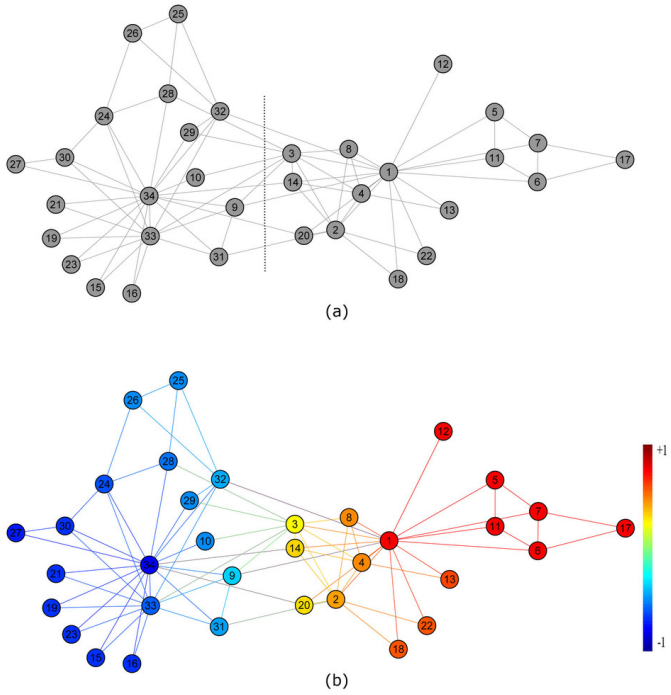

Figure 1.3: The modularity of the Karate Club network. Nodes are coloured by their association to the blue and red modules. The two modules correspond to the actual division of the club into two separate karate clubs (from ©Zhao et al. (2014))

As in clustering methods, modularity algorithms can have a top-down approach in which case graphs are partitioned iteratively or a bottom-up approach where subgraphs are merged iteratively. One advantage of the bottom-up approach is that it doesn't rely on knowing how many modules there might be in the graph [120, 176]. Some modularity algorithms are NP-hard, others have a complexity as high as $\Omega(E^2N)$ and thus many are not applicable to large networks with over tens of thousands of nodes [120, 176]. One low complexity algorithm with a bottom-up approach which has been evaluated on protein-protein interaction (PPI) networks is a fast agglomerative algorithm called FAG-EC [103]. This algorithm measures the modularity of a subgraph by comparing the in-degree of a subgraph to its out-degree [103] where the in-degree of corresponds to the number of edges between nodes within the subgraph and the out-degree corresponds to the number of edges

with nodes outside the subgraph. If the in-degree d^{in} is greater than the out-degree d^{out} of a subgraph S by a multiplicative factor λ, than the subgraph is a module:

$$\sum_{i \in S} d_i^{in} > \lambda \sum_{i \in S} d_i^{out} \tag{1.7}$$

where the value of the parameter λ can be adjusted to obtain a stricter definition of a module [103].

The algorithm builds and evaluates the modularity subgraphs by starting with singleton subgraphs (i.e. each node is its own subgraph) and merging subgraphs when their ratio of in- and out-degree increases. So as to reduce the complexity of the algorithm, the order in which subgraphs are evaluated as merge-able relies on the strength of the clustering coefficient of an edge (see Equation 1.5) between two nodes, one in each subgraph: the higher the clustering coefficient of an edge the higher the probability that the two nodes connected by that edge will be in a module [103]. Thus the edges with the highest clustering coefficient are used to merge subgraphs earlier in the algorithm ensuring FAG-EC a complexity of $\Omega(cE)$ where c is a constant, which is relatively low compared to other algorithms [120, 176].

FAG-EC was tested on PPI networks to find groups of proteins that perform specific biological functions in a cell through these network modules [103]. This method and other modularity algorithms have also been used to find functional modules in other systems such as trophic networks [24, 88, 122], and social networks including the Karate Club network described above [162] (see Figure 1.3).

1.1.3 Biological network complexity

Complex biological systems are teeming with interactions at different scales: from molecules to cells to tissues to organs to organisms to species to environment to ecosystems. Networks have been used to model these complex interactions in biological systems at every level (Figure 1.4). For instance, at the molecular level, PPI networks are constructed to model the relationships between protein-protein interactions; at the ecosystem level, foodwebs are built to model trophic interactions between species. As with other types of systems, biological networks are far from random and have topological features that have informed researchers on how they function and their dynamics. For example, PPI networks from different organisms

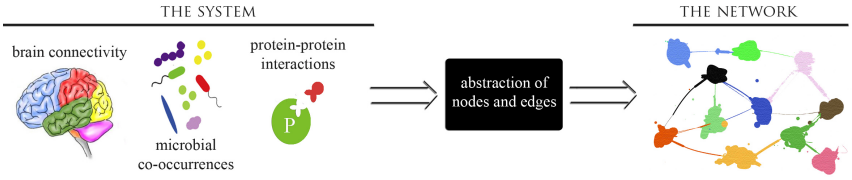

Figure 1.4: Complex biological system modelling through networks: a variety of systems can be described using networks by abstracting system agents and relationships between agents.

have been found to have modules that match cellular and metabolic functional units within that organism [24, 88]. Here we present a few recurring and non-random global and local structural patterns in biological networks followed by the different challenges faced when visualizing and interpreting them.

Common structures

As a first assessment of the a network model, the topology of biological networks is tested against the topology of random networks. For a biological network with N nodes and E edges, a random network can be built with the same number of nodes and edges [35, 118]. The structure of the two networks can be compared by evaluating their degree distribution, average shortest path length, diameter, modularity, global clustering coefficient, assortativity, etc [118].

The degree distribution of most biological networks, including protein-protein networks, gene expression networks, and foodwebs, is typically a power law distribution [24, 88]. As described in Section 1.1.2, a power law distribution implies several specific topological properties, two of which are discussed here in the context of biological networks. First, networks with a power law distribution have few high degree nodes commonly called hubs. In PPI networks, different methodologies have been used to characterize the essential proteins, proteins that are important to the proper function of the system, and have found that these proteins are often hubs in the network [24, 88]. While these well-connected nodes can play differential roles in biological networks, they consistently display properties that distinguish them from other nodes in the network. Second, power law networks with a scaling factor $2 < k < 3$ are called scale-free networks that manifest a hierarchical

connective structure: a high degree node is typically connected to medium degree nodes which are connected to low degree nodes [24, 88] (Figure 1.1B). foodwebs from different habitats and of different species richness, from tens to thousands of species, have a scale free network structure which renders them robust to disturbances (e.g., change in climate) such as the extinction of species (i.e. the removal of nodes in the network) [45, 92, 122]. Given such a low proportion of high degree species in the foodweb, removing a single species will rarely result in network collapse. The topology of foodwebs confers an adaptive connected structure that is robust to perturbation, a property of scale-free networks that is conserved across systems [118].

Complexity and systems

Biological systems are dynamically interconnected within and between hierarchical levels. This complexity makes them difficult to understand even with the use of graph theory to construct networks. Consider that biological data often faces accuracy and reproducibility issues which limit the power of the models used to study them. For instance, brain imaging technology has enabled the measurement of activity of regions of the brain at a macroscopic level and this data is used to reconstruct brain networks called connectomes despite the fact that the activity in the brain occurs at the level of individual neurons [25]. Therefore the construction and interpretation of the connectome is dependent on and limited by the resolution of the data and the fact that connection patterns typically vary between individuals [25].

Beyond data challenges associated with replication and resolution, biological networks containing tens to tens of thousands of nodes and edges are difficult to navigate. In particular, most network visualization schemes are inappropriate for very large networks as we will see in Section 1.2.2. Moreover, because of the great number of nodes and edges, characterizing local structures in the network such as cliques, modules, and triangles, as well as manually evaluating the construction of the network on an individual node basis is not feasible. One alternative is to find and evaluate local patterns such as repeating connective structures or motifs. Although algorithms exist to identify motifs, these must be specified a priori limiting

the discovery of new or unexpected patterns.

In addition to being hierarchical, biological systems are multivariate and thus highly dimensional with respect to intrinsic and extrinsic factors. Each factor and its influence can be encoded in the nodes and edges of the biological networks as quantitative and qualitative properties. For example, individual species in foodwebs have diets, population sizes, seasonal habits, etc. which impact their feeding behavior and are tightly linked to their role in the foodweb. At the same time, environmental parameters such as weather and geography also influence trophic interactions [122]. Taking all of these dimensions into account in building a complete model of the system leads to a multitude of possible association patterns. In order to meaningfully model these patterns, a network must integrate the dimensionality of biological systems. Accordingly, both quantitative graph theory measures and qualitative methods such as visualization need to accommodate for this complexity in their design and implementation.

1.2 Network exploration

Network analysis typically involves quantitative measures and methods as well as visualization. However, most network visualizations such as force directed layouts are not suitable for visualization tasks when scaled up to large networks because they are inconsistent and difficult to interpret [96, 115], often resembling "hairballs" [96] (Figure 1.7). Biological networks in particular are very large networks. Appropriately visualizing these networks could assist in the interpretation of network properties relevant to system function. Here we motivate visualization as a means for system exploration and present current network visualization procedures.

1.2.1 Visualization as a means for data exploration

Humans are visual creatures with powerful pattern detection capabilities [50, 125, 147, 151]. While some tasks are best accomplished by computational techniques, others are difficult to abstract quantitatively and are well suited for visualization purposes [115]. A visualization tool can thus become medium to explore a dataset and create a transition between raw data and the formulation of a problem or hy-

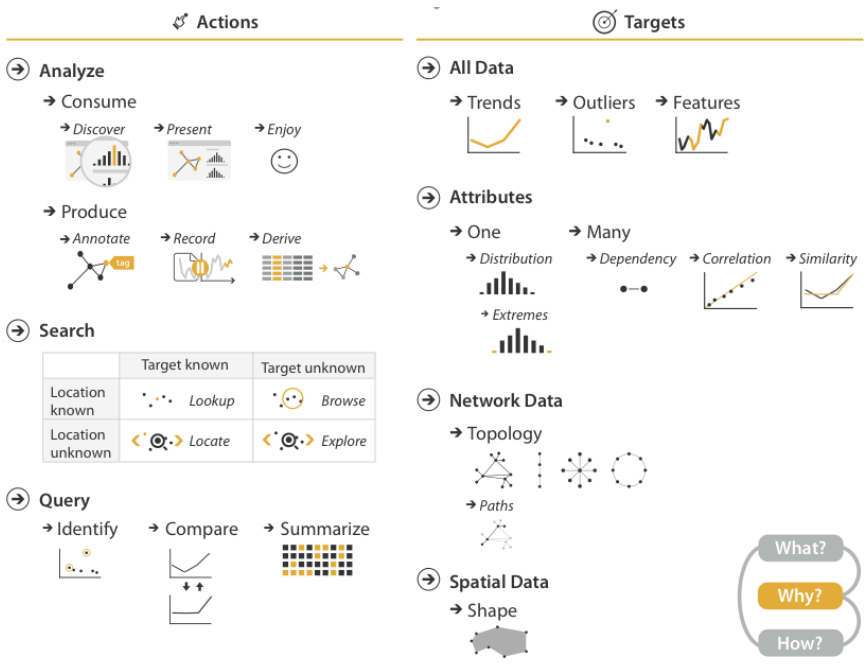

Figure 1.5: Overview of tasks accomplished by visualizations. These tasks are composed of visual actions applied to data targets (from ©Munzner (2014)).

pothesis [115, 125]. Visualizations can be designed for different types of data, and different tasks (Figure 1.5). For example, explorative visualizations have been built to identify outliers, global and local patterns, similarities, and in general novel features of the visualized data [115].

Interactivity can further enable a user's exploration experience of a visualization compared to a static graphic [50, 115, 147]. Shneiderman proposed a visualization mantra to optimize a user's interactive experience: "Overview first, zoom and filter, then details-on-demand" [147]. In sum, a visualization is designed to optimally display the different dimension of dataset in a way that is easily navigable by a user in the context of an interactive and explorative visualization. Furthermore, Coleman proposed a set of requirements for a design to produce an aesthetic and comprehensive visualization [37]

Generality in its application to different datasets

Flexibility in its the range of tasks that can be accomplished

Transparency in its layout to ensure its interpretability

Competence in the number and quality of the features it reveals

Speed in its rendering time

Particularly in the case of biological systems which vary in size, complexity and number of dimensions, explorative visualizations are invaluable tools to the discovery of interpretable biological patterns and the development of hypotheses which can drive subsequent experimental and computational analysis [115].

1.2.2 Current network visualizations

Though a multitude of heuristic and rule-based network visualization methods exist, here we present three different methods to illustrate the challenges in visualizing and exploring networks. The pitfalls of these visualization schemes demonstrate that current network visualizations are not designed to facilitate the exploration and discovery of novel global and local patterns in complex systems [96, 115].

Adjacency matrices

Adjacency matrices are both the linear algebraic formulation and a visual representation of graphs. Nodes are encoded as row and column labels while edges are encoded as entries in the matrix. This representation is suitable for the visualization of directed networks in which case the matrix is asymmetric, of weighted networks in which case entries encode the weight of an edge, and of connected components in which case the matrix is sparse.

This visual representation is particularly useful for showing the components, modules and cliques of a network by applying an appropriate node ordering (Figure 1.6) [115]. However it does not scale to large networks and is not suitable for looking at individual node topologies such as clustering coefficient.

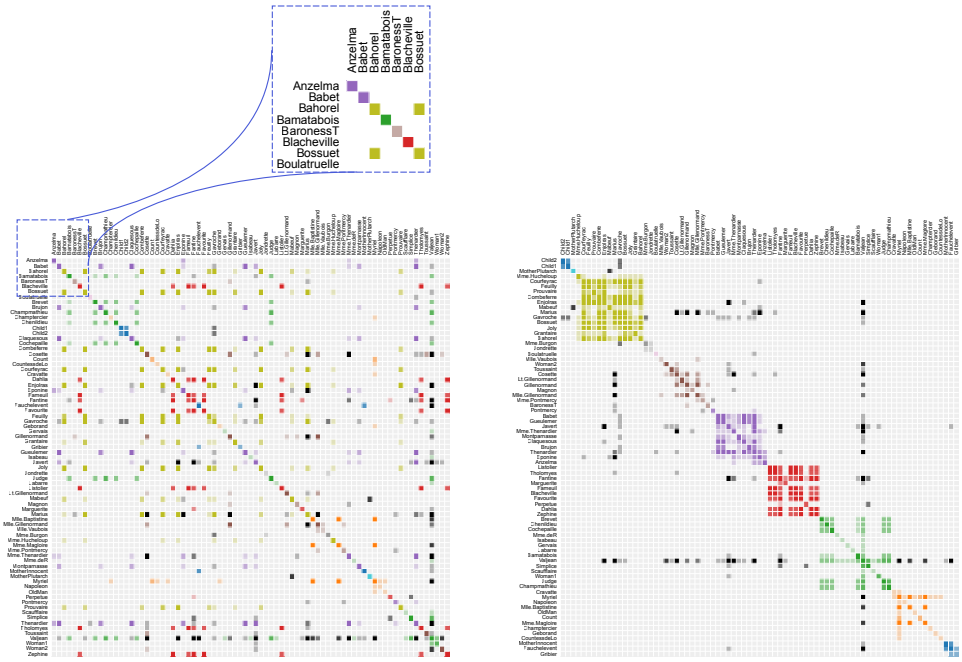

Figure 1.6: Adjacency matrices, a tabular representation of graphs. Rows and columns are labeled by nodes and each entry in the matrix corresponds to the presence or absence of an edge between the corresponding nodes. The two adjacency matrices, one ordered alphabetically, one ordered by clustering node connectivity patterns, lays out the social network of the characters in Victor Hugo's *Les Miserables* (from ©Bostock (2015)).

Force-directed layouts

Force-direct layouts encode nodes as circles and edges as links between them. The layouts are obtained by applying physical rules to the nodes and edges to place them on a plane in a way that minimizes overlap and the number of crossing edges [64]. For example spring-embedded layouts model the edges as springs with different spring coefficients relating to edge weights in the case of weighted networks [64]. The nodes are modelled as particles with repulsive forces to avoid overlap [64].

Figure 1.7 illustrates force directed layouts of a large network. While these layouts are suitable for showing modules, cliques and triangles in small networks, the inconsistencies due to their heuristic algorithms produces network layouts which

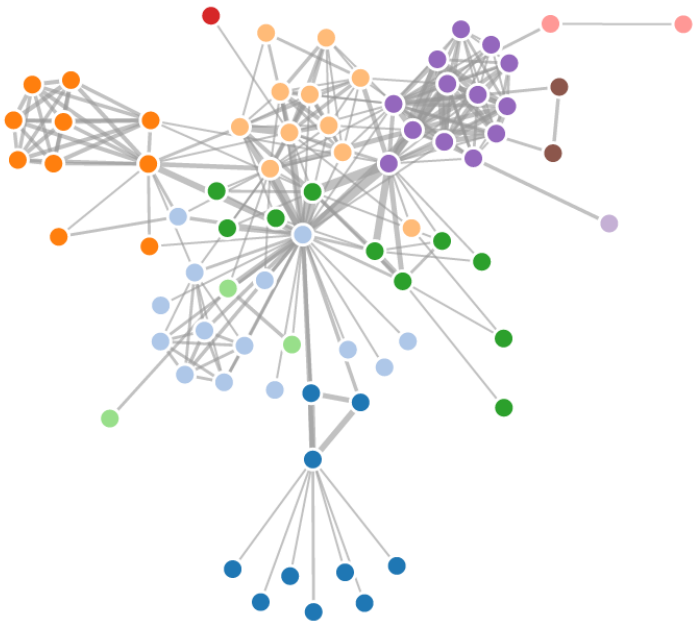

Figure 1.7: Force directed layouts, an intuitive and planar visual representation of graphs. This force directed layout shows the social network of the characters in Victor Hugo's *Les Miserables* coloured by cluster (from ©Bostock (2015)).

are inconsistent and thus do not allow for the comparison of networks.

Hive plots

Hive plots are a consistent and coherent rule-based layout and are an appropriate visualization for comparing and visualizing structural patterns in large networks across different data dimensions. They provide an interpretable visualization while leaving several visualization channels such as colour, size and rule choice, to encode additional data properties [96]. Hive plot's design scales to large networks as it handles visual occlusion and other potential visualization design pitfalls [115] by organizing the layout of nodes and edges given their attributes and network properties [96]. Despite their flexibility, hive plots can be a daunting visualization technique given the large range of options and combinations of layout rules that

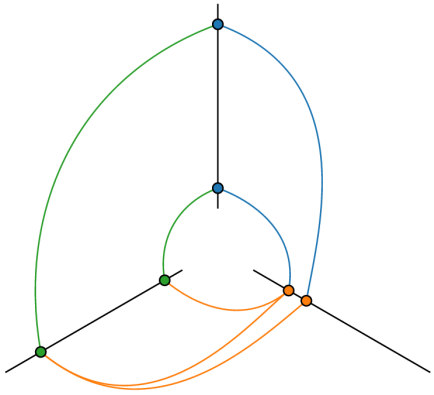

Figure 1.8: Hive plots, a circularly organized representation of graphs. This simple hive plot has 6 nodes arranged on three axes (from ©Bostock (2015)).

must be chosen by the user. In addition, they are not suitable for exploring certain network topological features such as connected components and cliques.

In summary, each network visualization has its strength and weaknesses. Depending on the topology of networks being studied and their size, different network layouts will be more suitable than others. Overall, these strategies are best used in combination to evaluate and explore networks and the systems they model.

1.3 Charting microbial community structure and function

As alluded to earlier, charting microbial community metabolism is extremely challenging because of the cultivation gap between indigenous microorganisms and laboratory settings [114]. There are several reasons for this cultivation gap, including the inability to reproduce in situ physical, chemical and ecological conditions in a laboratory setting. Thus, culture-based techniques capture only a small fraction of microbial diversity. Plurality sequencing bridges the cultivation gap by providing direct access to the genetic material of indigenous microorganisms. The genetic material stored in nucleic acids (DNA and RNA) contains the necessary information for an organism to grow and reproduce [91]. A genome contains the genes necessary to encode the organism's metabolic functions and reproduction.

Accordingly, the collection of genetic material in an environment defines the ensemble of the community's genomes: the metagenome [29]. Metagenomic studies resolve the taxonomic diversity and functional activity of a community by decoding this genetic information.

Next-generation sequencing techniques have been developed to measure the genetic material in environmental samples in a high-throughput manner. Given the seemingly infinite diversity of microbial life and the great variation in genetic encoding, plurality sequencing studies have generated a great abundance of microbial community data from a variety of natural and engineered ecosystems [29]. This surge of information has driven the development of different solutions to store, manage, analyze and present this "big data" [160]. For instance, publicly available databases permit the annotation of nucleotide and amino acid sequences to known genes and gene products as well as the assessment of their phylogeny [22, 29, 32, 67, 137, 144, 160]. Software solutions are used to align, cluster, and manipulate sequences to provide analytical frameworks to characterize the taxonomic, genetic and metabolic potential of a microbial community [27, 29, 75, 83, 93, 160]. In addition, visualization techniques have been developed to present and illustrate ecological findings [78, 83, 89, 95, 99]. Finally, as the quantity of environmental sequence information increases, these solutions are required to scale to the task to effectively study the organization of microbial communities and their relationship with their environment.

1.3.1 Soil ecology

Soil harbors the most diverse microbiome [167]. From boreal forests to arctic sediments, one gram of soil containing an estimate of up to tens of thousands of unique species [167]. Assessing the quality and type of soil involves measuring soil properties called edaphic factors, including soil moisture, porosity, temperature, and acidity, which are affected by abiotic and biotic factors such as agricultural practices, climate, plant and fungi growth. However, the distribution of microbes throughout the soil profile is influenced by both edaphic factors and interactions between microbial community members [3, 57, 74, 76, 100, 168] (Figure 1.11). Nevertheless, to date, most research has focused almost exclusively on the role of

edaphic factors by relating this measurable information to community composition data using multivariate methods [3, 57, 74, 76, 100, 168]. For example, studies have revealed soil acidity [100] as well as carbon and nitrogen pools and cycling [39] to be strong indicators of soil microbial community structure and diversity. In addition, decreases in microbial biomass and community diversity with soil depth [108, 113] have both been attributed to concurrent changes in carbon resources and soil acidity throughout the soil profile. Current models of soil communities paint an incomplete picture of this complex system as few studies combine microbial interactions with environmental parameter data [105, 138, 139].

The long term soil productivity project

The LTSP project is a multidisciplinary effort to monitor the impact of forestry practices on North American soil productivity that was initiated by the United States Forest Service 25 years ago [136]. The project spans ten North American ecozones: biogeographic regions manifesting particular temperature ranges, precipitation patterns and tree species. Today, the LTSP study remains one of the world's largest coordinated research networks including over 110 sampling locations in the United States and Canada [135, 136] (Figure 1.9). Research on LTSP sites is primarily focused on impacts of tree harvesting practices related to soil organic matter (OM) removal and soil compaction [76, 136]. Each LTSP site uses a randomized and replicated factorial design with three levels of OM removal (OM1-OM3) in $40x70m^2$ plots (Figure 1.9). A control plot, representing natural reference forest (OM0) is also included at each site. In OM1 plots, tree boles have been removed but tree crowns, felled understory, and forest floor material is retained resulting in minimal soil OM removal [136]. In OM2 plots, aboveground vegetation is removed but forest floor material is retained resulting in intermediate soil OM removal [136]. In OM3 plots, all surface organic matter is removed leaving bare soil exposed [76, 135].

Recent efforts to study microbial community responses to soil perturbation in the LTSP network has resulted in an archive of samples spanning 5 ecozones [76]. Extant microbial studies have focused on community composition from different soil types, at different depths and different levels of soil organic removal and com-

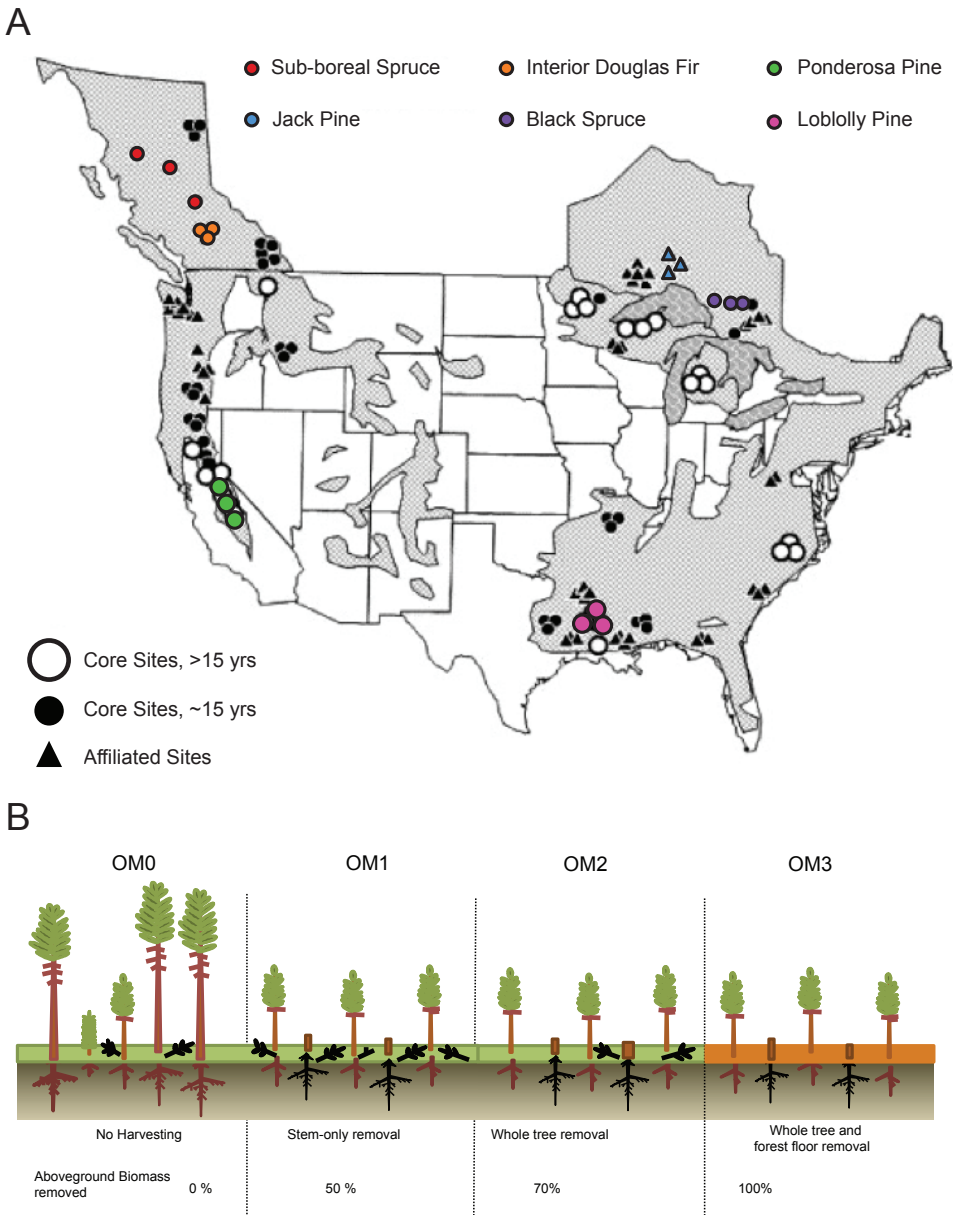

Figure 1.9: Overview of the LTSP ecozones and treatments where microbial communities were sampled. A) The age and geographic location of sites per ecozone is marked. B) Different OM treatments were conducted in these forests and harsher treatments reflect increased biomass removal [31].

paction, as described above. These forestry practices perturb the soil's physical, chemical and ecological conditions which cause a disturbance in the microbial community with resulting feedback on biogeochemical cycling [76]. Given that soil microbes recycle the carbon content of the soil and produce climate active gas such as carbon dioxide, methane, and nitrous oxide, this study has the potential to help assess the impact of forestry practices on forest ecosystem health and climate change [97]. With samples from a variety of ecozones, this impact can be measured on a large geographic scale. However, current analytical and visualization methods do not facilitate the interpretation of these large datasets and the microbial communities they represent across ecological scales.

1.3.2 Taxonomic assessment

As previously described, microbial community diversity can be assessed using SSU rRNA gene sequencing [141, 156]. The SSU rRNA gene is a highly conserved gene with sufficient taxonomic resolution. Its hypervariable regions V1-V9 serve as the genetic markers to taxonomically identify each organism while the conserved regions enable primer binding and sequencing [141, 156]. High quality sequences are recovered and can then be clustered at different percent identity thresholds against a curated database, for instance the Green Gene database [22]. Matches between the database entries and clustered sequences create an operational taxonomic unit (OTU) that can be assigned taxonomy to different levels in the taxonomic hierarchy [67].

An identity threshold of 97% clusters SSU rRNA sequences with sequence variability expected between organisms of the same species [18]. Lower thresholds capture higher taxonomic levels such as phylum, class, order, family, and genus [18]. Given the number of clustered sequences for a given OTU, the relative abundance of that OTU in the community can be estimated and used to infer its abundance pattern across the sample profile. Studies of large collections of SSU rRNA samples from varying ecosystems (e.g., soil, water, organisms, atmosphere) facilitate the characterization of the Earth's microbiome [67].

1.3.3 Community composition

Quantitative ecology transformed ecological research from a primarily descriptive science to an analytical science with hypothesis formulation and testing [101]. Models are developed and validated using numerical approaches; methods from fields such as mathematical algebra, statistics, information theory, and chaos theory have been utilized to resolve ecological and spatio-temporal patterns in complex and highly dimensional datasets. By measuring dependencies, similarities, correlations, and other complex relationships in the ecosystem, these quantitative procedures resolve the influence of abiotic (i.e. non-living) factors such as sunlight and biotic factors (i.e. living) such as plant growth in an ecosystem [79]. Many of these methods have been adapted from macroecology to characterize the composition of microbial communities. Here, three ecological quantitative measurements that are generally used in ecological studies and that evaluate different features of community composition are described: community diversity, clustering of environmental samples, and indicator species analysis.

Community diversity

Community ecology often focuses on determining relationships between species diversity and environmental factors. Particularly in macroecology, measuring the diversity over spacial and temporal scales has been used to assess the effect of environmental changes on ecosystems [79, 101]. For example, the BIODEPTH experiment in Europe compared the above-ground plant biomass and plant species count, a measure of diversity called species richness, over environmental and spatial gradients and showed a log-linear increase in biomass with species richness [79]. Similar relations have been found in microbial communities inhabiting soil [76, 108, 113]. Thus, diversity appears to be a quantitative indicator of ecosystem perturbation across macro and micro scales.

As a key measure of community structure, many different diversity metrics have been developed, four of which are summarized in Table 1.1. Diversity quantifies the distribution of species in a collection of samples. As described by Pierre and Louis Legendre, community diversity is “a measure of species composition, in terms of both the number of species and their relative abundances” [101].

Table 1.1: An overview of different of ecological diversity metrics where D is diversity, q is the total number of unique species, i represents the i^{th} species, p_i is the relative abundance of the i^{th} species, f_1 is the number of singletons, and f_2 is the number of doubletons.

Metric name	Formula	Description
species richness	$D = q$	number of unique species
Shannon's entropy	$D = -\sum_{i=1}^{q} p_i \log p_i$	measure of disorder in species distribution
Simpson's index	$D = 1 - \sum_{i=1}^{q} p_i^2$	measures concentration of species
Chao 1	$D = q + \frac{f_1(f_1-1)}{(2f_2+1)}$	skews the species richness by an estimate of the number of unsampled species

It is important to note here that microbial diversity is an operational and probabilistic measure due to the polyphasic nature of the definition of species. Namely, microorganisms of the same species can differ phenotypically and genetically [149] and this differentiation can in return blur the distinction between species. In practice the species concept can be defined given a measurement of genetic variation to evaluate the genetic composition of a community and consequently capture its ecological diversity [149] . Here the species concept definition is based on percent similarity in SSU rRNA sequences, as described above.

Species richness measures the count of unique species in a sample collection. This measure is highly affected by rare species and sampling depth (i.e. the number of sample units collected); however this issue is resolved using the rarefaction method which calculates the number of species given constant sample unit size [101]. Richness measures and rarefaction curves are used to quantitatively evaluate the recovery of the diversity of an environment through sampling [101].

Shannon's entropy measures how evenly species are distributed by taking into account their relative abundance: high values of this measure correspond to most species having similar abundances and low values typically correspond to a few species dominating the sample units [101].

Simpson's index corresponds to the sum of probabilities that two randomly

chosen organisms belong to the same species; the lower this probability, the higher the overall value of diversity becomes [101]. This measure is highly affected by changes in rare species and is relatively stable with increasing sample unit sizes [101].

The Chao 1 diversity measure differs from the others in that it takes into account f_1, the number of singletons (species only found once), and f_2, doubletons (species only found twice). In the context of measuring microbial diversity with SSU rRNA sequences, singletons are OTUs for which only one sequencing read is recovered. This measure is based on the idea that the rare species can tell us about how many species environmental sampling may have missed, and the added factor to q, the species richness, helps estimate this contribution to the diversity [101]. In the case of microbial species diversity measurements, next generation sequencing technologies can capture erroneous sequences and therefore most SSU rRNA sequencing studies do not include singletons in their analysis.

Given that ecological research is often focused on the spatial organization of an ecosystem, diversity measures are applied to partitions of the sample collection to assess the distributions of species through spatial components. Whittaker described three spatial levels of diversity: alpha, beta and gamma diversity. Alpha diversity (α) represents the diversity at each sample site, gamma diversity (γ) represents the diversity of the whole sample collection, and beta diversity (β) measures the per sample variation in diversity [101]. The three diversity levels are related through the relation $\beta = \gamma/\alpha$ [101]. Different metric, such as the ones in Table 1.1 can be used to calculate the α, β and γ diversities.

To conclude, these metrics quantify diversity based on different ways of assessing abundance and distributions of species and can be used to evaluate spatial and temporal variations within and between sample units.

Sample clustering

Environmental samples can be grouped into clusters from their respective species composition. Samples with similar compositions based on similarity metrics are grouped into the same cluster based on an operational threshold. These clusters are characterized to resolve ecological patterns, such as species niches [2, 44, 59],

groups of species that co-occur across geographically distinct sampling sites, and in general to assess the similarities within and between sample units. Clustering analysis is thus a knowledge discovery method.

There are many ways to measure the similarity between samples and many ways to group samples into clusters. Similarity metrics include the Euclidean distance and the Manhattan distance [101]. Each distance will weigh abundant and rare species composition differently. Here we provide an overview of hierarchical clustering which is commonly used in ecological studies and can be used with any distance metric. There are two types of hierarchical clustering procedures. Both take as input a distance matrix: the distance metric chosen is used to assess the similarity between all pairs of samples thus building a sample distance matrix. Agglomerative clustering iteratively groups pairs of samples to form clusters and then forms clusters between the initial clusters, and so on [101]. Divisive clustering is the equivalent "top down" approach: the ensemble of the samples is partitioned into clusters which are subsequently divided iteratively [101]. Both procedures rely on a greedy algorithm [101]: they pick the best way to merge or divide clusters in order to maximize the similarity within or distance between new clusters, respectively. The hierarchy of clusters produced by both methods can be different and is often visualized using a dendrogram.

Ecological analysis of clusters includes the evaluation of common environmental factors such as site locations. This idea is based on the fact that related samples through common environmental factors such as climate will have similar species compositions and will cluster into the same clusters [101]. Furthermore, hierarchical clustering can be used to assess the quality of an environmental sampling experiment. Finally, hierarchical clustering is one method used to evaluate sample compositionality which is can be used to expand the ecological understanding of an environment.

Indicator species analysis

The composition of a community can be characterized on a species level by evaluating the ecological relevance of species compared to some environmental factor. One simple example is a particular species which is consistently found in samples

of a particular habitat. Significant associations between species abundance profiles and environmental factors can help assign ecological meaning to samples and sample sites [101].

Different methods evaluate the association between species abundance and environmental factor distributions. One way to measure these associations is to use correlation indices. Methods differ by how they handle the variance and distribution of species and samples [101]. Here we provide an overview of an ecologically motivated method called indicator species analysis.

Once a partition of the samples is established, using a predetermined ecological factor or a clustering method, indicator species analysis can be applied to discover which species are indicative of the "condition" of sample partitions. The indicative value of a species for each condition is calculated using:

$$IndicatorValue_{ij} = A_{ij} * B_{ij} \tag{1.8}$$

where A_{ij} is the specificity of species i to the cluster j and B_{ij} is its fidelity [101]. The specificity is calculated by dividing the average abundance of species i in the samples belonging to cluster j by its average abundance in all samples for all clusters k, as follows:

$$A_{ij} = p_{ij}/p_{ik} \tag{1.9}$$

High specificity is obtained when a species is highly abundant in all samples of cluster j and rare in other samples. The fidelity is calculated by dividing the number of samples in cluster j where species i is found divided by the total number of samples in cluster j, as follows:

$$B_{ij} = samples_{ij}/samples_j \tag{1.10}$$

High fidelity is obtained when a species is present in all samples within a cluster. Figure 1.10 illustrates how a species can have high fidelity, a high specificity or both.

Once indicator values are measured for all combinations of species and sample clusters, the significance of the results are evaluated by taking into the composi-

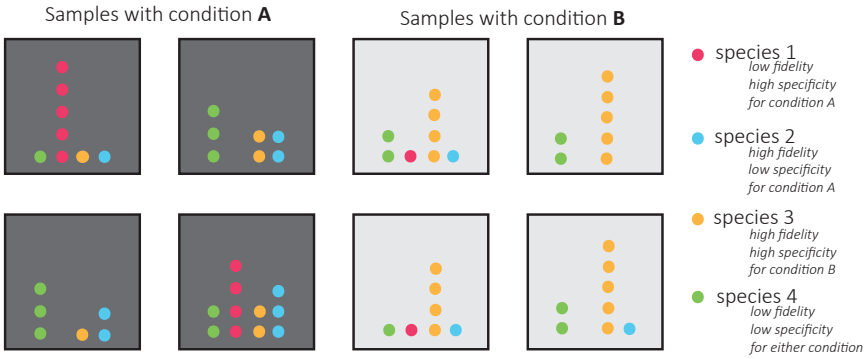

Figure 1.10: An illustration of the specificity and fidelity of species to environmental conditions. The eight samples are partitioned under two different conditions.

tional bias in species abundance profiles. One example of compositional bias is the presence of a very abundant species which lowers and skews the relative abundance of other species when they occur in the same samples [101]. The significance is evaluated for each species i by permuting sample counts of other species and recalculating the indicator value of species i. This permutation procedures obtains a distribution of indicator values for each species and by comparing the actual value found with this distribution, a p-value is obtained which indicates the probability that this value occurred by chance. Typically p-value thresholds of 0.05% significance and lower are used to filter out poor values [101].

Finally, for each cluster a number of indicator species is obtained that can be used in environmental surveys of the condition implied by the sample cluster. In particular, hierarchical clustering can be applied to find the sample clusters before applying indicator species analysis. In this case, indicator species and expert knowledge can help assess the ecological properties of the clusters. However, an important consequence of this procedure is the fact that more indicator species will be found than expected by chance since the sample partitioning was conducted using the same species abundance data [101]. The lack of independence between the two methods implies the need for a thorough examination of p-values before ecological interpretation.

1.4 Microbial co-occurrence networks

Microbial ecologists adopt macroecology quantitative and qualitative methods to analyze, visualize and investigate interaction patterns in microbial communities by constructing microbial community networks [11, 54, 54, 105]. modelling the community as an inter-connected system can give insight into the community's functional characteristics related to, for instance, the biogeochemical processes it performs [105, 150]. Structural properties of microbial community networks have been visualized and characterized to infer different biological attributes of the community such as its resilience to disturbance [5, 11, 47]. However, the interpretation of global and local network properties from an ecological standpoint, as is done in macroecology particularly with foodwebs [48, 49, 52, 122], remains difficult [54, 134]. Inferring and interpreting microbial community networks faces many challenges some of which are common to all biological network models and others which are specific to the microcosmos, such as:

1. the selection of a procedure among a multitude of methodologies used to resolves between taxa [11, 54]
2. the statistical obstacles in assessing the significance of inferred interactions [16]
3. the lack of standard procedure to analyze the constructed network [16, 54]
4. the difficulty in ecologically interpreting resolved global network structural properties [24, 54, 88]
5. the difficulty in relating environmental factors to resolved global network properties [54]
6. the difficulty in ecologically motivating and validating the analysis of local network patterns [18, 54]

In this section, we motivate the construction of microbial co-occurrence networks, describe the possible procedures and pitfalls in their construction and provide an overview of current microbial co-occurrence network studies and their findings.

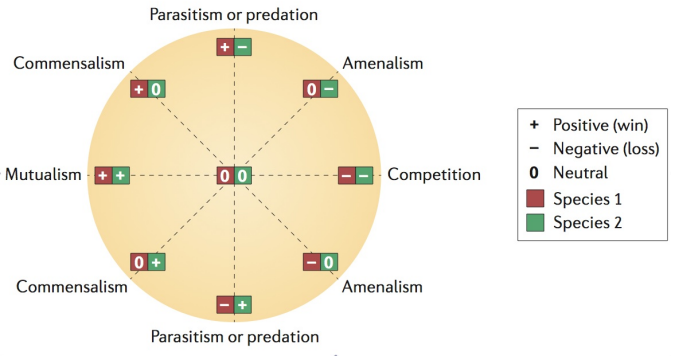

Figure 1.11: Overview of different ecological interactions between microbial community members. Pairwise interactions can have a positive, negative or neutral impact on the two participant (from ©Faust and Raes (2012)).

f

1.4.1 Symbiosis and inter-taxa interactions

Only recent efforts have begun to investigate community structure through the characterization of inter-taxa interactions. These interactions have primarily been resolved using co-culture studies where the stability of an artificial community is tested against the addition and removal of species [113]. Major findings of these efforts include the mutualistic interactions between microorganisms in which a metabolic factor produced by one microorganism is utilized by a second and who then performs a reciprocal service [54, 73, 113, 174]. The different types of interactions, as defined by how they benefit or impair the participating microorganisms are summarized in Figure 1.11 [54].

These interactions directly influence the composition of a community. Since environmental factors can affect the abundance of individual taxa (and vice versa), changes in an environment can have an impact on these interactions. Therefore a complete model of the system should include interactions between taxa and relationships between taxa and their environment.

1.4.2 Microbial network inference

In ecology, the choice of methods used to resolve relationships relies on the assumption that community dynamics can be either stochastic or non random [101, 158]. It is important to note here that Hubbell's unified theory of biodiversity proposes that community composition can be explained by random processes affecting the birth, growth and death of taxa [81]. This hypothesis is called "neutral theory" and has been verified in some ecosystems [33, 81] and contradicted in others [101, 104]. Given the evidence of non random and even causal relationships in microbial communities, we focus on measuring these complex relationships even though we acknowledge that random processes and stochasticity in general also plays a role in shaping microbial community structure. Here we use the term *relationship* to designate association patterns between taxa and their environment while the term *interaction* denotes the resolved inter-taxa associations.

These relationships can be quantitatively measured to build networks and model the community structure. Network inference relies on different quantitative measurements to detect pairwise or complex inter-taxa interactions. For instance, correlation measures assess the similarity in the abundance pattern of microbes: positive and negative correlations detect co-occurrence and mutual exclusivity, respectively [11, 16, 54]. Methods such as regression analysis and rule association mining can uncover complex interactions involving 3 or more taxa [54]. Other models have been adapted to model dynamic interactions that can evolve over time [54]. Here we focus on pairwise interactions resolved through correlation measures due to its flexibility in uncovering many different interactions across spatial gradients and because it is a computationally feasible method to apply to large SSU rRNA datasets, unlike methods such as rule association mining [54]. Notably, correlation-based network inference is also used to construct other biological networks such as gene expression networks [69].

Before a network can be inferred and interpreted from resolved pairwise interactions, many potential pitfalls need to be addressed: normalization bias, similarity measure bias, and multiple testing issues [16, 54]. The normalization of abundance patterns on a per sample basis can skew the relative abundance of certain taxa: the presence of a highly abundant taxon in a few samples can cause the relative abun-

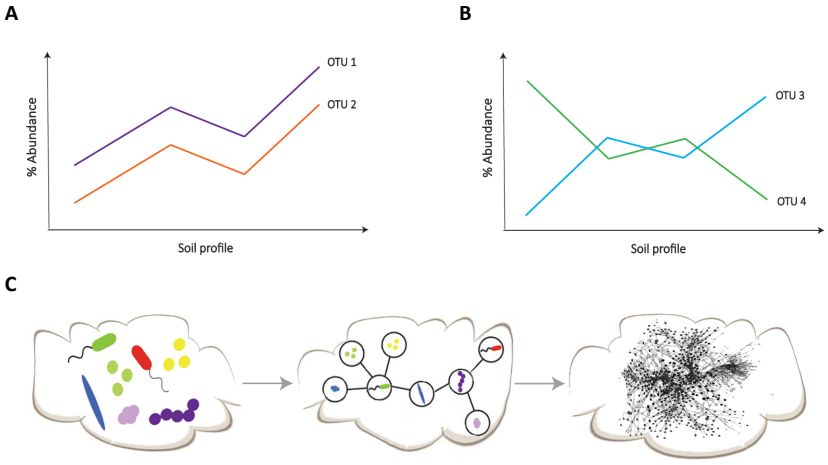

Figure 1.12: An illustration of microbial network inference through co-occurrence patterns. A) OTUs whose relative abundance are similar throughout the sample profile are said to co-occur. B) OTUs whose relative abundance profile are inversely correlated are said to be mutually exclusive. C) From a dataset of community composition, a network can be abstracted by defining OTUs as nodes and co-occurrence and mutual exclusion interactions as edges [54].

dance of other taxa to be significantly lowered in those samples. In order to avoid this compositionality bias and to identify correlations which may have been evaluated as significant because of skewed abundances, a permutation procedure is conducted [16, 53, 54]. This procedure helps remove spurious correlations based on the assumption that permuting the abundance of other taxa and thus varying their relative abundance shouldn't affect the predictive power of significant pairwise correlations.

Similarity measure bias is caused by the use of particular correlation measures which may resolve only specific kinds of interactions. For instance, some nonlinear associations can be detected by Spearman's correlation coefficient, a rank-based correlation, but not Pearson's [148]. In order to maximize the variety of pairwise interactions measured, we can detect correlations using several distance and correlation measures and by combining their results, as is done in the co-occurrence network inference software CoNet [53].

Multiple testing bias is due to the fact that the probability of finding spurious interactions increases with the number of tests and becomes prominent for a very large number of tests [4]. Multiple testing correction controls the number of false-positive interactions and produces a p-value for each interactions whose significance under the null hypothesis can then be assessed [54]. Finally, significant interactions are collected to build microbial community co-occurrence networks where nodes are OTUs and edges represent co-occurrences or mutual exclusions.

1.4.3 Validating network inference models

Though experimental validation is extremely challenging for uncultiaved microbes [29], simulation experiments have been used to both validate network inference methods as well as propose a sampling procedure to produce datasets appropriate for co-occurrence analysis. Berry and Widder simulated microbial communities using generalized Lotka-Volterra equations, calculated the resulting community composition, and inferred microbial co-occurrence networks from the produced abundances patterns [16]. By varying both experimental and ecological parameters and measuring the specificity and sensitivity between the simulated communities and inferred network models they evaluate the conditions under which network inference is an appropriate model to study inter-taxa interactions [16]. Figure 1.13 illustrates the variation in the specificity and sensitivity of the networks models when varying the number of samples, the correlation measure used, and the abundance measure used for communities with 100 species [16]. Figure 1.14 demonstrates how ecological parameters can affect the performance of co-occurrence network models. In general, experimental parameters that increase the modelling performance are the use of several samples, a combination of correlation measures, and the use of compositionality bias-corrected relative abundances. In addition, sampling designs can help optimize this performance by ensuring a high species richness and low β diversity.

Finally, these simulations and further theoretical and experimental studies will help provide a standardized sampling procedure and analytical procedure for building accurate community networks that can model known and capture novel inter-taxa interactions.

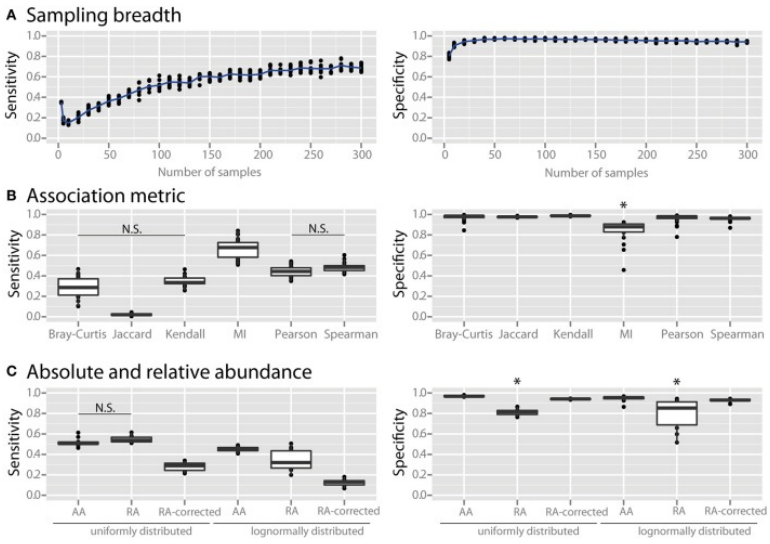

Figure 1.13: The effect of experimental parameters on co-occurrence network modelling performance on simulated communities measured using the sensitivity and specificity of the networks. The experimental parameters varied were A) the number of samples, B) the correlation measure used (MI = mutual information score), and C) the use of absolute abundance (AA), relative abundance (RA), or sparCC-corrected relative abundance (RA-corrected) data, compared for communities with uniformly- or log-normally-distributed species abundances (from ©Berry and Widder (2014)).

1.4.4 Current applications of microbial co-occurrence networks

Just as ecological quantitative methods model ecosystems and their relation to their environment, graph theory measures have been applied to microbial co-occurrence models to evaluate the community's inter-connected structure and its relation to its environment. In particular, network analysis methods have been adapted from foodwebs studies where interactions between species model trophic relations [89, 105, 138, 150]. The graph theory measures presented in Section 1.1.2 that evaluate global topological structures have been applied to co-occurrence networks however interpreting these findings ecologically remains difficult.

Node-based measures such as centrality measures have been used to identify keystone species: taxa whose presence in the community are essential to its

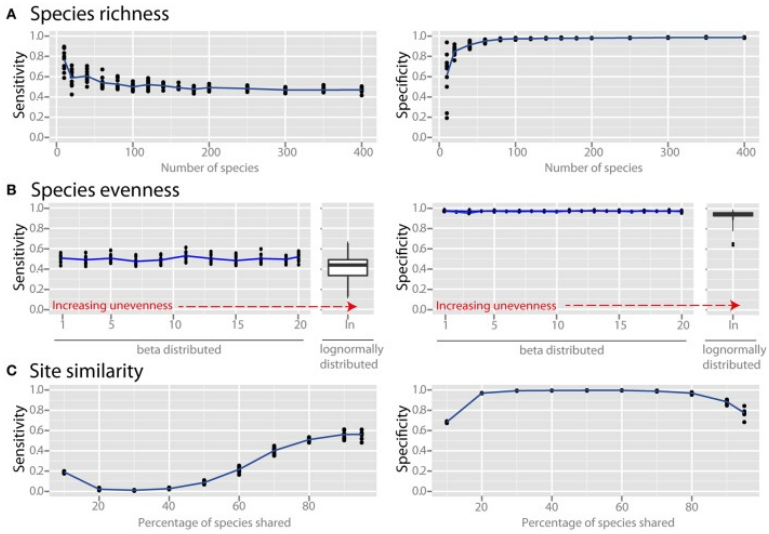

Figure 1.14: The effect of ecological properties on co-occurrence network modelling performance on simulated communities measures using the sensitivity and specificity of the networks. The parameters varied were the A) species richness, B) species evenness defined as Shannon's diversity normalized by the maximum entropy of a community, $\log(q)$, and C) the β diversity of sampled sites (from ©Berry and Widder (2014)).

proper functioning [16, 105, 138]. The current literature which applies methods from macroecology to study keystone microorganisms has not settled on a methodological procedure, in particular which centrality measure to use, to identify keystone species in uncultivated communities through co-occurrence networks analysis. However, network robustness analysis is a method which has shown much promise in finding an appropriate centrality measure to identify OTUs with interesting topological positions in the network that may be keystone OTUs. Network robustness is measured by iteratively removing nodes and evaluating the structure of the network resulting from this removal. The resilience of foodwebs to species extinction is evaluated this way by measuring the number of secondary extinctions from the iterative removal of species in the foodweb [45, 48, 107, 122, 134]. Network robustness simulations have been applied to microbial co-occurrence networks with promising results. For example, network analysis in gut microbiome

data has revealed that healthy subjects have more robust networks than diseased subjects [116]. In another co-occurrence network study, robustness simulations were conducted by removing OTUs using decreasing centrality to attempt to identify key microbial genera in soils from natural forests and agricultural plantations [105, 138, 150].

Though community resilience to species extinction can be motivated ecologically in microbial communities, there is a lack of evidence to support the use of network robustness simulations on co-occurrence networks to test community resilience to environmental changes. This motivation gap is a reflection of the fact that to date no studies have applied network inference methods to study and compare communities from disturbed and natural environments. However there is evidence that the community structure captured through network inference may reflect the presence of disturbance: bacterioplankton communities from a fresh water lake demonstrated an increase in species richness and network connectance in the spring compared to the summer and fall [89]. Since increased topological connectance typically indicates an increase in network robustness [45, 85], and that the communities from the spring have endured the harsh conditions of the winter [89], the results of this study suggests that the structure of a microbial co-occurrence network may reflect the impact of environmental pressures on community composition. Therefore such impacts may be measured through network robustness simulations. Further evidence of environmental changes impacting microbial community structure through co-occurrence network studies would validate the use of network robustness analysis which would in turn motivate the identification of keystone species using centrality measures.

1.5 Research questions

The following research questions have driven the development and application of the methods and analytical procedures presented in Chapter 2 and Chapter 3.

1. What visualization design and interactive features are appropriate for exploring biological networks such as microbial co-occurrence networks?

2. What kind of patterns, including associations between network properties

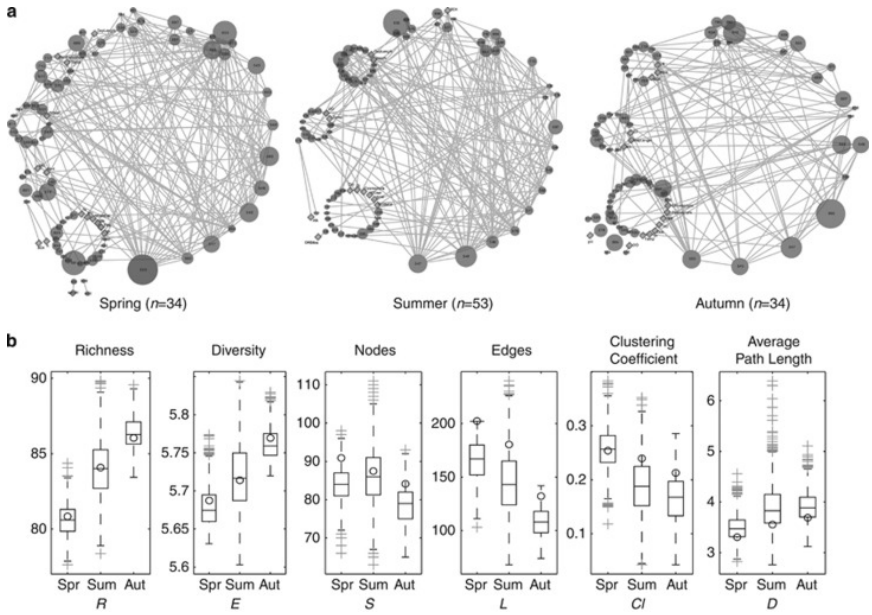

Figure 1.15: Co-occurrence network visualization and properties for a decade long time series of bacterioplankton communities in Lake Mendota in the United States of America. From spring to autumn, the diversity and richness of the community increased while the complexity of the inferred network decreased (from ©Kara et al. (2013)).

and community biotic and abiotic factors, can be resolved in microbial co-occurrence networks? What structural or functional features do these associations imply? At what scales, global or local (i.e. taxonomic), do these patterns and associations occur and are they conserved across different geographical locations or environments?

3. How can the exploration of co-occurrence networks enable the identification of global and local community structures such as microbial keystone species? What is the inferred or hypothesized relationship between the presence of these structures and the community response to change in environmental factors?

4. Do the characterized global and local patterns resolve both ecological conserved principles that are shaping the community and the functional roles of

community structures, such as key microbial species, that are driving it?

1.6 Research overview

Understanding the organization of microbial communities is an important step towards gaining predictive power in microbial ecology. This thesis tackles this challenge by studying microbial community structure and stability across different geographic locations using SSU rRNA sequences and network analysis.

Chapter 2 describes the design of Hive Panel Explorer, a data-driven, interactive and explorative network visualization. Its design is formulated to appropriately adapt to the high dimensionality, complexity and size of biological networks. Its effectiveness in revealing known and novel topological and data association patterns is tested and demonstrated on the *C. elegans* connectome.

Chapter 3 attempts to provide a standard ecologically driven analysis of microbial co-occurrence networks which model the inter-connected communities from the LTSP project. In particular, the robustness of networks obtained from soil communities that have undergone different levels of organic removal will be measured to evaluate the applicability of robustness simulations on microbial co-occurrence networks and to assess the effect of disturbance on community structure. Furthermore, robustness simulations can help identify individual taxa with central positions relative to the community's structure. These taxa's soil profile and taxonomy will then be characterized to evaluate their role in the community. In sum, the ensemble of network analysis conducted on the LTSP dataset has the potential to reveal patterns within and between locations and taxonomic groups and give insight onto the functional roles of individual taxa.

Finally, Chapter 4 concludes with a discussion of the assumptions and shortcomings of the current visualization and analytical methods outlined in this thesis, and lays out future work and improvements to Hive Panel Explorer and the study of microbial community structure through network analysis.

Chapter 2

Hive Panel Explorer: an interactive visualization tool to explore topological and data association patterns in large networks

Networks are used in a variety of fields to relate topological structure to system dynamics and function. Network analysis is often motivated and complemented by network visualization. However, the visualization of large networks is challenging due in part by the abundance of data needing to be visually organized and the difficulty in finding a suitable network layout to resolve patterns associated to system properties. Hive plots provide a general, consistent and coherent rule-based visualization alternative to force-directed layouts and are appropriate for assessing and comparing structural patterns within and between large networks. Despite their flexibility, hive plots can be a daunting visualization technique given, for example, the great number of possible combinations of layout rules that the user must choose from. Here we present HYPE, a visualization idiom and a D3 based tool consisting of a grid of hive plots, whose design follows the visualization mantra

"Overview first, zoom and filter, then details on demand". HYPE aims to make hive plots accessible to the broader scientific community by expanding on the original design and providing a data-driven procedure to construct hive panels and explore large networks interactively. HYPE allows the user to discover topological and data specific patterns across several dimensions simultaneously. Here, we evaluate the different features of hive panels and outline the navigation of a system through its network using HYPE's interactive features by exploring and characterizing the structure of the *C. elegans* neural connectome. HYPE is available for download on Github under the GNU license: https://github.com/hallamlab/HivePanelExplorer.

2.1 Introduction

2.1.1 Network science and visualization

Network approaches are widely used to study relationships between the structure and the function of a system in the social, biological and technological sciences [118]. Networks are composed of nodes and the relationships between them called edges. For instance, nodes represent people in social networks [162] and species in foodwebs [45, 122], while the edges represent their social and trophic interactions, respectively. modelling the relationships of a system, such as a social or ecological community, using networks can describe and characterize the system's structure and dynamics [68, 117, 117]. Network measures such as degree distributions, modularity and connectance, can help formulate associations between elements of structure to notions of function in the system [68, 117, 118]. Several review papers provide an overview of different network analysis and graph theory measures [117, 118, 120] some of which are briefly described here and are illustrated in Figure 1.1.

Visualization idioms are developed to accomplish different visualization tasks [115, 147]. Current network visualizations are designed to present, summarize, annotate, illustrate, investigate or explore the system modelled by the network (Figure 1.5). From force-directed layouts to circular layouts, each design offers insights into different structural elements of a network and the system it represents by highlighting different topological features. For example, force-directed layouts

give relative positions to the nodes encoded as circular marks and edges encoded as links (Figure 2.1A) to show paths (a set of successive edges connecting two nodes) and modules (a subnetwork with increased connectivity within compared to with the rest of the network) [64, 115]. In contrast, adjacency matrices are appropriate visualizations for presenting modules and cliques (a fully interconnected group of nodes) [56, 115]. Many other network visualizations exist such as circular layouts [112] and spectral layouts [65]. These visualizations can help researchers infer system structure and dynamics through network properties. For example, modules can be characterized to assess the biological and functional properties of a group of interacting proteins in protein-protein interaction networks in the context of interactome research [87, 88, 102]. Similarly, cliques are identified and analyzed to gain insight into the tightly knit social circles in the context of social networks [162]. Network visualizations such as force-directed layouts and adjacency matrices were designed to present such patterns in the network [65, 115, 147] but they are limited in the types and number of patterns they resolve.

In order to go beyond presenting known patterns and reveal new ones, visualizations need to be designed to permit an exploration of the system. The discovery of topological features and patterns can help drive a quantitative analysis of the network and formulate hypotheses on a system's structure and function [118, 159]. The discovery process can be accomplished using visualizations that have been designed for data exploration [124–126, 131, 147]. However, the pitfalls of these visualization schemes demonstrate that current network visualizations are not designed to facilitate the exploration and discovery of novel global and local patterns in complex systems [96, 115] Here, we provide an overview of current visualization pitfalls, introduce hive plots as a flexible and versatile network layout and describe how to expand hive plots' potential to build a data-driven interactive visualization for network exploration.

2.1.2 Current network visualization pitfalls

As models of complex systems, networks come in all shapes and sizes, from small networks with a few to a hundred nodes to large networks with hundreds to thousands of nodes and with each node or edge having a virtually unlimited number of

data properties. When scaling up to large networks and to higher levels of multivariate data complexity, effective network visualization becomes an increasingly challenging task. A suitable visualization must be flexible enough to enable the discovery of both global and local patterns across node and edge properties.

Current approaches do not scale effectively when rendering large networks. Specifically, the amount of data displayed shadows its interpretation, particularly in the case of force-directed layouts, which suffer from data occlusion and the high likelihood of pattern misinterpretation [96, 115]. Furthermore, overlaying additional information about the system by colouring or varying the sizes of nodes and edges given data properties, is often impossible without further cluttering the display. Beyond scaling additional visualization pitfalls have been demonstrated. For example, when using hierarchical layouts, finding the location of a deleted node by comparing two visualizations of the network, with and without the missing node, can be a difficult task [96]. In the case of algorithmic-based visualizations, these fallacies stem from the absence of a coordinate system for node positioning. For instance, in force-directed layouts node positions are simulated based on the arrangement of edges connecting them and node positions can change from one simulation to another. These variations can cause inconsistencies between two layouts of the same network, which can lead to different interpretations of the layout. In addition, since the layout is built based on a heuristic algorithm, this visualization is not suitable for the task of visually comparing different networks. These issues make it difficult to explore and interpret network visualizations. Finally, current network visualizations include a lack of generality (applicable to different types of networks), a lack of flexibility (can support different purposes), and their inability to complement other displays [37]. While layouts based on heuristic algorithms are not suitable for large networks, a rule-based layout includes some of the features necessary for building an interpretable, general, flexible and comparable network visualization.

2.1.3 Hive plots

Krzywinski developed hive plots as an alternate network visualization to force-directed layouts, circular layouts, hierarchic layouts, etc. [96]. Hive plots are a

rule-based layout that attempts to provide a coherent and interpretable network visualization [96] while leaving several visualization channels such as colour, size and rule choice, to encode additional data properties. Hive plot's design scales to large networks as it handles visual occlusion and other potential visualization design pitfalls by positioning the nodes using a coordinate system [96].

The nodes and edges' data properties and network measures are used to organize the nodes and edges in the coordinate system. A link is drawn between two circular marks if the nodes represented by these marks have an edge connecting them in the network (Figure 2.2). The nodes are placed onto circularly arranged axes and edges are drawn between nodes using Bezier curves (Figure 2.2). The user defines i) a rule designating a node's axis assignment and ii) a rule designating a node's position along an axis, as illustrated in Figure 2.2. These two rules are chosen using node properties. While axes are used to group nodes with similar properties, node positioning along the axis distributes nodes according to node property values. This coordinate system has many similarities with parallel coordinate plots [84, 96], only in hive plots the axes are organized in a circular fashion and the nodes are not represented on all axes but are assigned to a particular axis.

Any node property indicating the node's position in the network (i.e. degree, clustering coefficient, betweenness value, etc.) or the node's role in the system (i.e. gender in social networks, expression value in gene networks, protein family in protein-protein networks, neuron cell type in connectomes, species diet in foodwebs, etc.), can be used to construct rules designating axis assignment. A visual overview of node network measures is presented in Figure 1.1. In the social network displayed in Figure 2.1, the individuals are assigned to the hive plot's axes according to their gender (boy, girl or alien) and are positioned along their respective axis by the number of relationships they have in the network i.e., degree. This layout allows the viewer to investigate relationships between gender and the degree of an individual in the social community and quickly answer questions such as "Is the individual with the most relationships a boy, girl or alien?". The hive plot in Figure 2.1 shows that aliens generally have a high degree (more social relationships than boys and girls) and that the person with the highest degree is an alien called Zans (pattern 1). An even more striking pattern is that boys and girls share no connections (pattern 2) and that all boys who share relationships are en-

emies (pattern 3). Resolving the same patterns in the force-directed layout would involve searching the layout and counting degrees and edge types. Thus, Figure 2.1 demonstrates that certain patterns are often more difficult to discern using the force-directed layout compared to a hive plot, even for such a small network.

Difficulty in pattern resolution increases with network size and system complexity. Hive plots facilitate pattern resolution by projecting the network using node properties to create a consistent, interpretable and flexible layout that is appropriate for interactive network exploration.

2.1.4 Hive Panel Explorer

Despite the fact that hive plots were demonstrated to satisfy the requirements for an effective aesthetic layout (i.e. are general, flexible, reproducible, comparable etc.) [37] and to be an appropriate visualization for networks [96], they have been criticized for their lack of accessibility. Specifically, though the coordinate system ensures that the plot can be interpreted, users have had difficulties understanding how to harness the versatility of hive plots for their particular networks and research questions. HYPE attempts to address this criticism by providing the user with a data-driven approach to constructing hive plots and interactive methods to explore them.

Hive plots are appropriate for visualizing networks of all types: weighted or un-weighted, directed or undirected, complete or not. Several hive plots have been used in scientific literature to describe gene expression networks [130], splicing patterns in RNA sequencing data [172], and neural connectomes [153]. In each case, the users have chosen the rules to assign and position nodes from a large set of combinations to present the dimensions of the system they were interested in. The layout's flexibility and adaptability empowers the user to explore both network specific and data specific properties. At the same time the user must determine optimal assignment and positioning rules to produce a single hive plot, potentially discouraging the use of different combinations.

We have developed HYPE to produce a matrix layout of multiple hive plots called a "hive panel" [96]. This design circumvents the need to determine the optimal set of rules and enables the user to explore multiple combinations of network

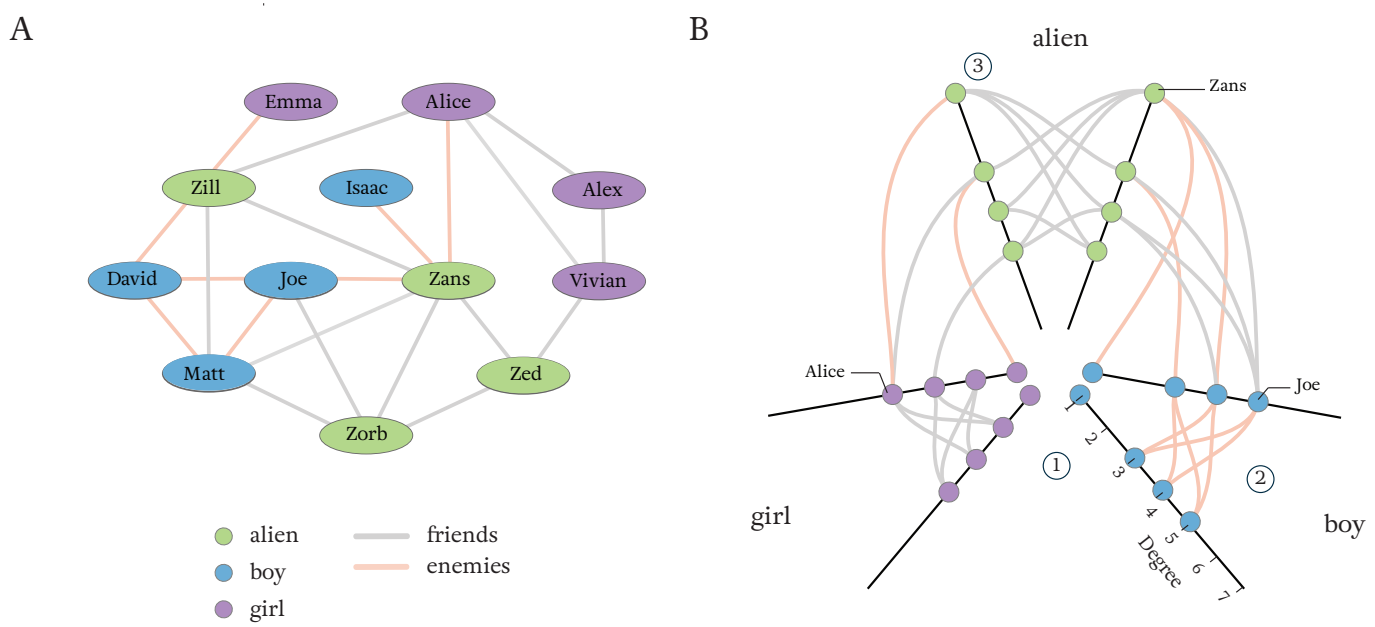

Figure 2.1: A comparison of a force directed layout and hive plot of a social network. A) Aliens, boys and girls with relationships are laid out to minimize node overlay and edge crossing. B) A hive plot of the same social network shows 3 friendship patterns. Friends were grouped onto axes by gender and were positioned along axes by degree.

measures and data dimensions. HYPE also provides different data transformations to scale the layout of nodes and edges according to the attribute values used to place them. We argue that visualizing and exploring multiple projections of the network in a coordinated layout facilitates the discovery of topological and association patterns in the network. Interactive features such as colouring, highlighting, look-ups and selective filtering enable the exploration of the system. Additional colour encoding of the circular marks and links, representing the nodes and edges respectively, draws attention to certain nodes or edges to both facilitate the network exploration and produce publication ready figures. To our knowledge HYPE is the first idiom and tool to utilize hive panels and to provide interactive features as an integral part of its design to visually present and explore networks.

Here we rationalize HYPE's visual idiom as an appropriate visualization for network exploration using a general visualization paradigm [115] and we motivate its interactive features using Shneiderman's visualization mantra and associated methodology [147]. We then provide guidelines for building informative panels, exploring them, and finding known and novel patterns in the network. Finally we benchmark HYPE on data from the *C. elegans* connectome to demonstrate it capacity to find meaningful relationships in model systems.

2.2 Methods

2.2.1 Visualization idiom and design

HYPE's design is based on a matrix of hive plots and was developed with networks in mind though any dataset with relationships between data items can be appropriately visualized in a hive panel. Table 1 outlines the general design of HYPE according to a general visual paradigm and language [115] and describes data transformation, visual network encoding, visualization tasks, and the size of the dataset it is appropriate for.

Hive plots were designed to permit the organization of nodes according to node properties. From here on we denote an attribute as either a network property such as degree (number of edges a node has) or as a node property that is inherited from associated multivariate data. These attributes can be categorical (e.g. source or

sink node in directed networks, neuron cell type in connectomes, etc.) or quantitative (e.g. node's clustering coefficient, age of an individual in a social network, etc.). Using these attributes, the nodes can be grouped and sorted before they are organized on the axes.

Typically, a layout of 3 axes is chosen to allow for edges between nodes on any axis to be drawn without crossing over another axis. When using a layout with 4 or 5 axes, HYPE doesn't draw edges between non-neighbouring axes as these curves would affect the interpretability of the visualization. In order to view edges between nodes that have been assigned to the same axis, a mirror image of each axis is produced with edges draw between reflected nodes. Figure 2.2 presents a skeleton of this rule-based visualization with the two layout schemes: single axes (Figure 2.2A) and doubled axes (Figure 2.2B).

Each hive plot visually encodes node attributes and thus by comparing multiple hive plots one can assess associations between pairs of node attributes simultaneously. To facilitate this comparison we construct a hive panel, a set of multiple hive plots organized on a grid. This visualization then becomes a matrix type layout where each column denotes the use of a particular axis assignment rule and each row denotes the use of a particular axis positioning rule (Figure 2.5). Since multiple node attributes are used as layout rules, different visual projects of the network onto the axes are presented and compared.

2.2.2 Designing a hive panel

HYPE enables users to construct hive plots and panels using a data-driven approach. Here we present the different features of HYPE as well as how to best utilize them depending on the type of system properties being explored.

Choosing assignment and position rules.

In a multivariate dataset, nodes can have numerous categorical and quantitative properties that can in turn be used as one of the two plotting rules. Certain node properties are more suitable as axis assignment rules than axis position rules and vice versa, as summarized in Table 1. For example, a categorical attribute with three categories is most suitable to group the nodes by their attribute onto the 3

Table 2.1: HYPE's visual design idiom

Idiom		Hive Panel Explorer	
What?	The data	Networks where nodes and edges have attributes which can include calculated network properties (ordinal, quantitative and categorical properties)	
	Deriving the data	1. Calculate network properties (i.e. degree) 2. Organize nodes by desired attributes into groups 3. Normalize/scale node attributes	
How?	Encoding the data	Data attribute	Mark or channel
		Nodes	Circle
		Edges	Link
		Node attributes (mostly quantitative)	Position on axis
		Node attributes (any type of attribute)	Grouping on axis
		Node or edge attribute	colour, visibility
Why?	Visualization tasks	Actions	Targets
		Present and summarize	Distribution of nodes properties
		Discover	Topology, outliers and patterns
		Explore	Characteristics of topology, outliers, and patterns
		Compare	The position of grouped nodes and edges in different hive panels of one network. Topologies and distribution of node properties between two networks.
Scalability		Nodes: dozens to thousands edges: hundreds to few tens of thousands.	

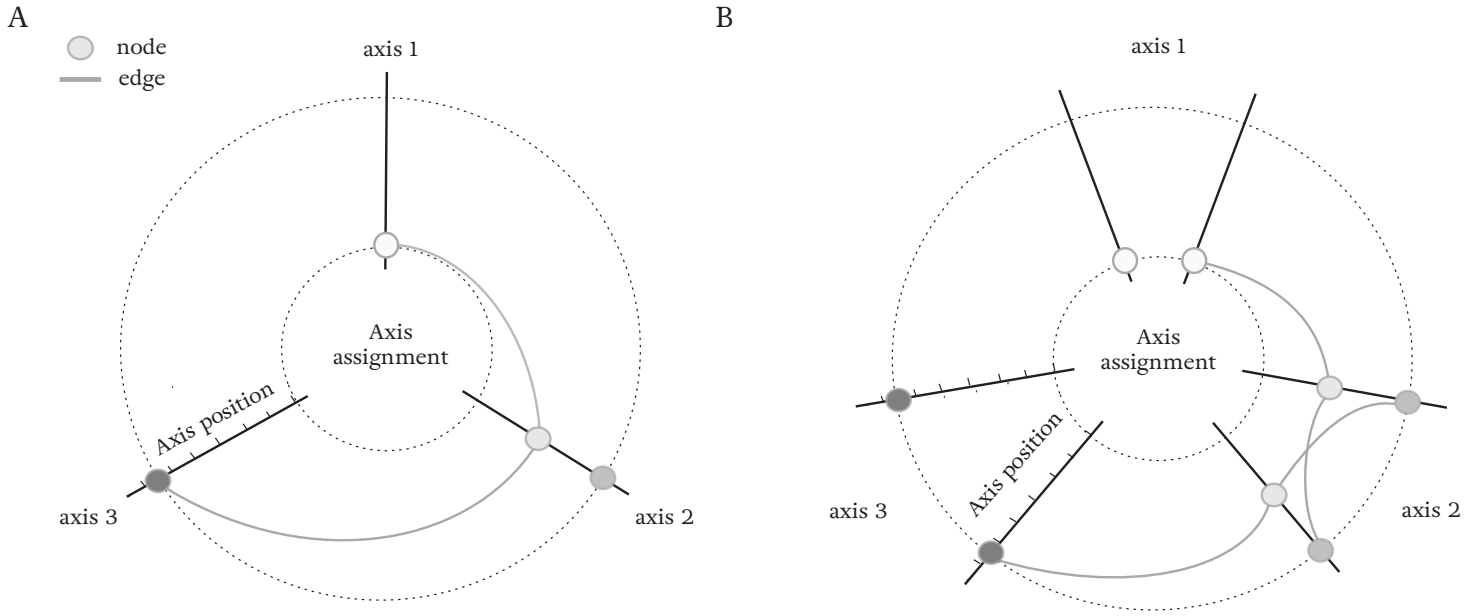

Figure 2.2: A schematic layout of single and double axis hive plots. A) Nodes are grouped onto and positioned along three axes. B) Double axes can be used to view edges between nodes grouped on the same axes.

axes. However, categorical attributes with several categories (greater than 3) can be used as axis positioning rules in which case they are ordered alphabetically and positioned accordingly along the axis (Figure 2.3B). On the other hand, quantitative attributes can be used for either rules: while axis assignment rules organizes the nodes into low, medium and high values of this attribute (Figure 2.3A), axis position shows the distribution of nodes given this attribute in more detail (Figure 2.3). Using these rules of thumb, a panel can rapidly be constructed. If an attribute is not suitable for either plotting rule, or if the user wishes to compare this attribute across all hive plots, then it can be used to colour the nodes. Taken together, any node attribute can be presented in three ways: as an axis assignment rule, as an axis positioning rule or as a colouring rule.

Data transformations through partitions and scales.

In order to adapt to different distributions of node quantitative attributes, the attribute's values can be mapped to plotting positions using a **linear**, **even** or **logarithmic** partition for nodes' axis assignment and a **linear**, **rank** or **logarithmic** scale for nodes' axis positioning (Figure 2.3A and 2.3B). If unspecified, a linear partition or scale is used for either plotting rule. An even partition determines the cut-offs to evenly distribute and assign nodes to an axis so that all axes contain the same number of nodes (give or take one node) (Figure 2.3A). Rank based scales are used for axis positioning to ensure that no two nodes overlap, even if they have the same attribute value: if "degree" is the axis position attribute and three nodes of have a degree of 5 then they are placed in succession and in an arbitrary order along the axis (Figure 2.3B). A logarithmic partition or scale is best suited for attributes that are distributed exponentially. If a user is interested in displaying all node attribute values, without overlap, than a rank or even scale should be used. On the other hand, a linear or logarithmic partition or scale are appropriate for showing outliers, nodes whose attribute values differ drastically from the other nodes' values. The choice of different partitions or scales allows the user to construct panels in a data-driven manner.

Known network topologies are best displayed using specific data transformations. A linear partition or scale is appropriate when plotting nodes by degree in

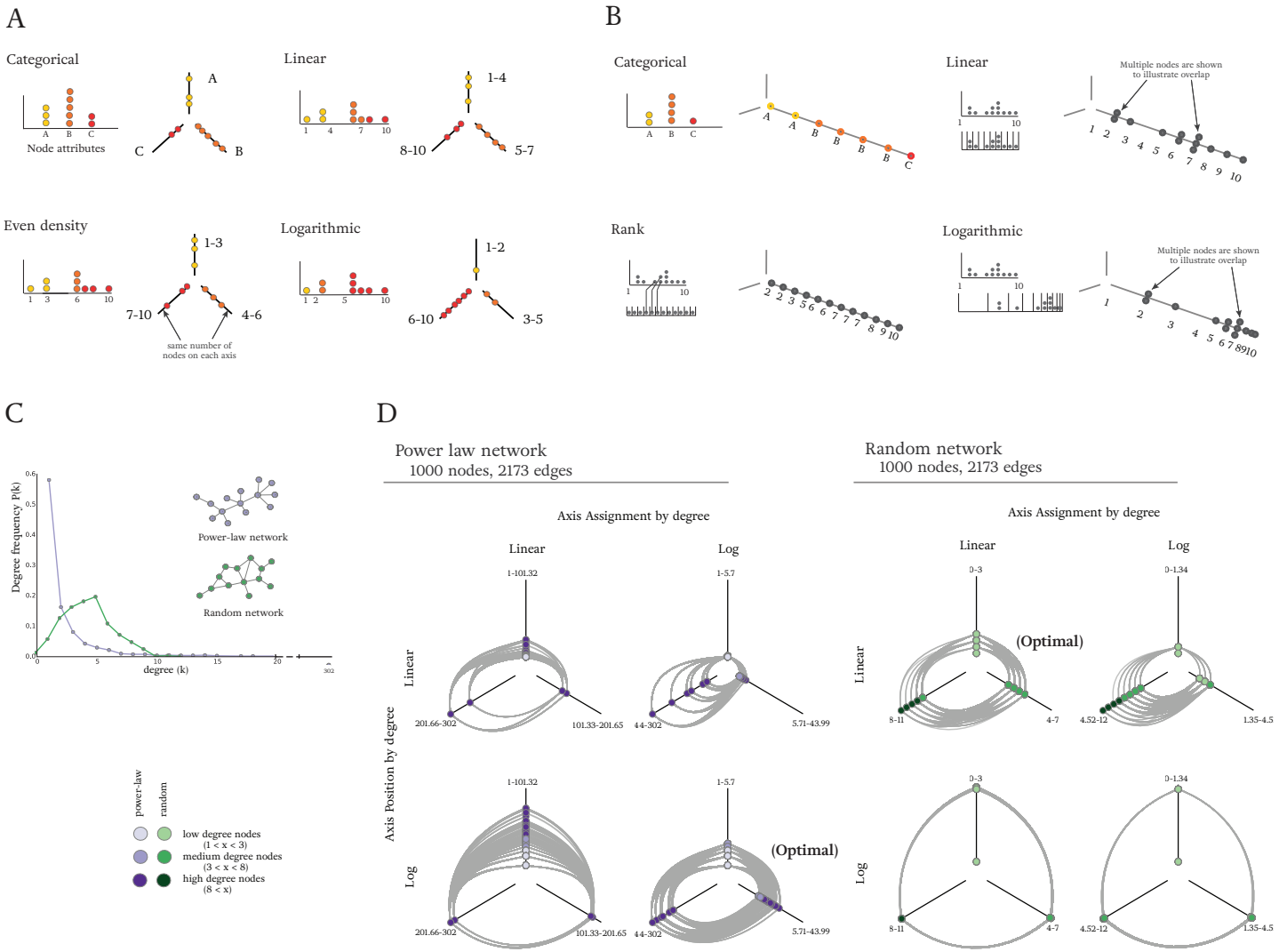

Figure 2.3: An overview of the possible partitions and scales driving node assignment and positioning. A) A schematic of the four possible axis assignment partitions. B) A schematic of the four possible axis positioning scales. C) The degree distributions of a random network a power law network with the same number of nodes and edges. D) Hive panels of the random and power law networks showing the efficacy of different partitions and scales used to organized nodes based on their degree.

networks with a binomial degree distribution, such as random networks. A log scale is better suited for networks with an exponential degree distribution, such as power-law networks. Figure 2.3C shows the degree distributions of these two common network types and the optimal choice of partition and scale for drawing these networks in hive plots is illustrated in Figure 2.3D. Determining the appropriate plotting method can both facilitate the interpretation of node positions as well as avoid having circular marks and links overlapping. When used appropriately, axis assignment partitions and position scales help maximize the total visual space occupied by the circular marks.

Once assignment and position rules have been chosen along with their partition scheme and scales, the user obtains a hive panel with each hive plot providing different visual projections of the system. Though performing successive refinements of the initial layout can be tempting, preliminary exploration of the network with the first constructed panel is encouraged to guide future iterations.

2.2.3 Navigating a hive panel

The layout and interactive features of the HYPE allow the user to explore their data by following Shneiderman's visualization exploration mantra "Overview first, zoom, filter, then details on demand" [147]. The organization of the layout fulfills the first two components by creating an overview of the network and allowing the user to "zoom in" on a subset of the hive plots. Once the layout is selected, there are five ways to interact with the system using HYPE: colouring, highlighting, searching, selecting and filtering. colouring and filtering effectively increase and decrease the visual salience of a node or edge, respectively. A user can then visualize "details on demand" in three ways: searching or highlighting a single node or a edge, and selecting multiple nodes or edges. An organization of HYPE's interface and features is illustrated in Figure 2.4.

Overview. A large panel size such as a 4*x*4 hive plots provides an overview of the network: the 16 unique hive plots present different visual projections of the system against different node attributes. Quickly, the user can assess global patterns in the network such as particular degree distributions apparent from the layout of nodes and edges in a Degree (linear) by Degree (linear) or a Degree (log)

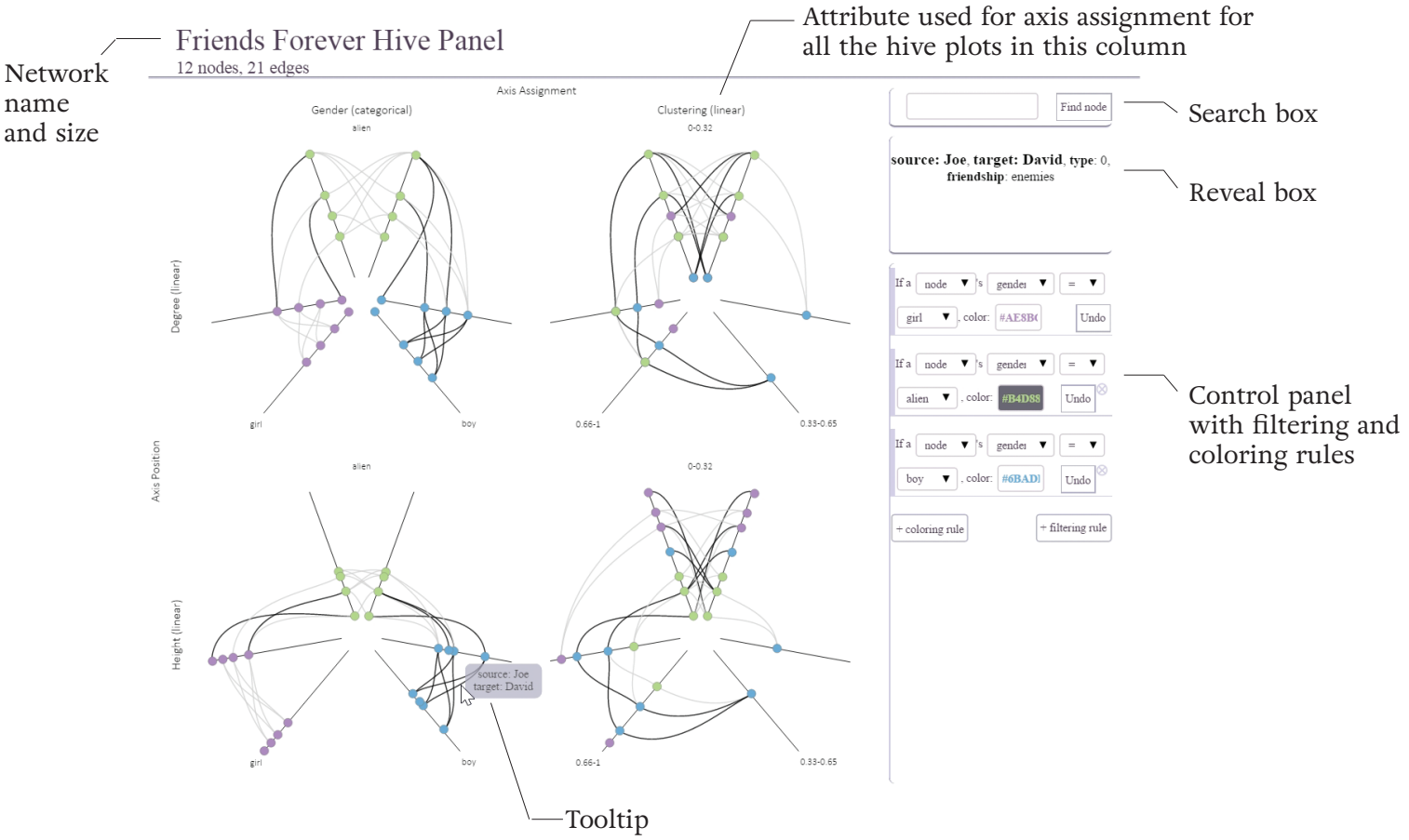

Figure 2.4: An overview of HYPE's interface.

by Degree (log) (Figure 2.3D).

Zooming. Starting with a large panel, one can reduce the amount of data shown by decreasing the number of hive plots. Since the size of the visualization area doesn't decrease, one effectively zooms in on the hive plots remaining, which become larger in size.

colouring. Nodes and edges can be given visual emphasis by colouring them according to their data or network attributes. Simple equality and inequality expressions, such as "equal to", "greater than", and "less than", are used to colour nodes and edges by quantitative attributes (Figure 2.4). The position of the coloured nodes or edges in each hive plot can then be compared. colouring is especially useful to show or compare the position of particular groups of nodes in each hive plot by assigning a colour to each group. This feature thus permits additional comparisons across network and data properties.

Filtering. While colouring can be used to draw attention to nodes and edges, filtering can be used to hide nodes and edges that are not of interest and that may be cluttering the display by sharing positions in the layout with other nodes. Hiding these nodes and edges may facilitate the resolution of hidden patterns. Given their attributes, nodes and edges can be filtered using two modes "keep" (hide all but the selected nodes or edges) and "hide" (hide the selected nodes or edges). When nodes are filtered, the links encoding their edges are also hidden.

When nodes or edges are coloured or filtered, the number of objects and the action chosen is shown in the reveal box in the top right of the interface. This information helps the user assess how many nodes or edges have the attribute values selected and have received a particular visual encoding. However, it is important to note that in a double axis hive plot certain nodes and edges are drawn twice and in single axis hive plot certain edges are not drawn at all. In other words, the number of nodes or edges coloured or filtered using the colouring or filtering rules may not be equal to the number of circular and link marks that have been visually encoded. The reveal box relates the number of selected nodes or edges independently of their visual encoding.

Highlighting. By hovering over nodes or edges, the name of the mark and the value of its attributes used for the layout rules are revealed in a tooltip window (Figure 2.4). This feature allows for rapid identification of nodes or edges given

their position in a plot. In particular, highlighting can be used to identify outlier nodes and edges.

Selecting. Clicking on a node or edge will cause each instance of the corresponding visual mark to increase in size and in opacity. This selection creates a "pop out" effect in the entire panel (Figure 2.4) allowing the user to compare the position of a node or edge in different hive plots. In addition to being visually revealed in the panel, the select node or edge's attribute values are shown in the reveal box (Figure 2.4). Several circular marks and links can be clicked successively to select multiple nodes and edges. Once one of the selected marks is clicked on again, all of the selected marks become "unselected" and return to their normal size and opacity.

Searching. Nodes and edges can also be "searched" using the search box in the top right corner of the interface. Identified nodes will be selected and "popped out" in the same way as clicking on the node would (Figure 2.4). Searching for a node is useful when the user is interested in locating a particular node whose position in the hive plots is not known.

Examples of the different applications of these seven features to present a network, explore its structure and generate hypothesis are demonstrated on the *C. elegans* connectome, a well characterized and studied neural network.

2.2.4 HyPE as a web tool

The HYPE tool was built using D3 [20, 21], JavaScript [133] and Python [142]. D3 was used for building and rendering the data-driven interactive graphics. D3 was an ideal candidate to build HYPE because it produces scalable, interactive, data-driven and web-based graphics [20]. Mike Bostock's hive plot plug-in was used to generate the positioning of the nodes and edges and its license is included in HYPE's documentation. All scripts, a wiki and tutorials are available on Github under the GNU license: https://github.com/hallamlab/HivePanelExplorer.

HYPE takes as input a tabular file in .csv format to facilitate the addition of node and edge properties. While many file formats have been developed to encode networks, these can easily be converted to .csv files using export functions in software such as Gephi [15] and Cytoscape [146].

Once the hive panel is explored and patterns are identified, users may want to use the panel to present their findings. The export functionality allows the users to obtain publication ready figures in a SVG format to allow for further editing in vector graphics manipulation software. In addition, the set of colouring and filtering rules applied can be exported in text format to help users keep track of the visual encodings they used while permitting figure reproducibility.

2.3 Results: the structure of the *C. elegans* connectome

2.3.1 The system

C. elegans is a model organism whose nervous system can be serially reconstructed using imaging technologies to provide a comprehensive wiring diagram of neuronal connectivity over developmental time [25, 164, 166]. Moreover, the anatomical location, the developmental history and the functional role of all neurons within the nervous system is recorded in public databases such as the Worm Atlas [8]. The *C. elegans* nervous system can be modelled with nodes representing neurons and edges representing synaptic connections. The resulting network is commonly called a connectome [25]. The *C. elegans* connectome is a logical and informative connectomics model that has been extensively studied using both experimental and quantitative modelling approaches [23, 71, 72, 159, 161, 163, 165].

Several studies have applied network visualization and graph theory measures to analyze different properties of the connectome including wiring efficiency and cost [34], the relation between connectivity and neural development [159], the rich club [157] and small world structure [163]. Here we use HYPE to explore the *C. elegans* connectome. We demonstrate the application of HYPE's data-driven design and interactive features to reveal both known and novel properties of the nervous system.

2.3.2 The network

The connectome studied here is that of a hermaphrodite worm with 279 somatic neurons and 2,287 synaptic connections. The initial construction, successive refinements and limitations of this dataset were previously described by [159] and

several studies have analyzed the resulting network [157, 159]. The nodes (neurons) and edges (synaptic connections) have categorical and quantitative attributes. Notably, the neurons' location along the posterior-anterior axis (head to tail) of the worm has been measured. Neurons come in three types: motor neurons connect muscular cells to the nervous system, sensory neurons connect sensory cells to the nervous system, and interneurons connect two neuronal cells. Synapses come in two types, chemical or electrical, which correspond to the signal being transmitted using either neurotransmitters or an electric potential, respectively [13, 25, 164, 166]. The direction of individual synaptic connections was not encoded in this visualization. The connectome is a complete network (all pairs of nodes are connected by some finite path) and has a small world structure (a combination of a high average clustering coefficient and low average path length) [159, 163].

2.3.3 Constructing the hive panel

The hive panel in Figure 2.5 was constructed using combinations of the neuronal attributes described above and different network properties as axis assignment and positioning rules. Since there are exactly three types of neurons (motor, sensory and inter), neuronal cell type can be used to group neurons onto axes as an axis assignment rule. Somatic position is a quantitative attribute that could be investigated using an axis position rule. Other node attributes, such as cell class and associated neurotransmitter, are all categorical attributes with several possible values and thus are best investigated using colouring and filtering rules once the panel is built.

In previous studies, the *C. elegans* connectome was shown to be a scale-free network with a power-law degree distribution [159, 163]. We therefore opted for using a logarithmic scale to position the nodes by degree. This attribute will allow us to evaluate a possible relation between the number of connections of a neuron and its role in the connectome. To compare the degree of nodes to other attributes that will also be used as a position rules, we select degree as both an axis position and assignment rule. We use an even partition to distribute the nodes onto axes equally.

There are many other network properties we can explore. In particular, previ-

ous studies have found significant numbers of three node cliques, or triangles, in the *C. elegans* connectome [159, 166]. To investigate patterns relevant to neuron cell clusters we include clustering coefficient as a plotting rule. Since we are more interested in the magnitude of the clustering coefficient (high or low) versus the absolute value, we select clustering coefficient as an axis assignment rule. In order to focus on nodes with a high clustering coefficient, we use a logarithmic partition.

Rich club structure [157] and wiring efficiency [34] studies have demonstrated that the *C. elegans* connectome has a relatively efficient structure: neurons are strategically connected and positioned along the worm's posterior-anterior axis to minimize the wiring cost (relative to the total number of synapses) of each neuron and synapse [34, 157]. In particular, the position and connectivity of interneurons suggest that they play the role of information highways in the connectome [157, 159]. Accordingly, we expect interneurons and their connections to decrease the path length needed to connect two neurons and therefore to have high betweenness centrality values. Because we have determined three axis assignment rules (cell type, degree, clustering coefficient) and two axis position rules (somatic position and degree), we choose betweenness centrality as our third axis positioning rule to complete the example.

Using our knowledge of the system, we chose 2 system properties and 3 network properties to construct a $3x3$ hive panel and explore the network (Figure 2.5). Unless specified, we used a linear partition or scale for the layout rules.

2.3.4 Exploring the *C. elegans* hive panel

In the following section, **bolded key words** express the use of different interactive features. When focusing on individual hive plots in Figure 2.5, these key words are designated using the following plotting rules: *Assignment rule by Position rule*. Before initiating exploration of the network, we **colour** neurons by cell type to compare their distribution in the panel. Observed patterns are interpreted in the Discussion section.

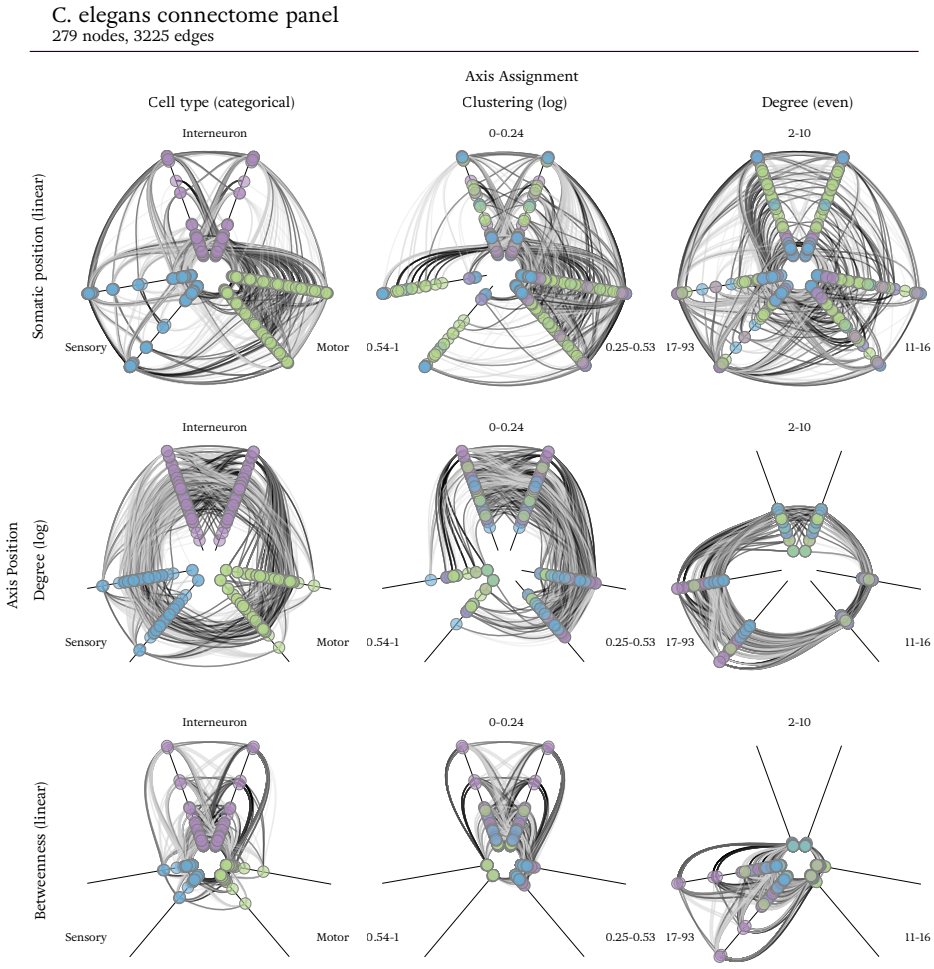

Figure 2.5: The *C. elegans* hive panel with nodes and edges representing neurons and synapses, respectively. Neurons are coloured by cell type and synapses are coloured by synapse type: electrical synapses in black and chemical synapses in grey.

Overview of cell types and positions

Looking at the whole panel, one can discern the presence of edges connecting nodes from all axes in each hive plot. For instance, the *Cell type by Somatic position* hive plot shows that all three types of neurons share connections within and between types. If sensory neurons and motor neurons weren't connected through synapses, we wouldn't see any edges between the two axes where each type of neuron is represented.

Looking at the somatic position of the nodes by their position on the axes in the *Cell type by Somatic position* hive plot, we observe that sensory neurons and interneurons occur at the head, tail and at few discrete positions along the posterior-anterior axis whereas motor neurons cover the whole length of the worm. Effectively **zooming in** by only showing *Cell type by Degree* hive plot, we can get a better look at the distribution of neurons across the length of the worm. Looking at the distribution of synapses in this hive plot, we find a high density of synaptic connections between neurons located in the head and between neurons located in the tail. Furthermore, many synaptic connections between motor neurons start and end at similar somatic positions. These observations are consistent with a study by Varshney and colleagues who illustrated the same patterns using an adjacency matrix [159] and points to coordination between adjacent motor neurons to facilitate sinusoidal movement.

In contrast, interneurons connect to each other and to sensory or motor neurons from varying and often opposing somatic positions (i.e. head to tail and tail to head): interneuron-interneuron, interneuron-motor and interneuron-sensory synaptic connections link nodes near the center of the hive plot and nodes near the outer edge. This pattern suggests that interneurons connect physically distant neurons in the connectome. To further characterize the connectivity of interneurons and infer their role in the system, we look at another network measure: betweenness centrality.

Interneuron connectivity

A node's betweenness centrality measures the importance of its position in the network relative to other nodes. As illustrated in Figure 1.1C, nodes whose con-

nections reduce the paths between other nodes or whose absence would create disconnected subnetworks have a high betweenness centrality value. We can observe the centrality of neuronal cell types in the *Cell type by Betweenness Centrality* hive plot. We first notice that on average interneurons have higher centralities than sensory and motor neurons. Using the **tooltip** we can find the maximum betweenness centrality value per cell type: sensory, motor and interneurons have a maximum of value of 0.028, 0.036 and 0.103, respectively. Wiring efficiency studies have found that interneurons and their synapses reduce the path length between other neurons [34, 159]. The hive panel resolves this pattern by illustrating how interneurons connect physically (across the worm's length) and topologically (across the network's path structure) distant nodes.

Clustering coefficient and connectivity patterns

The clustering coefficient expresses the connectivity between neighbours: if all of its neighbours are connected than a node has a clustering coefficient of 1. We can observe clustering between neuronal cell types in the *Clustering coefficient by Somatic position* hive plot. Here neurons that were grouped on the axis with high values (i.e. whose neighbours are connected at a rate of over 54%) are primarily motor and sensory neurons in the body and the tail. Using the tool tip we can quickly survey their synaptic connections to low and medium clustering coefficient neurons on the other axes. If we look at the *Clustering coefficient by Degree* hive plot, we notice that these neurons have medium to low degree and share synapses with medium to high degree nodes. For example, using the tool tip we find that the motor neuron DB06 has a degree of 7, a clustering coefficient of 1 and is therefore part of a 7-node clique. These highly connected cliques are characteristic of a small world network [118, 163]: a significant small world coefficient implies that the path from any two nodes is relatively short despite the large number of nodes in the network.

We can **colour** the neurons with clustering coefficient of 1 using alternative colours to those used for neuronal type: the **reveal box** indicates that we have coloured 9 neurons. These high clustering coefficient neurons are thus part of cliques whose other members have high degrees and are very connected in the

network. This topological pattern suggests that these motor and sensory neurons might relay signals from sensory cells or body wall muscles to the rest of the connectome. Evaluating the direction, strength and the type of synapses between the members of the clique could further resolve this pattern.

Using filtering to partition the system and study subnetworks

To further understand differences between neuronal connectivity patterns along the anterior-posterior axis, we **filter** the head neurons (somatic position < 0.2, 140 nodes) and the tail neurons (somatic position > 0.65, 51 nodes). Filtering only the head neurons, we notice that all of the interneurons with degree greater than 43 are missing from the *Cell type by Degree* hive plot. Filtering only the tail neurons, we observe that a few low and medium degree interneurons and one high degree sensory neuron and motor neuron with high betweenness centrality are missing. We can identify these neurons using the **tooltip** and then **selecting** them. The sensory neuron is identified as PQR (degree = 54, betweenness = 0.028) involved in several processes including aerotaxis and social feeding [70]. The motor neuron is called DD06 (degree = 50, betweenness = 0.036) and innervates dorsal body walls muscles along with DD1-DD5 neurons [8, 177]. Filtering out both head and tail neurons, we notice that many medium degree nodes with high betweenness centrality remain. These observations suggests that high degree interneurons play important roles relaying signals at the head of the worm, and a few medium to high degree interneurons along with one key sensory and one key motor neuron permit centralized signalling in the tail.

Characterizing synaptic connections

Next, we explore network edges by colouring them to gain insight into synaptic connections and their distribution in the network. We use 2 distinct colours to distinguish between electrical and chemical synapses. We notice that two bundles of electrical synapses seem to be connecting several neurons to two interneurons with high degree. **Selecting** these neurons reveals that they are interneurons AVAL and AVBL. These are command neurons responsible for forward and backward locomotion, respectively [8, 177]. We can locate these nodes at the head of the

worm in the *Cell type by Soma position* plot and assess that they have high degree betweenness centrality from the *Cell type by Betweenness centrality* hive plot, as expected.

Next we study the weights of synapses, which simply represent the number of synaptic connections between neuron pairs. To do so, we **filter out** the synapses with low weights (weight ≤ 10) (Figure 2.6). Looking at the information in the **reveal box** we know that we have coloured 51 synapses. In the *Cell type by Somatic Position* and the *Cell type by Degree* hive plot we notice that these synapses primarily connect low to medium degree motor neurons (node degree between 2 and 21) (Figure 2.6A). To further assess the connectivity of these motor neurons, we **filter out** sensory neurons and interneurons and their edges, so that only heavy weighted synapses between pairs of motor neurons are **revealed** (Figure 2.6B). Looking at the *Cell type by Degree* hive plot and the *Cell type by Somatic position* hive plot, we observe that most connections occur between motor neurons with a degree between 7 and 21 and are located primarily in the body of the worm.

2.4 Discusion

2.4.1 Assessing patterns and generating hypotheses

Using the different interactive features of HYPE we explored known local and global topological patterns in the *C. elegans* connectome. We focused our exploration on somatic position, neuronal type, degree, betweenness centrality, and clustering coefficient by setting these attributes as plotting rules. We found associations between these attributes to reflect known properties of the system such as its wiring efficiency. Quantitative analysis of these patterns in developmental time or between mutant and wild type strains can be used to evaluate the significance of these patterns and generate testable hypotheses. For example, we found a few outlier neurons with higher betweenness centrality values than other neurons of the same cell type. To assess if these values are expected in networks with similar topologies, we can compare these values to the maximum betweenness centrality value found in simulated networks with a similar structure to the *C. elegans* connectome. We find that randomly generated scale-free networks with the same

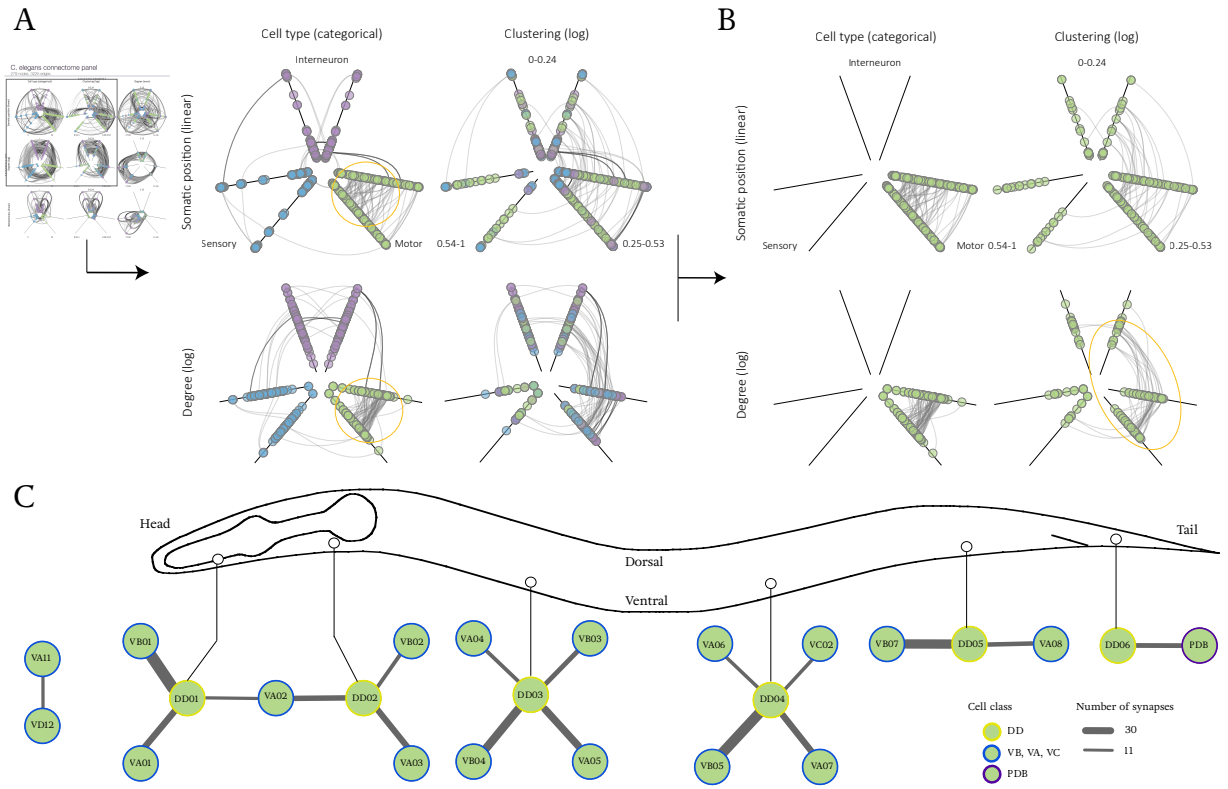

Figure 2.6: A schematic of the filtering procedure used to reveal motor neurons connected by more than 10 synapses. A) From the original panel, edges representing 10 or less synapses are filtered out. The numerous synapses left between motor neurons is highlighted by a yellow circle. B) All interneurons and sensory neurons are filtered out. C) A closer look at the connections between these motor neurons reveals all DD neurons and a part of the dorsal muscular inhibitory circuit activated during worm locomotion. Motor neurons are coloured by cell class and the number of synapses of each edge is illustrated.

number of nodes as the connectome have much smaller betweenness centrality values: the interneurons AVAL and AVAR with a betweenness centrality of 0.103 and 0.101 respectively, are much more central than expected ($p < 0.05$).

Whereas many connectome studies focus on circuits within the connectome, HYPE invites the user to investigate the connectivity and roles of individual neurons, as well as compare the connectivity of neurons with the same cell type. For example, the sensory neuron PQR, which was found to have a betweenness centrality much higher than the other sensory neurons has a connectivity pattern that suggests it plays a more central role than the other sensory neurons. The PQR neuron has been implicated in different physiological aspects and behavioral phenotypes of the worm, including oxygen sensing, innate immunity modulation, social feeding and locomotion related to feeding [70, 165]. The other individual neuron that HYPE revealed as an outlier in its connectivity pattern was the motor neuron DD6 with a betweenness centrality of 0.036. While this motor neuron shares certain characteristics with other DD motor neurons, it also has a higher degree (50 compared to DD1 which has a degree of 21) and exhibits an alternative gene expression profile [106]. The difference in connectivity pattern and biological associations of these individual neurons compared to their other neurons of the same type suggests they play multi-functional roles in the system.

In our exploration we also found that heavy weighted synaptic connections tend to connect pairs of motor neurons. To put this pattern in perspective, considering the six possible ways of connecting three cell types as well as the number of neurons in each cell type, the probability of one of these 51 highly weighted synapses to be between two motor neurons is about 0.1. Therefore the large proportion of weighted motor-motor neuron connections (36%, or 18 out of the 51 synapses) is significant considering the null hypothesis that the weighted synapses are distributed evenly between neuron cell types ($p < 0.001$).

To interpret these results, one must consider functional implications of multiple synapses between pairs of neurons. Though there is a huge variation both in the number and the physical size of synapses, the morphology of synaptic connections between pairs of neurons has been found to be related to the functional strength of the interaction between the neurons [13, 14, 86]. The motor neurons connected through multiple synapses resolved by our exploration include DD1-DD6, several

VA and VB, one VC and the PDB neuron (Figure 2.6C). The motor circuits involving DD, VA and VB found in repeating structures along the body of the worm are responsible for propagating sinusoidal movement [177]. Specifically these multiple synapses occur in the motor circuits responsible for inhibiting dorsal muscle contraction and innervating ventral muscle contraction [43, 177]. The complementary circuits are composed of the VD, DA and DB motor neurons that inhibit ventral muscle contraction and innervate dorsal muscle contraction [43, 177].

Knowing that both complements of these motor circuits are required for the sinusoidal movement of the worm, why do neurons in one complement have more synapses per connection than the other? To the best of our knowledge this pattern has not yet been characterized, however, biological differences between the neuron cell types involved in these circuits have been studied. First, whereas DD and VD neurons play similar roles as inhibitors of dorsal and ventral muscles respectively, the VD1-VD13 neurons develop post-embryonically whereas the DD1-DD6 neurons develop in the embryo and change their synaptic connections after the birth of VD neurons [72, 161, 164]. Second, these two motor neurons classes differ in their expression of certain genes, including a gene related to acetylcholine receptor subunits [61]. Third, DD motor neurons may regulate the amplitude of the sinusoidal movement as suggested by UNC-25 and UNC-30 gene mutants [109]. Therefore, the differences in their connectivity pattern motivates the investigation of possible differences in the biological roles of DD and VD neurons, some of which have been presented here, and these functional differences may be associated with their role in locomotion.

Models of *C. elegans* locomotion propose an asymetry in the neuromuscular system of the ventral and dorsal sides of the body to explain the initiation of locomotion from any intial worm shape [23, 90]. While one model suggests this asymetry may be facilitated by non equal numbers of VD (13) and DD (6) neurons [90], the other model suggests it can be facilitated through a lower activation threshold for neuron firing in VD neurons [23]. The difference in the connectivity pattern of VD and DD neurons resolved here suggests that the physiology of their synaptic connections may also play a role in this asymetry and should be included in locomotion models. Further characterization of these neurons' synapses may reveal a relationship between the embryonic and post-embryonic development, the

varying response to gene expression, and differing role in locomotion of the DD and VD motor neurons.

Despite the fact that the C elegans connectome has been thoroughly characterized in previous studies of the network's structure, we demonstrate that HYPE can reveal new local and global patterns. Specifically, we suggest the possible association between the multi-functional roles and differential connectivity of individual neurons, and between the asymetry of locomotion models and the difference in the number of synapses between VD and DD motor neuron circuits. As observed, the patterns resolved through network exploration motivate quantitative analysis and the results thereby produced can help formulate hypothesis relating the structure and function of the system.

2.4.2 A flexible and adaptive visualization tool

Using the *C. elegans* connectome we demonstrated how HYPE uses system and networks properties to resolve known and novel patterns. Though the use case was an undirected weighted network, we explain how the layout rules can be adapted to a directed network by assessing source, sink, or transit roles of nodes and using this categorical attribute as a plotting rule. The partitions and scales used permit the organization of nodes given a variety of distributions of quantitative and qualitative node attributes. Moreover, additional attributes can be created to enhance the exploration of system properties. For instance, in weighted networks, a node's average weight can be measured by averaging the weights of its edges and this property can be visually encoded. Moreover, modularity analysis can be used to find subnetworks and module membership can be encoded as a node attribute. Similarly to how each hive plot is compared to resolve attribute associations, subnetworks can be compared by building two different hive panels. Finally, HYPE's versatility in its layout rules and its adaptability to different types of system properties and networks distinguishes it from other network visualizations.

2.4.3 A scalable tool

HYPE was able to resolve patterns such as outliers, trends, similarities, and distributions of the nodes, edges and their attributes. While HYPE was demonstrated on

a network with hundreds of nodes, we argue that this rule-based layout scales to larger networks with thousands of nodes. Though in larger networks there is an increased occurrence of overlap of nodes and edges, the proposed exploration approach based on Shneiderman's mantra circumvents this issue [147]. Namely, the user can start with an overview of the network using a $4x4$ panel, investigate the global trends in the visual signatures provided by 16 hive plots, narrow the number of attributes of interest, and zoom in on the specific hive plots using a $3x3$ or $2x2$ hive panel. In this way, patterns can be resolved at all levels from the whole network topology to subnetworks to individual nodes. In addition, the filtering rules can be used to study specific subnetworks of the network and compare different subnetworks. Therefore, HYPE's interactive features allow the user to navigate large networks to resolve both local and global patterns in the system.

2.5 Future directions and conclusions

Despite the fact that HYPE is based on an intuitive encoding of nodes and edges using circular marks and links, navigating a network through a hive plot or a panel of hive plots requires a certain familiarity with layout rules and overall set up. For this reason, we have made available the Friends network shown (Figure 2.1), and the *C. elegans* panel (Figure 2.5) on our git repository for interested users to interact with. In our experience, we have found that users quickly become accustomed to layout rules and can then take full advantage of the tool's features to investigate patterns of interest. As demonstrated on biological networks with resolved structures, HYPE allows users, which may or may not be experts in the system modelled, to find known and reveal novel topological and data association patterns. In the future, we envision HYPE as a web application in the cloud with the purpose to enhance user experience by creating a community of hive panel builders and to increase the accessibility of different network types of varying complexity. To accomplish this it will be necessary to provide embedded settings in the hive panel output and a log file for generating reproducible visualizations.

Chapter 3

Characterizing robustness and centrality in microbial co-occurrence networks from natural and disturbed soil communities

Microbial communities form distributed networks of genetic and metabolite exchange shaped by horizontal gene transfer and symbiotic interactions [40, 51, 73, 82, 169]. These networks can be reconstructed based on co-occurrence patterns and environmental sequence information [55, 93, 110, 175]. The topological properties of microbial co-occurrence networks including the centrality, connectance, and clustering have the potential to reveal ecological design principles that ultimately drive ecosystem functions. Indeed, network centrality measures have been used to identify important components of systems such as "essential" proteins in protein-protein interaction networks [87], neurons acting as information highways in connectomes [157] and keystone species in food webs [134]. Recently, different centrality measures have been used to identify microbial keystones in marine and soil environments using co-occurrence networks [11, 12, 80, 105, 138, 139, 150, 169].

However, the lack of consistency and methodology in the selection of centrality measures limit the interpretation of the derived results in these networks [16, 54]. Adopting methods from macroecology, we provide a novel way in which to select centrality measures to identify taxa which may play structurally important roles in co-occurrence networks. Our procedure relies on robustness simulations to test network structural integrity and quantify the structural importance of taxa in co-occurrence networks. We demonstrate this approach using clustered SSU rRNA tag sequences sourced from timber harvested forest soils spanning three biogeoclimatic zones within the Long Term Soil Productivity study (LTSP). We show that robustness analysis reflects the impact of disturbance from the structure of the networks inferred from natural and disturbed communities and that the identified central taxa may play a role in community stability and resilience.

3.1 Introduction

Genetic and metabolic exchanges have been well documented in natural and engineered ecosystems [108, 110, 113, 128, 174, 175, 178] and the absence of taxa involved in these exchanges can impact the community's dynamics in varying amounts [46, 128]. In this way, some species play key roles by providing essential nutrients to the community or maintaining the appropriate environmental conditions [46, 108, 113, 128, 174, 178]. In macro-ecology, these functionally important species are denoted "keystone species" [134]. Evidence of keystone species is also found in microbial communities despite being a much more diverse system than food webs [105, 128, 138]. For example, low abundant but highly active sulfate reducers were found to mediate a major and essential biogeochemical process on which the rest of the community relies [128]. Identifying these keystone species is critical to understanding community genetic and metabolic processes integral to ecosystem functions as well as microbial community response to disturbance in a time of global climate change [145]. However, in highly diverse ecosystems such as those inhabiting soils resolving keystone connectivity in relation to microbial community structure and function is a challenging enterprise.

High-throughput technologies such as SSU rRNA sequencing enables the characterization of community membership and bridges the cultivation gap of mi-

croorganisms [67, 141, 156]. Pairing this high resolution community composition data with network inference analysis has captured potential inter-taxa interactions [105, 129, 138, 139, 150]. Co-occurrence networks are a type of network inference model where nodes represent individual taxa and edges represent correlations between taxa. Co-occurrence between two taxa can be interpreted as mutualism, niche overlap, commensalism etc., and a mutual exclusion can be interpreted as amensalism, competition, alternative niche preference etc. [46, 54, 55]. The topological structure of microbial co-occurrence has been related to environmental properties such as seasonal disturbance in lake water communities [89], enterotypes in the human gastrointestinal tract microbiome [10], and the effect of animal feeding activity on soil communities [139]. Co-occurrence networks have also been used to first infer and then to isolate symbiotic microorganisms [46]. These studies suggest that co-occurrence networks combining individual taxonomic composition information and inter-taxa relationships can illuminate ecological design principles organizing microbial community structure and function across ecological scales [10, 54, 89, 105, 150, 169].

In microbial co-occurrence networks studies, different node centrality measures have been used to identify keystone species [16, 105, 138, 139, 150, 169] based on the idea that structurally critical taxa play important ecological roles given that their removal leads to network fragmentation [46, 138, 145]. Several network centrality measures have been applied in robustness analysis: degree centrality, betweenness centrality, eigenvector centrality, closeness centrality, etc (Section 1.1.2). For example, Lupatini and colleagues identified key microbial taxa in soils from natural forests and agricultural plantations using betweenness centrality and closeness centrality [105]. However, the centrality measure used to identify these taxa are inconsistent across studies as is the underlying reasoning used to determine the appropriate centrality measure. For example, one study rationalizes the use of certain centrality measures by choosing those that agree in the way they rank taxa by centrality value [139]. Other studies pick centrality measure arbitrarily [105, 169].

In addition, the experimental design and statistical validation of the network construction methods used in these studies to obtain co-occurrence networks could be improved. For example, two of these studies use fewer than 10 samples to

measure correlations [105, 139] and network inference on such low samples numbers could produce many false positive co-occurrences [16]. Favorable sample compositions such as levels of sample heterogeneity have been proposed by Berry and Widder to maximize the sensitivity and specificity of co-occurrence analysis [16]. Moreover, some these studies do not employ statistical methods to filter false positive co-occurrences [105, 139] despite known sources of error such as compositionality bias. Nonetheless, many softwares have been developed to address compositionality bias and other pitfalls of network construction [53, 54, 63]. Therefore, combining rigorous statistical methods and proper sample compositions can increase the sensitivity and specificity of co-occurrence analysis while creating a standard for future microbial co-occurrence network studies.

Studies in other biological systems have demonstrated that different centrality measures quantify different structural features of node positions relative to the entire network [19, 62, 123]. In addition, the applicability of a centrality measure relies on the network topology and the central characteristics of interest [85]. For example, a centrality measure that captures central characteristics in a local neighbourhood of a node may not be appropriate to compare the centrality of two nodes found in distinct regions of the network. Different methods have been proposed to identify the appropriate centrality measure depending on the global topology of the network and the type of functional role played by central nodes [5, 45, 85]. Iyer and colleagues demonstrated that robustness simulations can identify structurally important nodes by measuring the integrity of the network's structure against the removal of nodes ranked by centrality measures. In this way, the centrality measure that decreases the structural integrity of the network is the measure that identifies the nodes that are required to preserve network structure. It is reasonable to assume that if a network has been sufficiently fragmented by node removal than a functional process of the system modelled by the network will be less effective in the fragmented network [5, 85]. In foodweb studies, where nodes are species and edges are trophic interactions, network robustness simulations and quantitative measure of robustness have been shown to quantify ecosystem stability to species extinction [45]. Adopting robustness analysis methods and applying them to microbial co-occurrence networks has the potential to develop more rigorous selection of centrality measures, assess community stability and identify keystone taxa.

Here, we use network inference and robustness simulations to identify central taxa in microbial communities in timber harvested soil from the Long term soil productivity (LTSP) study. Recent efforts to measure microbial community responses to perturbation in the LTSP sites has resulted in an archive of SSU rRNA samples spanning biogeoclimatic ecozones sampled at different soil horizons and in locations with varying levels of timber harvesting [76]. Therefore, the resolved co-occurrence networks from the LTSP study represent variable topologies influenced by ecozone, soil properties and timber harvesting treatment. Given that soil communities are highly diverse they provide a real world use case to benchmark network robustness analysis in microbial ecology that is extensible to less complex communities.

In the process we ask the following questions:

1. What centrality measures driving robustness simulations are co-occurrence networks least robust to?

2. Are co-occurrence networks consistently more robust to the same centrality measure driving node removal?

3. How do the networks inferred from natural and disturbed communities differ in their topology?

4. How do the networks inferred from natural and disturbed communities differ in their robustness?

5. How do the networks inferred from communities from different biogeoclimatic zones differ in their topology, robustness and central taxa?

We find that, despite differences in community composition between LTSP sites and ecozones, the resolved networks from both natural and disturbed communities have similar topologies and are consistently less robust to the removal of taxa ranked by their betweenness centrality value. Finally, we characterize the identified central taxa to show that the chosen centrality measure captures taxa that could not be identified by their taxonomy or distribution in the soil profile.

3.2 Methods

3.2.1 LTSP sample collection and processing

The LTSP study is a multidisciplinary effort to monitor the impact of forestry practices on North American soil productivity [136]. Our study focuses on three LTSP ecozones in British Colombia, Ontario and California previously described by Hartmann and colleagues [76]. The three British Columbia sites are located in the Sub Boreal Spruce (SBS) biogeoclimatic zone and were harvested 15 years prior to sampling. The three California sites are located in the Mediterranean (MD) biogeoclimatic zone and were harvested 16 years prior to sampling. The three Ontario sites are located in the Jack Pine (JP) biogeoclimatic zone and were harvested 17 years prior to sampling.

At each site, samples were collected at different treatment plots ($40x70m^2$). Three levels of organic matter (OM) removal and one unharvested control. The plots were arranged in a randomized, full-factorial design. The three levels of OM removal were defined as stem-only harvesting (OM1), whole-tree harvesting (OM2) and whole-tree harvesting plus forest floor removal (OM3). These levels correspond to an increasing carbon source removal [77]. Table A.13 provides a summary of the number of samples recovered for each ecozone and each treatment. Table 3.1 describes the biogeoclimatic properties of individual sites within each ecozone including the tree species planted post-harvest. This data was collected from associated LTSP publications [76, 77, 132, 135, 136].

At each plot, samples were collected from the organic soil horizon (top layer of soil) and mineral soil horizon (bottom layer) randomly with 3 to 5 replicates per plot. Given the varying depth of organic horizon (typically between $0-20$cm) from one site to another, dimensionless quantities are used to indicate the horizon sampled: 1 for the organic and 2 for the mineral horizon. In the OM3 plots of the SBS ecozone, the forest floor removed during harvesting had not redeveloped 15 years post-harvesting and thus the organic soil horizon could not be sampled. This study includes a total of 326 samples.

Table 3.1: LTSP sampling sites' soil data for the SBS, MD and JP ecozones

Site	Zone	Province or State	Latitude	Longitude	Elevation (m)	Climatic Zone (life zones)
Wells	JP	Ontario	46.42	-83.37	228	Cool temperate moist
Superior 3	JP	Ontario	47.57	-82.85	426	Boreal moist
Eddy 3	JP	Ontario	46.75	-82.25	488	Moist
Lowell Hill	MD	California	39.26	-120.78	1270	Warm temperate dry
Blodgett	MD	California	38.88	-120.64	1320	Warm temperate dry
Brandy City	MD	California	39.55	-121.04	1130	Warm temperate dry
Log Lake	SBS	British Columbia	38.88	122.61	780	Wet cool
Topley	SBS	British Columbia	52.32	126.31	1100	Moist cold
Skulow Lake	SBS	British Columbia	52.32	121.92	1050	Dry warm

Site	Climatic Zone (Köppen classification)	Forest type	Tree Species
Wells	Humid Continental warm summer	Mixed pine	Jack pine, Black spruce, Red pine
Superior 3	Humid Continental warm summer	Jack pine	Jack pine, Black spruce
Eddy 3	Humid Continental warm summer	Mixed conifer	Jack pine, Balsam fir, White birch
Lowell Hill	Mediterranean hot summer	Mixed conifer	Ponderosa pine, Sugar pine, White fir, Giant sequoia
Blodgett	Mediterranean hot summer	Mixed conifer	Ponderosa pine, Sugar pine, White fir, Giant sequoia
Brandy City	Mediterranean hot summer	Mixed conifer	Ponderosa pine, Sugar pine, White fir, Giant sequoia
Log Lake	Boreal cool summer	Sub-boreal spruce	Subalpine fir, Douglas fir, Interior spruce
Topley	Boreal cool summer	Sub-boreal spruce	Lodgepole pine, Subalpine fir, Interior spruce
Skulow Lake	Boreal cool summer	Sub-boreal spruce	Lodgepole pine, Interior spruce

Site	Soil parent	Principal Soil Classification	Year established	Year sampled
Wells	Glacial outwash	Orthic Humo-Ferric Podzol	1993-1994	2011
Superior 3	Glacial outwash	Orthic Dystric Brunisol	1993-1994	2011
Eddy 3	Glacial outwash	NA	1993-1994	2011
Lowell Hill	Volcanic mudflow	Mesic Ultic Haploxeralfs	1995	2011
Blodgett	Volcanic mudflow	Mesic Ultic Haploxeralfs	1995	2011
Brandy City	Volcanic mudflow	Mesic Ultic Haploxeralfs	1995	2011
Log Lake	Glacial till	Orthic Humo-Ferric Podzol	1994	2008
Topley	Glacial till	Orthic Gray Luvisol, Gleyed Gray Luvisol	1994	2008
Skulow Lake	Glacial till	Orthic Gray Luvisol	1994	2009

3.2.2 Environmental DNA extraction and sequencing

The hypervariable region V1 to V3 of the bacterial (SSU rRNA) gene, PCR amplified from 50*ng* soil DNA and sequenced using the 454 platform as previously described by Hartmann and colleagues [77]. The resolved reads were processed as previously described by Hartmann and colleagues [77]. In brief, reads were filtered using MOTHUR [144]: reads with ambiguous base calls and average quality scores < 25 were eliminated. Sequences were clustered into operational taxonomic units (OTUs) at 97% sequence identity. Singletons, clusters with only one represented sequences were not included in the analysis. The number of sequences per sample recovered after quality control is summarized in Tables A.1 to A.12.

3.2.3 Microbial co-occurrence network inference

Samples were grouped by treatment and by ecozone to produce twelve microbial co-occurrence networks using the software CoNet (version 3.0) implemented in Cytoscape (version 3.1.0) [53, 55] (Table A.13). First, samples' composition data was combined to produce a matrix of OTU read counts per network. Read counts were filtered so that only OTUs occurring in 25% percent of samples (in sample grouping per network) were kept and normalized by total reads per sample. Pairwise correlations were calculate for each pair of OTUs using two different correlation measures: Spearman correlation coefficient and Bray Curtis dissimilarity. Pairwise correlations with an absolute value of 0.6 for Spearman and within the thresholds of 0.4 and 0.6 for Bray Curtis was used to reduce the number of correlations to be evaluated.

Once the initial network is constructed, different procedures are implemented to refine the network. To avoid compositionality bias, the network co-occurrences are recomputed for 1000 permutations: for each evaluated co-occurrence, taxonomic abundance profiles are shuffled and the abundance matrix is renormalized. Then the network is recomputed for 1000 bootstrapped matrices: the original matrix is sub-sampled with replacement and all correlation measures recomputed. This procedure provides a confidence interval around the co-occurrence score which is used to remove all co-occurrences not within the limits of the 95% confidence interval. From the bootstrap distribution and after applying a multiple-

test correction, a p-value is calculated per co-occurrence, per measure, and co-occurrences with a p-value of less than 0.05 are removed. Further details on the different statistical validations of each network construction step are available through the documentation of the CoNet software [53, 55].

The co-occurrence analysis and sampling composition of each network follows the recommendation provided by Berry and Widder [16]:

- high resolution of community composition was used and infrequent taxa were removed
- sample heterogeneity was minimized: samples were grouped by ecozone defined by biogeoclimatic conditions
- compositionality bias due to relative abundance data was accounted for and corrected
- Bray Curtis dissimilarity, which is robust to spurious correlations from presence-absence count data, was used
- several correlation coefficients were measured to increase the sensitivity of the inferred networks

Twelve LTSP networks are thus produced where nodes are OTUs and edges are positive co-occurrences and mutual exclusions (Table A.13). These networks and the collection of samples used to produce each of them are referred to using the name *Ecozone-OMX* in the rest of this analysis.

3.2.4 Ecological analysis

In order to be consistent, the composition data used to assess ecological diversity and clustering patterns is the same data used to compute the networks. Hierarchical cluster analysis of samples was conducted with the *R* package *pvclust* using the Bray-Curtis dissimilarity metric [152]. All clustering was conducted with a bootstrapping procedure of 100 permutations. Community diversity was measured using richness and Shannon's entropy (see Table Table 1.1).

3.2.5 Network analysis

Unless otherwise specified, the network and LTSP sample analysis was conducted using a collection of scripts in Python which are publicly available at https://github.com/hallamlab/network-robustness. A summary of the Python packages used is available on the github repository's front page.

Degree distribution fitting was conducted following the procedure outlined in [35, 36] and using the associated Python package *powerlaw* [7]. Modularity analysis of networks was conducted using our own implementation of the algorithm FAG-EX [103] in Python. The minimum proportion of in-degrees to out-degrees for a subgraph to be a module is suggested to be in the range $(1,3]$ [103, 120]. A higher value of this modularity factor corresponds to a stricter definition of module and thus modularity [103, 120]. Modularity analysis was conducted on all positive co-occurrences (not mutual exclusions) in the largest connected component (LCC) of each network with a factor of 2.

3.2.6 Using HyPE to visualize networks

Hive panel Explorer (HYPE) is a network visualization tool which presents and enables the exploration of complex networks in a data driven manner (Chapter 2). In order to construct informative hive panels and compare the topology of the LTSP networks, ecological and network measures were calculated which represent individual OTUs. Average abundances were computed per network by normalizing read counts per total sample counts. OTU' soil horizon was computed by weighing the sample horizon by the abundance of the OTU in that sample. Resulting values between $1-2$ correspond to the organic and mineral horizon, respectively. Networks measures were computed using the Python *networkx* package: degree, betweenness centrality, clustering coefficient.

The following six parameters were chosen as layout rules: average soil horizon, abundance, degree, centrality, clustering coefficient and phylum. The average soil horizon of a node reflects where it is predominantly located within the soil. Given the known stratification between organic and mineral horizons we therefore choose average soil horizon as an axis position rule and node degree as an axis assignment rule. To visualize interactions between phyla we use an OTU's phylum to rank

and position the nodes along axes. In order to assess the possible associations between the clustering coefficient of nodes and their average abundance, we set these properties as axis positioning and axis assignment rules, respectively. Finally the betweenness centrality is chosen as the third axis assignment rule. While several other centrality measures could have been used, previous studies of microbial co-occurrence networks and other biological networks have found this measure to be more informative that more local centrality measures such as closeness centrality [85].

Given the exponential type degree distribution of these networks (evaluated in Section 3.3), degree and betweenness centrality were plotted using a logarithmic partitioning scheme. Average soil horizon and clustering coefficient was plotted using a linear scale while abundance was plotted using an even partitioning scheme.

3.2.7 Network robustness simulations

Network robustness simulations can be conducted by removing nodes using different rankings [85]. The robustness at each removal step can be measured by assessing the change in different network properties, including the number of nodes disconnected and the diameter of the network [5, 85]. Here, network robustness simulations were conducted on the LCC of each network (the largest subgraph in which all nodes are connected by some path) by measuring the relative size of the LCC at each node removal step. Nodes were ranked randomly or by different network centrality measures: degree, betweenness centrality, eigenvector centrality and closeness centrality.

In order to obtain a quantitative measure of resilience to different node removals for each network, a robustness factor R is calculated:

$$R = r/|N_{lcc}| \tag{3.1}$$

where r is the number of nodes removed in the network such that the size of the LCC has decreased by 50% and $|N_{lcc}|$ is the total number of nodes in the LCC [45]. Dunn and colleagues proposed this robustness factor to measure the resilience of food webs to species lost by assessing how many extinctions leads the 50% of possible ecosystem extinctions [45]. By adopting this factor from this

macro-ecology study, we are assuming that fragmenting the network such that less than 50% of the OTUs take part in the original community will drastically affect the functional processes that rely on core community structure. R has a maximum value of 0.5 and a minimum value of $1/|N_{lcc}|$. This robustness factor is normalized and can thus be compared between networks with LCCs of different sizes.

3.3 Results

3.3.1 Ecological diversity within and between ecozones

We begin our analysis by assessing the ecological similarities and differences within and between ecozones. We expect sample compositions to differ based on biogeoclimatic conditions (Table 3.1). Figure 3.1 shows a hierarchical clustering analysis of all samples coloured by ecozone and demonstrates that dendrogram clusters distinguish ecozone effectively. The same hierarchical clustering dendrogram is shown in Figure 3.2 with leaves coloured by treatment to demonstrate that individual samples do not cluster by level of OM removal.

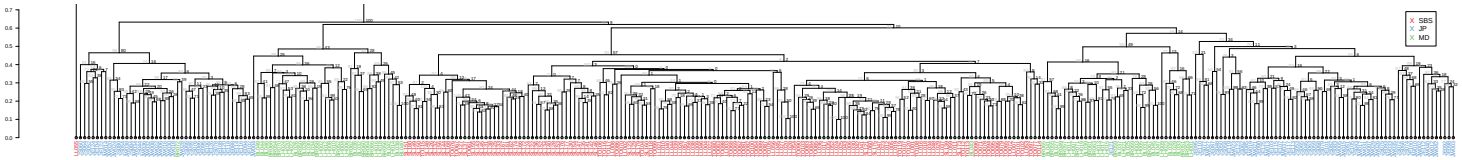

Figure 3.1: Hierarchical clustering of all samples coloured by ecozone

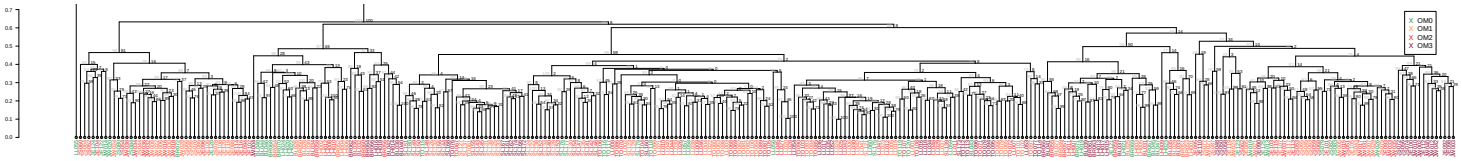

Figure 3.2: Hierarchical clustering of all samples coloured by OM treatment level

Table 3.2: Richness of LTSP samples grouped by ecozone and treatment

	OM0	OM1	OM2	OM3
SBS	370	536	489	615
MD	1337	1178	1375	1418
JP	1135	1014	936	1313

Table 3.3: Shannon's entropy of LTSP samples grouped by ecozone and treatment

	OM0	OM1	OM2	OM3
SBS	4.1	4.3	4.3	4.6
MD	5.8	5.7	6.0	6.1
JP	5.6	5.5	5.5	5.7

Within ecozones, hierarchical clustering analysis indicates that samples cluster by soil horizon (Figures A.1-A.9), except for samples from SBS which first cluster by the three sampling sites (Figures A.1-A.9).

The diversity of each sample group is quantified using species richness and Shannon's entropy and is summarized in Table 3.2 and A.14. Richness and diversity varies between ecozones more than within: the richness on the SBS communities is an order of magnitude below that of MD and JP. We also find an increased entropy in these two ecozones indicating a more heterogeneous composition than the communities from the SBS ecozone.

Overall, communities that have undergone OM3 treatments have a relatively higher richness and greater Shannon's entropy. These findings quantify the compositionality differences between ecozones and may reflect the fact that treatment effects vary under different biogeoclimatic conditions[77, 132].

3.3.2 Global network topology

Having found several factors driving ecological differences between sample groups, we begin our investigation of community structure by evaluating the differences and similarities in the global topology of our networks. The number of nodes in each network is of the same order of magnitude as their richness (Table 3.4). Table 3.5 and 3.6 summarize each network's average clustering coefficient and the

Table 3.4: The number of nodes $|N|$ and edges $|E|$ in the LTSP networks

Ecozone	OM0		OM1		OM2		OM3	
	$\|N\|$	$\|E\|$	$\|N\|$	$\|E\|$	$\|N\|$	$\|E\|$	$\|N\|$	$\|E\|$
SBS	278	1469	285	2253	270	1528	114	166
MD	1299	85832	1118	70489	1316	76511	1395	173591
JP	1057	52363	788	67035	762	51441	68	39

Table 3.5: Global clustering coefficient of the LTSP networks

Ecozone	OM0	OM1	OM2	OM3
SBS	0.335	0.462	0.429	0.153
MD	0.466	0.484	0.447	0.553
JP	0.448	0.618	0.581	0.044

size of LCC in terms of number of nodes and diameter (longest shortest path) (see Section 1.1.2 for a review of network measures). Overall we find a highly clustered topology (Table 3.5), as previously found in other microbial co-occurrence networks [11, 105, 139, 150]. It's important to note here that triangles, groups of three connected nodes, can only be achieved between three co-occurring OTUs since two mutually exclusive OTUs cannot by definition co-occur with a third OTUs.

Therefore the high clustering coefficient relates the connectivity of co-occurring OTUs. We also note that the number of nodes and edges in the LCC is of the same magnitude as in the entire network for all except *SBS-OM3* and *JP-OM3*, suggesting that the clustered topology is found on a global scale, instead of a local scale in which case there would be several smaller components (disconnected subgraphs).

As explained in Section 1.1.3, many biological networks have a power law de-

Table 3.6: Size of the largest connected component of the LTSP networks: $|N_{lcc}|$ and D correspond to the number of nodes in the LCC and its diameter, respectively.

	OM0		OM1		OM2		OM3	
	$\|N_{lcc}\|$	D	$\|N_{lcc}\|$	D	$\|N_{lcc}\|$	D	$\|N_{lcc}\|$	D
SBS	249	10	285	8	268	8	78	11
MD	1295	9	1108	9	1285	10	1391	10
JP	1023	10	778	7	744	9	3	2

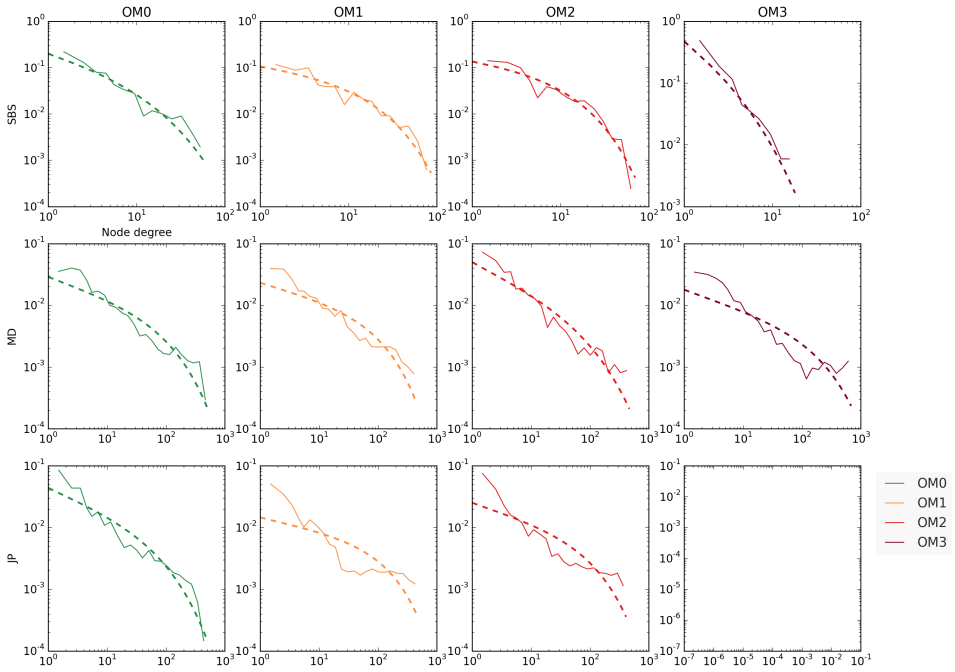

Figure 3.3: Probability distribution function of node degree for all LTSP networks with stretched exponential fitting. The *JP-OM3* network was omitted given its lack of structure.

gree distribution (Figure 3.3). We fitted the degree distributions of our networks to a power law using the procedure described in Clauset and colleagues and compared the fit to other distributions [35, 36]. We found that all twelve networks follow a stretched exponential distribution rather than a power law, exponential or lognormal degree distribution (Figure 3.3). This type of distribution is also known as a power law with exponential tail and was found in marine microbial time-dependent co-occurrence networks [150]. These kinds of distributions are not scale free and typically contain numerous high degree nodes [118]. In addition, the diameter of networks with such distributions scale sub-linearly with increasing network size [118] which explains why the networks have similar diameter (Table 3.6) despite having up to an order of magnitude difference in the number of nodes and edges.

Next, we evaluated the subglobal structure of communities by conducting a modularity analysis on positive co-occurrences and find highly connected clusters

of co-occurring OTUs. We find modules in most networks that are stratified between organic and mineral horizons as illustrated by the hive panel in Figure 3.5.

3.3.3 Visualizing microbial co-occurrence networks with HyPE

We next used HYPE to visualize the modularity of the twelve networks to explore topological and ecological associations. Hive panels of size $3x3$ were constructed for each network by using the six following parameters as layout rules: average soil horizon, abundance, degree, centrality, clustering coefficient and phylum. We provide a highlight of the different visually resolved patterns (as of yet quantitatively validated) illustrated by the hive panel of *SBS-OM0* in Figure 3.4:

- co-occurrences stratify by horizon and mutual exclusions connect OTUs with different average soil horizons
- high degree nodes have average soil horizons near the organic layer or the mineral layer but not in between
- there are many co-occurrences between OTUs of different abundances and from different phyla
- OTUs with lower abundances seem to have higher clustering coefficients
- the organic horizon module contains OTUs with a wider range of average soil horizon than mineral horizon modules
- co-occurrences between the two modules seem to be primarily between low and high degree nodes.

Given the focus on this chapter, we focus on two patterns in particular which are shown across all networks in Figure (3.5) and (3.6). Figure 3.5 illustrates the twelve networks' modular connectivity and Figure 3.6 shows the connectivity and centrality of OTUs categorized by phyla. Despite not finding two modules corresponding to the organic horizon and mineral horizon in all twelve networks, we do see a similar connectivity pattern: most co-occurrences occur within horizons and few co-occurrences are found between OTUs horizons.

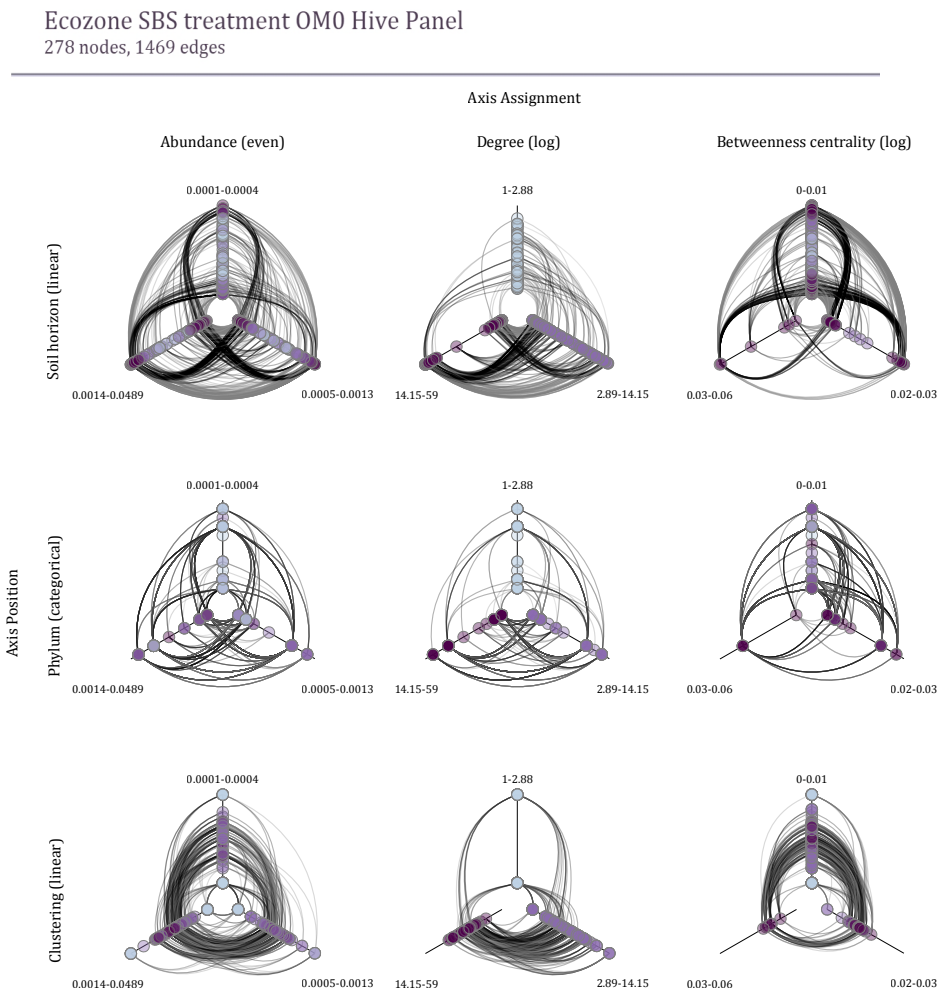

Figure 3.4: Hive Panel of the network from the SBS ecozone with treatment OM0. Nodes are coloured by degree.

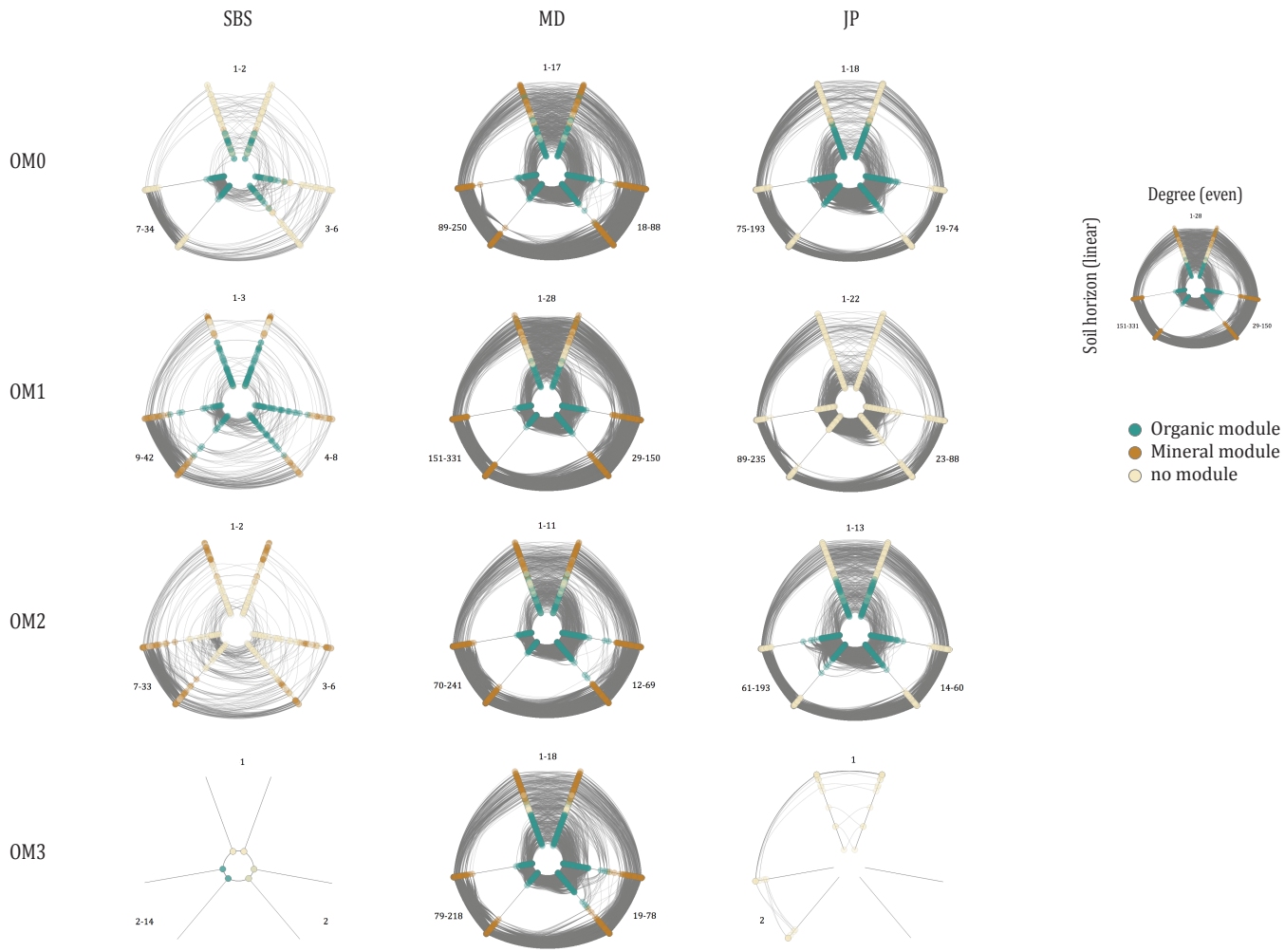

Figure 3.5: Hive panel of twelve hive plots showing the horizon modularity of the LTSP networks. Hive plots were constructed by partitioning node degrees logarithmically onto axes and linearly positioning the nodes by average soil horizon (the organic horizon to hive plot centers)

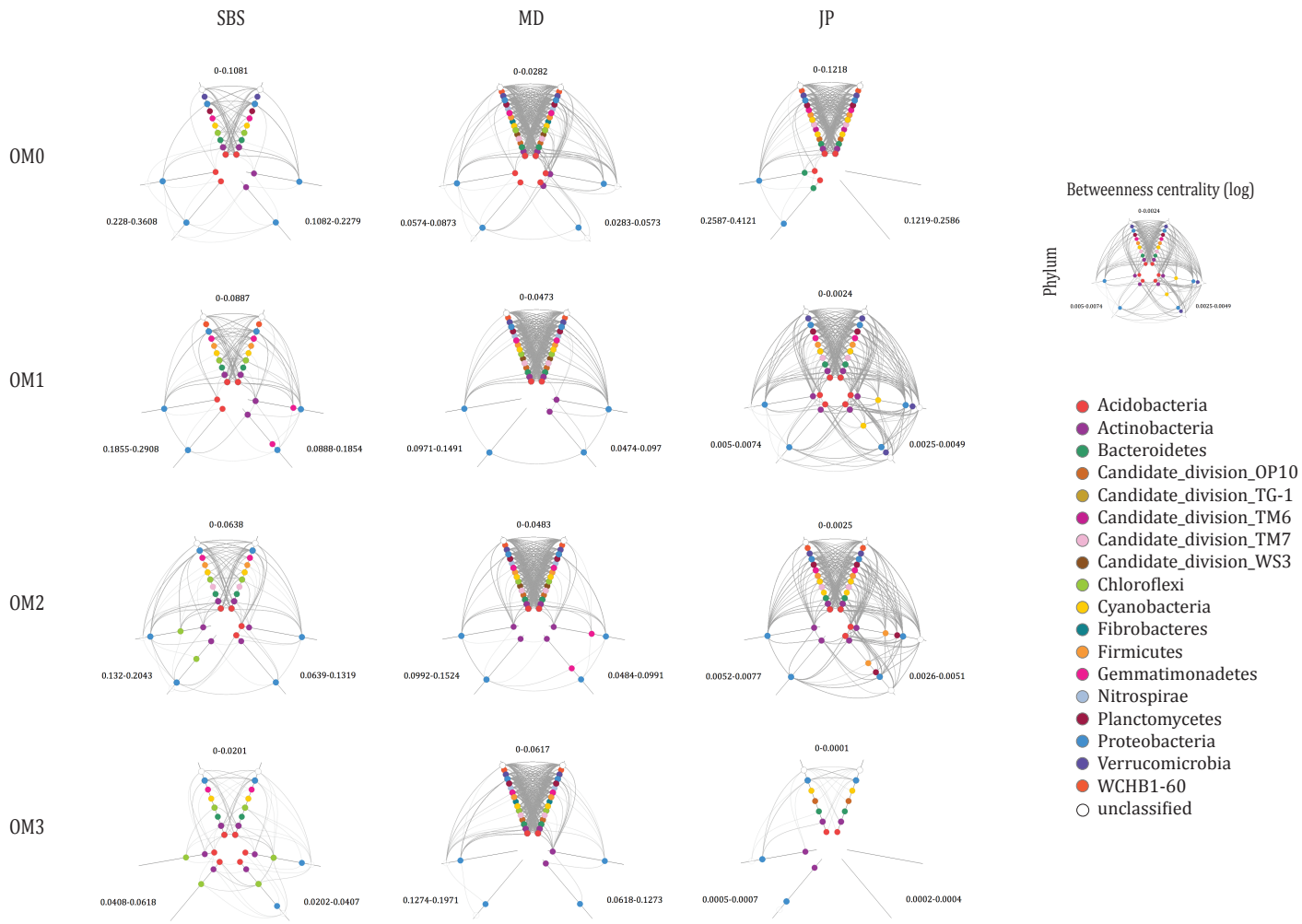

Figure 3.6: Hive panel of twelve hive plots showing the connectivity and centrality of OTUs' phyla of the LTSP networks. Hive plots were constructed by partitioning node betweenness centrality values logarithmically onto axes and linearly positioning the nodes by the alphabetical rank of their phylum

Figure 3.6 illustrates that OTUs with high betweenness centrality tend to come from the three most predominant phyla in this samples collection, Proteobacteria, Acidobacteria and Actinobacteria. However, other phyla also have high centrality OTUs in specific networks including Bacteriodetes, Chloroflexi, Gemmatimonadetes, Planctomycetes, Firmicutes, and Verrucomicrobia. Given that different centrality measures evaluate different features of node positions and that their applicability depends on the type of network and the task at hand, we use robustness simulations to determine the appropriate centrality measure given the topology of the twelve networks.

3.3.4 Network robustness simulations

Evaluating centrality driven robustness of networks

Robustness analysis tests the integrity of the network's structure to different types of node failures. Conducting different simulations by removing nodes ranked by their centrality value can identify the nodes playing key structural roles in the network. Figure 3.8 shows the robustness simulations on all twelve networks where nodes were removed either randomly and by ranked values of degree, closeness centrality, betweenness centrality and eigenvalue centrality. We notice that ranking nodes by betweenness centrality consistently fragments the LCC earlier in the simulations. In addition, many simulations show a sharp drop in the relative size of the LCC: the removal of certain nodes disconnects large subgraphs within the LCC. In particular, we notice that this drop occurs most precipitously in *SBS-OM0* and *JP-OM0* networks. We then measured the robustness factor for each node removal method and find that overall these networks are least robust to the removal of nodes ranked by betweenness centrality (Table 3.7, 3.8 and 3.9). Other centralities vary in their effect on robustness and often produce similar robustness factors as does the random removal of nodes. Comparing robustness to betweenness centrality node removal, we observe that the robustness of networks differs most across ecozones than within: SBS networks are on average less robust than networks from the MD and JP ecozones. Moreover *SBS-OM0* and *JP-OM0* are much less robust than the treatment networks from the same ecozone. This pattern suggests that treatments

effects on natural and disturbed communities are reflected in the co-occurrence network structure. However, MD networks vary very little in their robustness factors, as illustrated by the simulations driven by betweenness centrality.

Comparing centrality measures

Figures 3.7, A.10 and A.11 show the relations between each centrality measure and demonstrate that betweenness centrality captures different OTUs than the other centrality measures. In particular we notice that a high degree, closeness centrality or eigenvector centrality does not ensure a high betweenness centrality value. This trend indicates that, in the case of comparing degree and betweenness centrality, nodes with few co-occurrences can connect paths of multiple co-occurrences (i.e. a chain of co-occurrences). In the absence of these high betweenness centrality taxa, these paths would we be longer or nonexistent in its absence. Only two centrality measures, degree and eigenvector centrality, seem to have a linear relationship. We also notice that many nodes tend to have high closeness centrality values, which does not facilitate the selection of highly central OTUs. This trend is expected as a node's closeness centrality is hierarchically calculated from the closeness centrality of other nodes [118].

3.3.5 Characterizing central taxa

Having determined that betweenness centrality (BC) captures certain structural positions related to network robustness, we continue our investigation on the OTUs with highest BC values. A BC value is not as informative as its rank [85]; therefore we choose a percentile cut off to capture central OTUs. We select the OTUs with the top 10% percentile of BC values in each network, in combination with a cut off of 0.005. The highest and lowest BC values are 0.41 and 0.007, respectively. These values express that the corresponding nodes take part in $41\% - 07\%$ of all shortest paths in the network. To put this in perspective, a network with $|N|$ nodes has in the order of $|N|^2$ shortest paths. In this way, we collect from all networks despite their different sizes and capture a total of 265 central OTUs, which represents 8% of the total number of OTUs that co-occur in the networks.

We compare central OTUs to other network members by evaluating their aver-

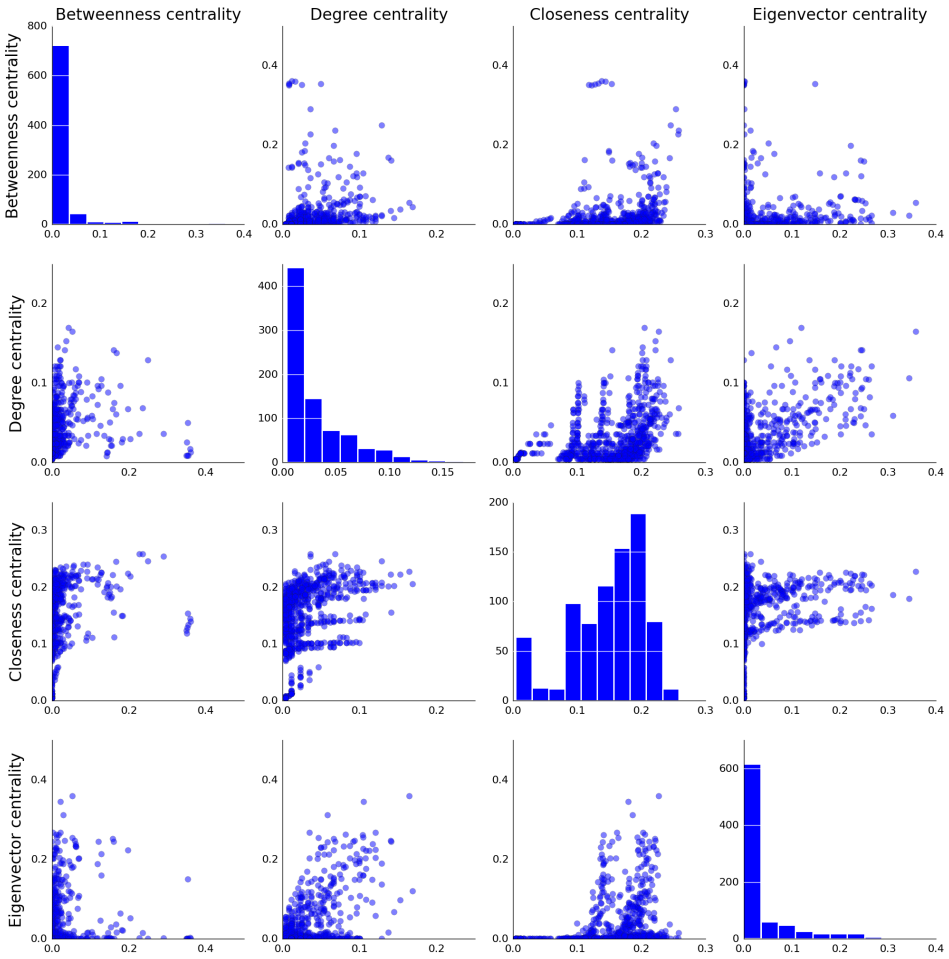

Figure 3.7: Scatter matrix plot of four centrality measures in the SBS networks. Histograms of each centrality measure is also shown. The different centrality values of OTUs for each treatment network was pooled to produce these plots.

Table 3.7: Robustness factor of SBS networks per node removal method

	Random	Betweenness centrality	Degree centrality	Closeness centrality	Eigenvector centrality
OM0	0.37	0.07	0.24	0.07	0.09
OM1	0.42	0.17	0.19	0.19	0.38
OM2	0.39	0.10	0.17	0.25	0.33
OM3	0.29	0.12	0.12	0.12	0.2

Table 3.8: Robustness factor of MD networks per node removal method

	Random	Betweenness centrality	Degree centrality	Closeness centrality	Eigenvector centrality
OM0	0.48	0.17	0.48	0.46	0.43
OM1	0.47	0.17	0.49	0.45	0.49
OM2	0.47	0.14	0.48	0.46	0.49
OM3	0.48	0.13	0.49	0.46	0.49

Table 3.9: Robustness factor of JP networks per node removal method

	Random	Betweenness centrality	Degree centrality	Closeness centrality	Eigenvector centrality
OM0	0.42	0.16	0.47	0.13	0.45
OM1	0.47	0.39	0.40	0.41	0.42
OM2	0.47	0.41	0.46	0.44	0.46
OM3	NA	NA	NA	NA	NA

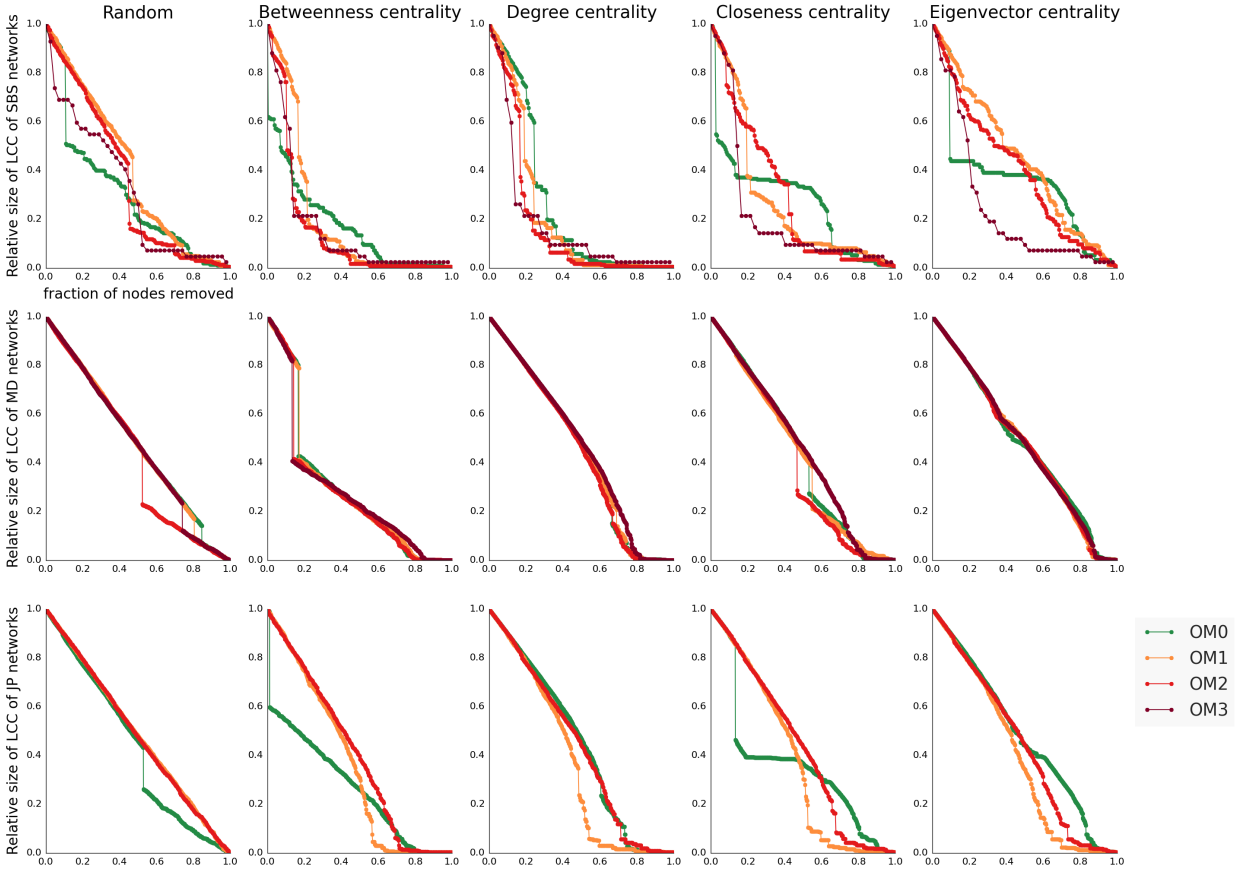

Figure 3.8: Robustness simulations of twelve LTSP networks driven by different centrality measures. The relative size of LCC of each treatment network is plotted against the number of nodes removed. Networks are coloured by associated treatment level.

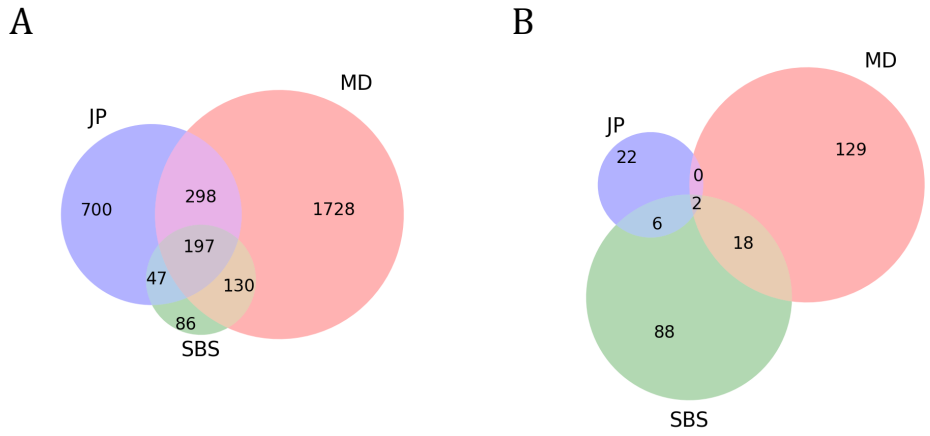

Figure 3.9: Venn diagram of all OTUs in ecozone networks (A) and all central OTUs (B).

age soil horizon and abundance. Figure 3.11 shows that central OTUs have a wide range of average soil horizons that are primarily not exclusive to either organic nor mineral horizons. In terms of their abundance, most central OTUs are rare ($< 0.1\%$) or of intermediate ($> 0.1\%$ and ($> 0.1\%$ and $< 1\%$) abundance [170], with a few exceptions (Figure 3.12).

Next we compare the taxonomic distribution of central and non-central OTUs. Figure 3.10 illustrates the overlap in taxonomic representations in each network at the phylum, order and class level: the overlap in taxonomies found in all ecozone groupings of networks decreases when comparing central OTUs. Taxonomic overlap at lower taxonomic levels was not evaluated as the number of unclassified OTUs at those levels drops from 10% to $20\% - 80\%$. Using counts of OTUs per taxonomic level per network, we evaluate the possible over-representation of taxonomies in central OTUs given the null hypothesis that central OTUs were randomly selected: no individual phylum, order nor class was over-represented in the central OTUs at a significance below $p = 0.05$ (Table A.15-A.22). Taxonomic representation was modelled using a hypergeometric distribution of taxonomic counts and over-representation p-values were produced using a Bonferonni correction.

Looking at the overlap in central OTUs between networks grouped by ecozone, we find that few OTUs are central in ecozones (0.7%), despite the fact that many

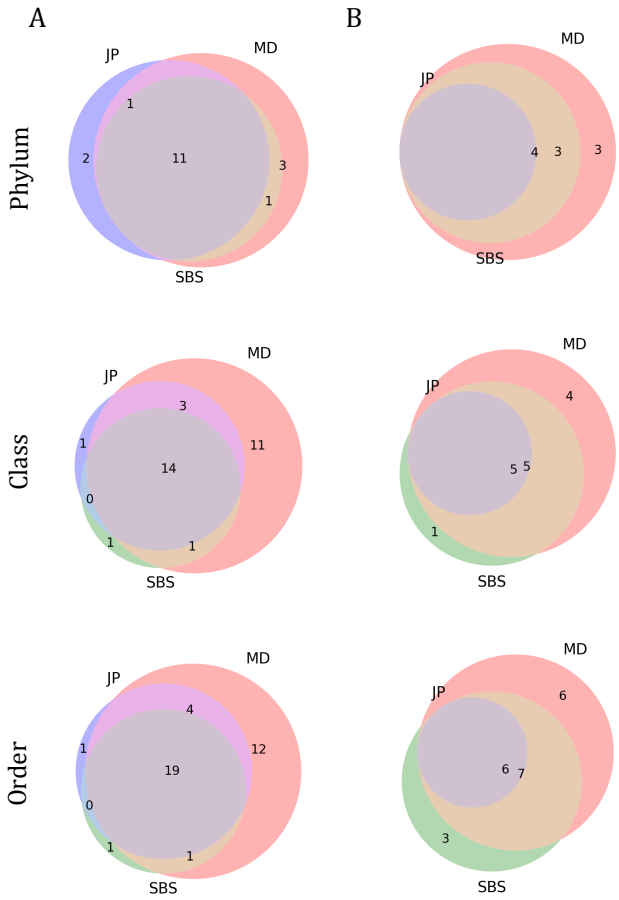

Figure 3.10: Venn diagram of the number of phylum, class and order shared across ecozone networks (A) and the number of central taxonomic levels shared (B).

OTUs are found in all ecozone groupings of networks (7%) (Figure 3.9). Given that functional relations can be associated to bacterial lineages [58], this pattern suggests that the role played by central OTUs is fulfilled at a higher taxonomic level instead of a species level.

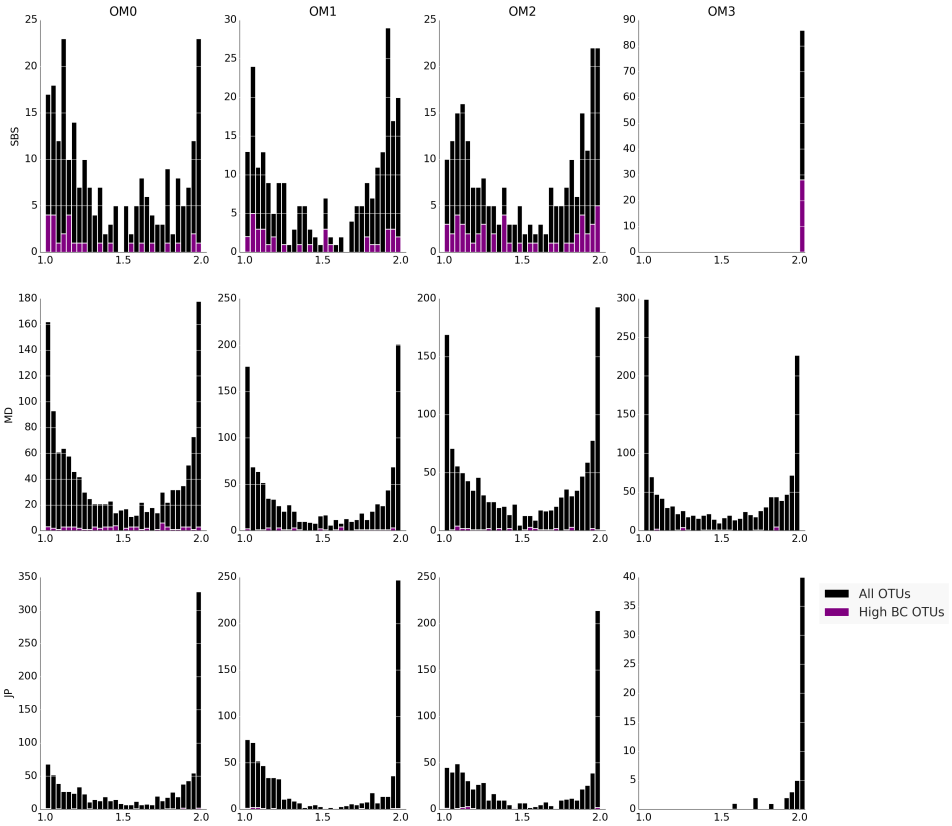

Figure 3.11: Histograms of the average soil horizon of OTUs with high BC values

3.4 Discussion

Previous studies indicate that community diversity and composition varies by biogeoclimatic conditions in LTSP sampling sites and soil horizons [76, 77]. Among these drivers, soil horizon consistently split samples in all ecozones and OM removal treatments. Despite these differences, the global topology, modularity and outcome of robustness simulations remained similar across all twelve LTSP networks. These results demonstrates that consistent ecological patterns can be resolved through network analysis despite the variability in community composition in forest soils.

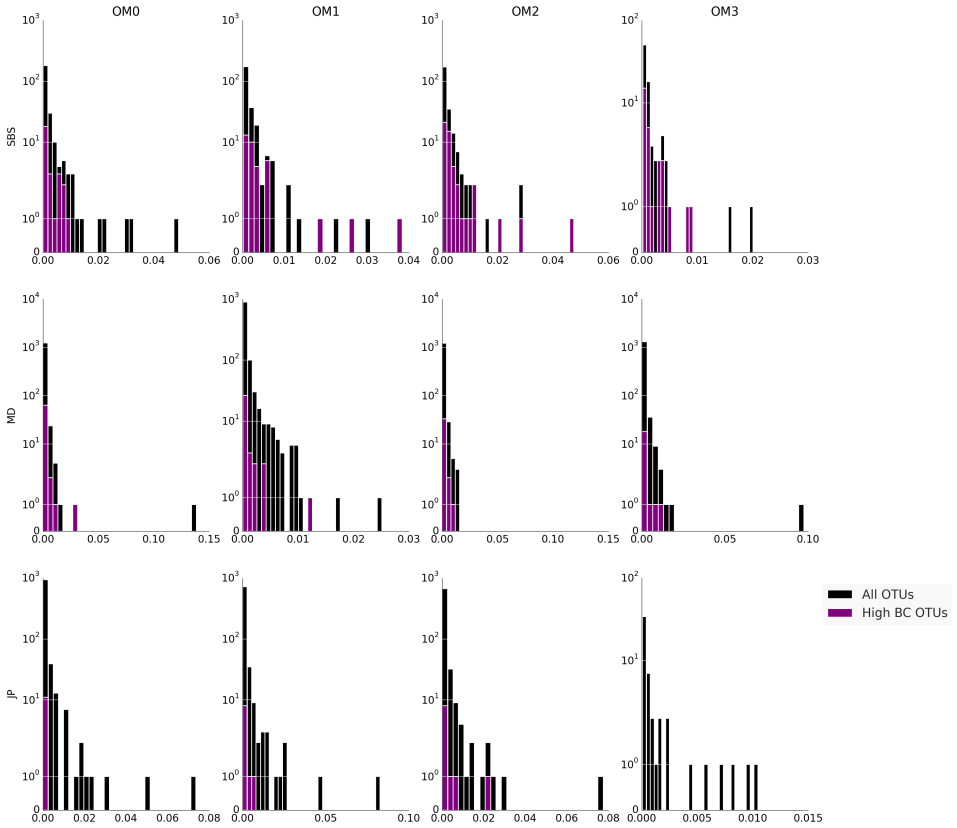

Figure 3.12: Histograms of the abundance of OTUs with high BC values

3.4.1 Soil microbial co-occurrence networks: a complex ecologically driven structure

Evaluating the global topology of each network, we find similar properties as other co-occurrence network studies: an exponential type degree distribution [150], a highly clustered [89, 105, 129, 139, 150, 178], connected [105, 129, 139, 150, 178], and modular structure [150]. Though most biological networks can be fitted to a power law distribution, other real world networks have stretched exponential degree distributions: science collaboration networks [117], certain foodwebs [122] and power grids [9]. The structural similarity between the LTSP co-occurrence networks and social, biological and technological networks suggests that the mode of network inference used in this study captures non-random relationships.

To illustrate how these topological properties relate to ecological properties, we visualized these patterns using HYPE, a data-driven and versatile network visualization tool that overcomes the difficulties in visualizing large networks. The hive panels indicated that centrality measures may be used to capture structurally important taxa in co-occurrence networks. Two of these patterns were presented in all networks and hive panels were constructed to demonstrate the stratification of co-occurrences between horizons and the possible association between taxonomy and centrality with and between LTSP sites.

3.4.2 Centrality and robustness across biogeoclimatic networks

Given the similarity in topological structures between networks inferred from natural and disturbed communities, we expected similar outcomes in robustness simulations between networks. Indeed, simulations confirmed that our networks were the least robust to node removal ranked by the same centrality measure, betweenness centrality. Certain networks were fragmented after the removal of only 10% of nodes with highest BC values, as reflected by their robustness factor. To further confirm that high BC values selects different structural positions than other centrality measures, we compared the centrality of taxa according to different centrality measures. High BC values were not consistent with high values of any other centrality measures. To put this in perspective, correlations between centrality measures have been recorded in randomly generated power law and exponential networks [85]. The lack of correlation observed here suggests that these co-occurrence networks have a more complex structure than that of randomly generated networks with similar degree distributions. This finding confirms that the mode of network inference used captures non-random relationships between taxa.

Overall taxa with high BC values were not distributed like other taxa. Their average relative abundance demonstrated that these taxa have rare or intermediate abundance, with a few exceptions. This trend agrees with experimental findings of low abundance keystone taxa in oral biofilms [46] and in fermentative mixed cultures [138]. For instance, Duran-Pinedo and colleagues showed that the addition of a rare taxa to a culture permitted the isolation of a previously uncultured microorganism [46]. Looking at the taxonomic distribution of central taxa, we found

that they originate from rare and abundant phyla, classes and orders. However, no taxonomy at the phylum, class or order taxonomic level was over-represented in the set of central taxa (Table A.15-A.22). The ecological attributes and topological positions of central taxa suggests that BC selects OTUs which are not artifacts of network construction and that could not have been resolved through ecological measurements: central taxa are predominantly rare or intermediate abundant, ambivalent in their average horizon, and have mixed centrality values according to other centrality measures.

3.4.3 Relating treatment effects to robustness analysis

Having confirmed that networks from both natural and disturbed communities were the least robust to BC driven node removal, we compared robustness factors within ecozones and between treatments. We found that SBS and JP networks' robustness factors decreased between untreated (OM0) and treatment networks whereas MD network robustness factors did not vary much between treatments (Table 3.7, 3.8 and 3.9). Looking at the classification of soils from different sampling sites, we notice that SBS and JP samples have glacial soil parents whereas MD soils originate from volcanic mudflow (Table 3.1). Given that short term (10 years) timber harvesting effects on forest productivity in LTSP sites depended on the susceptibility of different soil types [136], it is not surprising to find different treatment effects associated with robustness simulations between these ecozones.

Surprisingly, we find a counter-intuitive relation between robustness and treatment in SBS and JP networks. It is unclear why treatment networks from SBS and JP ecozones exhibited an increased robustness given the evidence of significant treatment effects in both ecozones. Specifically, ecological assessments of soils in the JP ecozone showed a significant disturbance in environmental conditions related to forest productivity [60, 132] and the impact of OM treatment was evident in changes in microbial community composition in soils from the SBS ecozone [76, 77]. The increase in robustness in JP and SBS OM1, OM2 and OM3 (for SBS only) networks compared to the controls (OM0) therefore demonstrates a shift in topology that could reflect a change in community structure. This change may echo either community instability, community resilience, or the achievement of an

alternate stable community structure [145]. Given the expected long-term effect of organic removal on soil conditions and microbiome [76, 77, 132, 135], a follow-up study of microbial community structure in the next decades could resolve wither the apparent shifts in co-occurrence topology capture a state of community adaptation or fragmentation. Moreover, multi-omics studies could help elucidate specific changes in community metabolic potential resulting from changes in microbial interactions [30].

We now turn to several LTSP studies which have measured the impact of organic matter removal on forests to infer why this robustness pattern was not found in the MD ecozone networks and evaluate if the results from our robustness analysis matches ecological findings from prior studies. The effect of organic matter removal has been evaluated based on several criteria including pre- and post-harvest biomass measurements (volume of organic matter per area) [94, 135, 136], soil carbon and nitrogen concentration [135, 136, 154], microbial biomass [26], carbon utilization [26], tree survival [60], tree growth [60], soil bulk density [121], microbial diversity and shifts in community composition [76, 77]. These studies confirm that MD forests, soil conditions and microbial communities were less impacted by organic matter removal than SBS and JP. First, Fleming and colleagues showed that despite similar responses in tree survival in five ecozones including the ones studied here, tree growth severely decreased in SBS conifers and JP black spruce and jack pine while MD giant sequoias had an increase in growth [60]. Second, statistical evaluation of treatment effect was measured on total biomass measurements and was found to be significant in JP but not MD sites [132]. Third, studies that quantified shifts in microbial community composition from these ecozones found significant perturbations in community structure and taxonomic composition in SBS communities using SSU rRNA sequencing [76, 77]. In contrast, measurements of microbial biomass, respiration and carbon utilization did not resolve any treatment effects in communities from the MD ecozone [26]. These results support the fact that robustness analysis of co-occurrence networks reflect ecological findings. Therefore the association between robustness and organic matter removal impact demonstrates the sensitivity of co-occurrence relationships in microbial communities to environment perturbation.

We have shown that central taxa can be captured by centrality measures chosen

using robustness simulations and analysis. Given the structural importance of central taxa and their role in maintaining network structural integrity, it is reasonable to infer that central taxa may play important functional roles in microbial communities and suggests that these OTUs could be keystones. In order to asses their functional (i.e., genetic, metabolic, and biogeochemical) importance, further experimental and quantitative analysis is required. Specifically, the functional roles of the central taxa can be assessed using plurality and singe-cell genomic sequencing or co-culture experiments using representative isolates [41, 46, 66, 113]. For example, assigning taxonomic information to population genome bins reconstructed from shotgun sequencing can determine metabolic potential of specific taxa within the co-occurrence network [42, 66]. The resolved functional associations between taxa has potential to illuminate distributed metabolic pathways linking taxa at community levels.

3.5 Conclusion

Microbial co-occurrence studies have adopted different network analysis methods to find potential keystone taxa. However, the concept of keystone taxa is difficult to tackle given the diversity and complexity of microbial communities [178] and the ambiguity of the species concept, as explained in Section 1.3.3. Understanding how different network measures, including centrality measures, can be interpreted in the context of co-occurrences networks can help identify keystone taxa and determine the impact of disturbance on microbial community structure and function. In the case of LTSP sites, we showed that robustness analysis resolved differential impacts of OM removal on microbial communities across ecozones and determined that these communities were similar in their inferred networks' topology and distribution of central taxa. Furthermore, we identified central taxa from a variety of taxonomies and characterized their soil profile and abundance. These findings demonstrate the capacity of network inference models in microbial ecology research to provide new insights into microbial interactions, community stability and resilience in forest soil ecosystems extensible to other natural and engineered ecosystems.

Chapter 4

Conclusion

This thesis described HYPE, an interactive and data driven exploratory tool for biological networks, and an analytical graph theory based approach to modelling microbial communities from environmental sequence information. This final chapter presents a high-level discussion of the assumptions and limitations HYPE's design and of SSU rRNA sequencing data, outlines the future of network visualization, and concludes by presenting the future integrative needs of microbial ecology.

4.1 Assumptions and limitations of sequencing approaches

High throughput sequencing technologies have bridged the cultivation gap and enable the characterization of microbial community composition and genetic potential. However, certain assumptions and limitation must be considered so as to appropriately analyze and interpret the produced data. First, particularly in diverse environments like soil, SSU rRNA sequencing under-samples the community capturing the most abundant community members [141]. Similarly, the possibility of sequencing errors in singletons, OTUs for which only one sequence has been recruited, challenges their credibility when in fact a singleton could represent a rare organism. Second, the resolution of OTUs' taxonomies at the family, genus and species level remains difficult as the Earth's microbial diversity has not yet been fully document in public databases. Furthermore, public databases of SSU rRNA

genes are biased towards culturable microorganisms. Finally, as the quality and quantity of environmental sequencing increases, microbial ecology research will gain traction in charting microbial diversity on Earth.

4.2 HyPE as a community tool

HYPE enables the exploration of complex systems and drives their quantitative analysis through hypothesis generation. As briefly described in Chapter 2, HYPE has a few usage limitations that need to be acknowledged. In particular, the efficient exploration of a system must adapt to its size and complexity and certain patterns may be visually hidden and require a deeper exploration to be uncovered. Fortunately, as a user gains experience in exploring their system they will find a combination of the colouring rules, filtering rules and interactive tools to use to ease their navigation of their network. In order to decrease the learning curve of the tool and facilitate this learning process, we envision a platform where a community of HYPE users can share their experience with the tool, their adventures in exploring their system, and the patterns they resolved. This type of social and community based learning approach has proved successful on web platforms such as Stack Overflow [1] where novice and expert users pose and answer statistical, mathematical and computer science related questions. Moreover, such a platform would help resolve recurrent usage patterns that can be analyzed to improve HYPE's interactive features and develop navigation guidelines and procedures for novice users. Finally, creating an inter-connected HYPE community can encourage interdisciplinary collaboration and research while helping users make the best out of the tool.

As a collaborative online code host, Github's code sharing features has already increased the awareness of HYPE as a novel network visualization tool. As of June 2015, several dozen unique visitors have visited the repository of which a few have requested features and cloned the repository (created a local copy). With the development of an online user interface, the publication of the tool in a peer-reviewed journal, the development of online use cases for novice users, and a social community platform for sharing hive panels and patterns, this user base will only increase.

4.3 Closing: cross-disciplinarity in microbial ecology

This thesis has demonstrated that the integration of methods from different disciplines can empower researchers studying complex systems such as the *C. elegans* connectome and microbial communities. In Chapter 2 we combined concepts and methods from the fields of information visualization, pattern recognition and networks science to develop a visualization tool. In Chapter 3 we combined sequencing methods, soil ecology, microbial ecology, macroecology methods and network science to demonstrate the applicability of graph theory methods and robustness analysis to evaluating microbial community stability at a taxonomic and community level. As motivated by Dorian Sagan (see Section 1.1) cross-disciplinary synthesis stimulates scientific research and creates scientific breakthroughs [143]. In environmental genomics in particular, the integration of multi-omic sequencing techniques, statistical methods, network science, complexity modelling, high-performance computing, and other disciplines will capacitate researchers to understand and harness the potential of the invisible majority of life on Earth [167].

Bibliography

[1] Stack Overflow, 2015.

[2] M. Achtman and M. Wagner. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6):431–440, June 2008.

[3] A. Agnelli, J. Ascher, G. Corti, M. T. Ceccherini, P. Nannipieri, and G. Pietramellara. Distribution of microbial communities in a forest soil profile investigated by microbial biomass, soil respiration and DGGE of total and extracellular DNA. *Soil Biology and Biochemistry*, 36(5): 859–868, May 2004.

[4] M. Aickin and H. Gensler. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health*, 86(5):726–728, May 1996.

[5] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

[6] M. Alcalde, M. Ferrer, F. J. Plou, and A. Ballesteros. Environmental biocatalysis: from remediation with enzymes to novel green processes. *Trends in Biotechnology*, 24(6):281–287, Jan. 2006.

[7] J. Alstott, E. Bullmore, and D. Plenz. Powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, Jan. 2014. arXiv: 1305.0215.

[8] Z. Altun and D. Hall. Wormatlas, 2002.

[9] L. a. N. Amaral, A. Scala, M. Barthlmy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, Oct. 2000.

[10] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dor, MetaHIT Consortium (additional Members), J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473 (7346):174–180, May 2011.

[11] A. Barbern, S. T. Bates, E. O. Casamayor, and N. Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, 6(2):343–351, Feb. 2012.

[12] A. Barbern, E. O. Casamayor, and N. Fierer. The microbial contribution to macroecology. *Evolutionary and Genomic Microbiology*, 5:203, 2014.

[13] C. I. Bargmann. Neurobiology of the Caenorhabditis elegans Genome. *Science*, 282(5396):2028–2033, Dec. 1998.

[14] C. I. Bargmann and E. Marder. From the connectome to brain function. *Nature Methods*, 10(6):483–490, June 2013.

[15] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media*, Mar. 2009.

[16] D. Berry and S. Widder. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5, May 2014.

[17] N. L. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory 1736-1936*. Clarendon Press, Dec. 1986. ISBN 978-0-19-853916-2.

[18] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, Oct. 2005.

[19] S. R. Borrett. Throughflow centrality is a global indicator of the functional importance of species in ecosystems. *Ecological Indicators*, 32:182–196, Sept. 2013.

[20] M. Bostock. Visualizations with D3, 2015.

[21] M. Bostock, V. Ogievetsky, and J. Heer. D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.

[22] J. L. Bowman, S. K. Floyd, and K. Sakakibara. Green GenesComparative Genomics of the Green Branch of Life. *Cell*, 129(2):229–234, Apr. 2007.

[23] J. H. Boyle, S. Berri, and N. Cohen. Gait Modulation in C. elegans: An Integrated Neuromechanical Model. *Frontiers in Computational Neuroscience*, 6:10, 2012.

[24] M. Brilli and P. Li. The Structural Network Properties of Biological Systems. *Briefings in Functional Genomics*, pages 9–32, 2009.

[25] E. T. Bullmore and D. S. Bassett. Brain Graphs: Graphical Models of the Human Brain Connectome. *Annual Review of Clinical Psychology*, 7(1): 113–140, 2011.

[26] M. D. Busse, S. E. Beattie, R. F. Powers, F. G. Sanchez, and A. E. Tiarks. Microbial community responses in forest mineral soil to compaction, organic matter removal, and vegetation control. *Canadian Journal of Forest Research*, 36(3):577–588, Mar. 2006.

[27] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pea, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010.

[28] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, 2011.

[29] E. Cardenas and J. M. Tiedje. New tools for discovering and characterizing microbial diversity. *Current Opinion in Biotechnology*, 19(6):544–549, Dec. 2008.

[30] E. Cardenas, J. M. Kranabetter, G. Hope, K. R. Maas, S. Hallam, and W. W. Mohn. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. *The ISME Journal*, Apr. 2015.

[31] E. Cardenas, J. M. Kranabetter, G. Hope, K. R. Maas, S. Hallam, and W. W. Mohn. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. *The ISME Journal*, Apr. 2015.

[32] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 36(suppl 1):D623–D631, Jan. 2008.

[33] J. M. Chase. Stochastic community assembly causes higher biodiversity in more productive environments. *Science (New York, N.Y.)*, 328(5984): 1388–1391, June 2010.

[34] B. L. Chen, D. H. Hall, and D. B. Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4723–4728, Mar. 2006.

[35] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), Dec. 2004. arXiv: cond-mat/0408187.

[36] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, Nov. 2009. arXiv: 0706.1062.

[37] M. K. Coleman and D. S. Parker. Aesthetics-based Graph Layout for Human Consumption. *Softw. Pract. Exper.*, 26(12):1415–1438, Dec. 1996.

[38] T. H. M. P. Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[39] W. R. Cookson, D. A. Abaye, P. Marschner, D. V. Murphy, E. A. Stockdale, and K. W. T. Goulding. The contribution of soil organic matter fractions to carbon and nitrogen mineralization and microbial community size and structure. *Soil Biology and Biochemistry*, 37(9):1726–1737, Sept. 2005.

[40] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews. Microbiology*, 2(5):414–424, May 2004.

[41] J. A. Dodsworth, P. C. Blainey, S. K. Murugapiran, W. D. Swingley, C. A. Ross, S. G. Tringe, P. S. G. Chain, M. B. Scholz, C.-C. Lo, J. Raymond, S. R. Quake, and B. P. Hedlund. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications*, 4:1854, 2013.

[42] J. Drge and A. C. McHardy. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, 13(6):646–655, Nov. 2012.

[43] M. Driscoll and J. Kaplan. Mechanotransduction. In D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess, editors, *C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2nd edition, 1997. ISBN 0879695323.

[44] A. J. Dumbrell, M. Nelson, T. Helgason, C. Dytham, and A. H. Fitter. Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, 4(3):337–345, Nov. 2009.

[45] J. A. Dunne, R. J. Williams, and N. D. Martinez. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, 5(4):558–567, 2002.

[46] A. E. Duran-Pinedo, B. Paster, R. Teles, and J. Frias-Lopez. Correlation Network Analysis Applied to Complex Biofilm Communities. *PLoS ONE*, 6(12):e28438, Dec. 2011.

[47] A. Eiler, F. Heinrich, and S. Bertilsson. Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal*, 6(2): 330–342, Feb. 2012.

[48] E. Estrada. Characterization of topological keystone species: Local, global and meso-scale centralities in food webs. *Ecological Complexity*, 4(12): 48–57, Mar. 2007.

[49] E. Estrada and r. Bodin. Using network centrality measures to manage landscape connectivity. *Ecological Applications*, 18(7):1810–1825, Sept. 2008.

[50] M. Fahle. Human pattern recognition: parallel processing and perceptual learning. *Perception*, 23(4):411–427, 1994.

[51] P. G. Falkowski, T. Fenchel, and E. F. Delong. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039, May 2008.

[52] S. L. Fann and S. R. Borrett. Environ centrality reveals the tendency of indirect effects to homogenize the functional importance of species in ecosystems. *Journal of Theoretical Biology*, 294:74–86, Feb. 2012.

[53] K. Faust. CoNet - A Cytoscape plugin that detects significant association in presence/absence and abundance matrices, 2014.

[54] K. Faust and J. Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, Aug. 2012.

[55] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol*, 8(7):e1002606, July 2012.

[56] J. Fekete. Visualizing networks using adjacency matrices: Progresses and challenges. In *11th IEEE International Conference on Computer-Aided Design and Computer Graphics, 2009. CAD/Graphics '09*, pages 636–638, Aug. 2009.

[57] N. Fierer, J. P. Schimel, and P. A. Holden. Variations in microbial community composition through two soil depth profiles. *Soil Biology and Biochemistry*, 35(1):167–176, Jan. 2003.

[58] N. Fierer, M. A. Bradford, and R. B. Jackson. Toward an ecological classification of soil bacteria. *Ecology*, 88(6):1354–1364, June 2007.

[59] B. J. Finlay, S. C. Maberly, and J. I. Cooper. Microbial Diversity and Ecosystem Function. *Oikos*, 80(2):209–213, Nov. 1997.

[60] T. Fleming, S.-C. Chien, P. J. Vanderzalm, M. Dell, M. K. Gavin, W. C. Forrester, and G. Garriga. The role of C. elegans Ena/VASP homolog UNC-34 in neuronal polarity and motility. *Developmental Biology*, 344(1): 94–106, Aug. 2010.

[61] R. M. Fox, S. E. Von Stetina, S. J. Barlow, C. Shaffer, K. L. Olszewski, J. H. Moore, D. Dupuy, M. Vidal, and D. M. Miller. A gene expression fingerprint of C. elegans embryonic motor neurons. *BMC Genomics*, 6:42, Mar. 2005.

[62] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[63] J. Friedman and E. J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput Biol*, 8(9):e1002687, Sept. 2012.

[64] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, Nov. 1991.

[65] H. Gibson, J. Faith, and P. Vickers. A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization*, 12(3-4):324–357, July 2013.

[66] E. A. Gies, K. M. Konwar, J. T. Beatty, and S. J. Hallam. Illuminating Microbial Dark Matter in Meromictic Sakinaw Lake. *Applied and Environmental Microbiology*, pages AEM.01774–14, Aug. 2014.

[67] J. A. Gilbert, J. K. Jansson, and R. Knight. The Earth Microbiome project: successes and aspirations. *BMC Biology*, 12(1):69, Aug. 2014.

[68] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, June 2002.

[69] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabsi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007.

[70] J. M. Gray, D. S. Karow, H. Lu, A. J. Chang, J. S. Chang, R. E. Ellis, M. A. Marletta, and C. I. Bargmann. Oxygen sensation and social feeding mediated by a C. elegans guanylate cyclase homologue. *Nature*, 430 (6997):317–322, July 2004.

[71] J. M. Gray, J. J. Hill, and C. I. Bargmann. A circuit for navigation in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3184–3191, Mar. 2005.

[72] S. J. Hallam. *Mechanisms of neuronal asymmetry and synaptic remodeling in the nematode Caenorhabditis elegans*. PhD thesis, 2000.

[73] S. J. Hallam and J. P. McCutcheon. Microbes don't play solitaire: how cooperation trumps isolation in the microbial world. *Environmental Microbiology Reports*, 7(1):26–28, Feb. 2015.

[74] C. M. Hansel, S. Fendorf, P. M. Jardine, and C. A. Francis. Changes in Bacterial and Archaeal Community Structure and Functional Diversity along a Geochemically Variable Soil Profile. *Applied and Environmental Microbiology*, 74(5):1620–1633, Mar. 2008.

[75] N. Hanson, K. Konwar, S.-J. Wu, and S. Hallam. MetaPathways v2.0: A master-worker model for environmental Pathway/Genome Database construction on grids and clouds. In *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–7, May 2014.

[76] M. Hartmann, S. Lee, S. J. Hallam, and W. W. Mohn. Bacterial, archaeal and eukaryal community structures throughout soil horizons of harvested and naturally disturbed forest stands. *Environmental Microbiology*, 11(12): 3045–3062, 2009.

[77] M. Hartmann, C. G. Howes, D. VanInsberghe, H. Yu, D. Bachar, R. Christen, R. Henrik Nilsson, S. J. Hallam, and W. W. Mohn. Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. *The ISME journal*, 6(12):2199–2218, Dec. 2012.

[78] M. Hartmann, P. A. Niklaus, S. Zimmermann, S. Schmutz, J. Kremer, K. Abarenkov, P. Lscher, F. Widmer, and B. Frey. Resistance and resilience of the forest soil microbiome to logging-associated compaction. *The ISME Journal*, 8(1):226–244, Jan. 2014.

[79] A. Hector, B. Schmid, C. Beierkuhnlein, M. C. Caldeira, M. Diemer, P. G. Dimitrakopoulos, J. A. Finn, H. Freitas, P. S. Giller, J. Good, R. Harris, P. Hgberg, K. Huss-Danell, J. Joshi, A. Jumpponen, C. Krner, P. W. Leadley, M. Loreau, A. Minns, C. P. H. Mulder, G. O'Donovan, S. J. Otway, J. S. Pereira, A. Prinz, D. J. Read, M. Scherer-Lorenzen, E.-D. Schulze, A.-S. D. Siamantziouras, E. M. Spehn, A. C. Terry, A. Y. Troumbis, F. I. Woodward, S. Yachi, and J. H. Lawton. Plant Diversity and Productivity Experiments in European Grasslands. *Science*, 286(5442): 1123–1127, Nov. 1999.

[80] M. C. Horner-Devine, J. M. Silver, M. A. Leibold, B. J. M. Bohannan, R. K. Colwell, J. A. Fuhrman, J. L. Green, C. R. Kuske, J. B. H. Martiny, G. Muyzer, L. Ovres, A.-L. Reysenbach, and V. H. Smith. A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology*, 88 (6):1345–1353, June 2007.

[81] S. P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press, Apr. 2001. ISBN 0691021287.

[82] B. L. Hurwitz, S. J. Hallam, and M. B. Sullivan. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biology*, 14(11):R123, Nov. 2013.

[83] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, Sept. 2011.

[84] A. Inselberg and B. Dimsdale. Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In *Proceedings of the 1st Conference on Visualization '90*, VIS '90, pages 361–378, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press. ISBN 0-8186-2083-8.

[85] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang. Attack Robustness and Centrality of Complex Networks. *PLoS ONE*, 8(4):e59613, Apr. 2013.

[86] T. A. Jarrell, Y. Wang, A. E. Bloniarz, C. A. Brittin, M. Xu, J. N. Thomson, D. G. Albertson, D. H. Hall, and S. W. Emmons. The Connectome of a Decision-Making Neural Network. *Science*, 337(6093):437–444, July 2012.

[87] H. Jeong, S. P. Mason, A.-L. Barabsi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[88] B. H. Junker and F. Schreiber. *Analysis of Biological Networks*. John Wiley & Sons, Sept. 2011. ISBN 9781118209912.

[89] E. L. Kara, P. C. Hanson, Y. H. Hu, L. Winslow, and K. D. McMahon. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *The ISME Journal*, 7(3):680–684, Mar. 2013.

[90] J. Karbowski, G. Schindelman, C. J. Cronin, A. Seah, and P. W. Sternberg. Systems level circuit model of C. elegans undulatory locomotion: mathematical modeling and molecular genetics. *Journal of Computational Neuroscience*, 24(3):253–276, June 2008.

[91] P. Khanna. *Essentials of Genetics*. I. K. International Pvt Ltd, 2010. ISBN 9789380026343.

[92] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov. 2002.

[93] K. M. Konwar, N. W. Hanson, A. P. Pag, and S. J. Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC bioinformatics*, 14:202, 2013.

[94] J. Kranabetter. Site carbon storage along productivity gradients of a late-seral southern boreal forest. *Canadian Journal of Forest Research*, 39 (5):1053–1060, May 2009.

[95] M. Krzywinski, J. Schein, n. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, Sept. 2009.

[96] M. Krzywinski, I. Birol, S. J. Jones, and M. A. Marra. Hive plotsrational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5): 627–644, Sept. 2012.

[97] R. Lal. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science*, 304(5677):1623–1627, June 2004.

[98] R. Lamendella, S. Strutt, S. Borglin, R. Chakraborty, N. Tas, O. U. Mason, J. Hultman, E. Prestat, T. C. Hazen, and J. K. Jansson. Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial communities. *Frontiers in Microbiology*, 5:130, 2014.

[99] M. G. I. Langille and F. S. L. Brinkman. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, 25(5):664–665, Mar. 2009.

[100] C. L. Lauber, M. Hamady, R. Knight, and N. Fierer. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Applied and Environmental Microbiology*, 75(15):5111–5120, Aug. 2009.

[101] P. Legendre and L. Legendre. Numerical Ecology. In P. a. L. Legendre, editor, *Developments in Environmental Modelling*, volume 24 of *Numerical Ecology*, pages 337–424. Elsevier, 2012.

[102] C. Lei and J. Ruan. A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. *Bioinformatics*, page bts688, Dec. 2012.

[103] M. Li, J. Wang, and J. Chen. A Fast Agglomerate Algorithm for Mining Functional Modules in Protein Interaction Networks. In *International Conference on BioMedical Engineering and Informatics, 2008. BMEI 2008*, volume 1, pages 3–7, May 2008.

[104] C. A. Lozupone and R. Knight. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27):11436–11440, July 2007.

[105] M. Lupatini, A. K. A. Suleiman, R. J. S. Jacques, Z. I. Antoniolli, A. de Siqueira Ferreira, E. E. Kuramae, and L. F. W. Roesch. Network topology reveals high connectance levels and few key microbial genera within soils. *Soil Processes*, 2:10, 2014.

[106] G. S. Maro, M. P. Klassen, and K. Shen. A -Catenin-Dependent Wnt Pathway Mediates Anteroposterior Axon Guidance in C. elegans Motor Neurons. *PLoS ONE*, 4(3):e4690, Mar. 2009.

[107] A. M. Martn Gonzlez, B. Dalsgaard, and J. M. Olesen. Centrality measures and the importance of generalist species in pollination networks. *Ecological Complexity*, 7(1):36–43, Mar. 2010.

[108] C. J. Marx. Getting in Touch with Your Friends. *Science*, 324(5931): 1150–1151, May 2009.

[109] S. L. McIntire, E. Jorgensen, and H. R. Horvitz. Genes required for GABA function in Caenorhabditis elegans. *Nature*, 364(6435):334–337, July 1993.

[110] M. T. Mee, J. J. Collins, G. M. Church, and H. H. Wang. Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 111(20): E2149–2156, May 2014.

[111] T. Mino and H. Satoh. Wastewater genomics. *Nature Biotechnology*, 24 (10):1229–1230, Oct. 2006.

[112] E. Mkinen. On circular layouts. *International Journal of Computer Mathematics*, 24(1):29–37, Jan. 1988.

[113] S. Mller, C. Sternberg, J. B. Andersen, B. B. Christensen, J. L. Ramos, M. Givskov, and S. Molin. In situ gene expression in mixed-culture biofilms: evidence of metabolic interactions between community members. *Applied and Environmental Microbiology*, 64(2):721–732, Feb. 1998.

[114] S. Mocali and A. Benedetti. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Research in Microbiology*, 161(6):497–505, July 2010.

[115] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. A K Peters/CRC Press, 2014.

[116] A. Naqvi, H. Rangwala, A. Keshavarzian, and P. Gillevet. Network-based modeling of the human gut microbiome. *Chemistry & Biodiversity*, 7(5): 1040–1050, May 2010.

[117] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, Jan. 2001.

[118] M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.

[119] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, Mar. 2004.

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

[121] D. S. Page-Dumroese, M. F. Jurgensen, A. E. Tiarks, F. Ponder, F. G. Sanchez, R. L. Fleming, J. M. Kranabetter, R. F. Powers, D. M. Stone, J. D. Elioff, and D. A. Scott. Soil physical property changes at the North American Long-Term Soil Productivity study sites: 1 and 5 years after compaction. 2006.

[122] M. Pascual and J. A. Dunne. *Ecological Networks: Linking Structure to Dynamics in Food Webs*. Oxford University Press, Nov. 2005. ISBN 9780199775057.

[123] G. C. Pereira, F. F. Santos, and N. F. F. Ebecken. Centrality and Network Analysis in a Natural Perturbed Ecosystem. In R. Menezes, A. Evsukoff, and M. C. Gonzlez, editors, *Complex Networks*, number 424 in Studies in Computational Intelligence, pages 217–224. Springer Berlin Heidelberg, Jan. 2013. ISBN 978-3-642-30286-2, 978-3-642-30287-9.

[124] A. Perer and B. Shneiderman. Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700, Sept. 2006.

[125] A. Perer and B. Shneiderman. Integrating Statistics and Visualization: Case Studies of Gaining Clarity During Exploratory Data Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 265–274, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1.

[126] A. Perer and B. Shneiderman. Systematic Yet Flexible Discovery: Guiding Domain Experts Through Exploratory Data Analysis. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 109–118, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-987-6.

[127] N. Perra and S. Fortunato. Spectral centrality measures in complex networks. *Physical Review E*, 78(3):036107, Sept. 2008.

[128] M. Pester, K.-H. Knorr, M. W. Friedrich, M. Wagner, and A. Loy. Sulfate-Reducing Microorganisms in Wetlands Fameless Actors in Carbon Cycling and Climate Change. *Frontiers in Microbiology*, 3, Feb. 2012.

[129] S. Peura, S. Bertilsson, R. I. Jones, and A. Eiler. Resistant microbial co-occurrence patterns inferred by network topology. *Applied and Environmental Microbiology*, pages AEM.03660–14, Jan. 2015.

[130] D. Pils, A. Bachmayr-Heyda, K. Auer, M. Svoboda, V. Auner, G. Hager, E. Obermayr, A. Reiner, A. Reinthaller, P. Speiser, I. Braicu, J. Sehouli, S. Lambrechts, I. Vergote, S. Mahner, A. Berger, D. Cacsire Castillo-Tong, and R. Zeillinger. Cyclin E1 (CCNE1) as independent positive prognostic factor in advanced stage serous ovarian cancer patients A study of the OVCAD consortium. *European Journal of Cancer*, 50(1):99–110, Jan. 2014.

[131] C. Plaisant. The Challenge of Information Visualization Evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM. ISBN 1-58113-867-9.

[132] F. Ponder Jr., R. L. Fleming, S. Berch, M. D. Busse, J. D. Elioff, P. W. Hazlett, R. D. Kabzems, J. Marty Kranabetter, D. M. Morris, D. Page-Dumroese, B. J. Palik, R. F. Powers, F. G. Sanchez,

D. Andrew Scott, R. H. Stagg, D. M. Stone, D. H. Young, J. Zhang, K. H. Ludovici, D. W. McKenney, D. S. Mossa, P. T. Sanborn, and R. A. Voldseth. Effects of organic matter removal, soil compaction and vegetation control on 10th year biomass and foliar nutrition: LTSP continent-wide comparisons. *Forest Ecology and Management*, 278:35–54, Aug. 2012.

[133] T. Powell, F. Schneider, and N. Maragioglio. *JavaScript: The Complete Reference, 2Nd Edition*. McGraw-Hill, Inc., New York, NY, USA, 2 edition, 2004. ISBN 0072253576, 9780072253573.

[134] M. E. Power, D. Tilman, J. A. Estes, B. A. Menge, W. J. Bond, L. S. Mills, G. Daily, J. C. Castilla, J. Lubchenco, and R. T. Paine. Challenges in the Quest for Keystones Identifying keystone species is difficultbut essential to understanding how loss of species will affect ecosystems. *BioScience*, 46 (8):609–620, Sept. 1996.

[135] R. F. Powers. Sustaining site productivity in North American forests: problems and prospects. In S. Gessel, D. Lacate, and G. Weetman, editors, *Porceedings from the 7th North American Soil Forests Conference*, Vancouver, BC, 1990. Faculty of Forestry, University of British Columbia.

[136] R. F. Powers, D. Andrew Scott, F. G. Sanchez, R. A. Voldseth, D. Page-Dumroese, J. D. Elioff, and D. M. Stone. The North American long-term soil productivity experiment: Findings from the first decade of research. *Forest Ecology and Management*, 220(13):31–50, Dec. 2005.

[137] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, page gks1219, Nov. 2012.

[138] Y. Rafrafi, E. Trably, J. Hamelin, E. Latrille, I. Meynial-Salles, S. Benomar, M.-T. Giudici-Orticoni, and J.-P. Steyer. Sub-dominant bacteria as keystone species in microbial communities producing bio-hydrogen. *International Journal of Hydrogen Energy*, 38(12):4975–4985, Apr. 2013.

[139] P. H. Rampelotto, A. D. M. Barboza, A. B. Pereira, E. W. Triplett, C. E. G. R. Schaefer, F. A. de Oliveira Camargo, and L. F. W. Roesch. Distribution and interaction patterns of bacterial communities in an ornithogenic soil of Seymour Island, Antarctica. *Microbial Ecology*, 69(3): 684–694, Apr. 2015.

[140] S. Rayu, D. G. Karpouzas, and B. K. Singh. Emerging technologies in bioremediation: constraints and opportunities. *Biodegradation*, 23(6): 917–926, Nov. 2012.

[141] L. F. W. Roesch, R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. H. Daroub, F. A. O. Camargo, W. G. Farmerie, and E. W. Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, 1(4):283–290, July 2007.

[142] G. Rossum. Python Tutorial. Technical report, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, The Netherlands, 1995.

[143] D. Sagan. *Cosmic Apprentice: Dispatches from the Edges of Science*. Univ Of Minnesota Press, Minneapolis ; London, May 2013. ISBN 9780816681358.

[144] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn, and C. F. Weber. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23): 7537–7541, Dec. 2009.

[145] A. Shade, H. Peter, S. D. Allison, D. L. Baho, M. Berga, H. Brgmann, D. H. Huber, S. Langenheder, J. T. Lennon, J. B. H. Martiny, K. L. Matulich, T. M. Schmidt, and J. Handelsman. Fundamentals of Microbial Community Resistance and Resilience. *Frontiers in Microbiology*, 3, Dec. 2012.

[146] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, Nov. 2003.

[147] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society. ISBN 0-8186-7508-X.

[148] R. R. Sokal and F. J. Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40, Feb. 1962.

[149] J. T. Staley. The bacterial species dilemma and the genomicphylogenetic species concept. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1475):1899–1909, Nov. 2006.

[150] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C.-E. T. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A. Caron, and J. A. Fuhrman. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, 5(9):1414–1425, Sept. 2011.

[151] N. S. Sutherland. Outlines of a Theory of Visual Pattern Recognition in Animals and Man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):297–317, Dec. 1968.

[152] R. Suzuki and H. Shimodaira. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling, Dec. 2014.

[153] S. D. P. r. B. Tabacof, Tim //r//nAU Larson. Beyond the connectome hairball: Rational visualizations and analysis of the C. elegans connectome as a network graph using hive plots. *Frontiers in Neuroinformatics*.

[154] X. Tan, S. X. Chang, and R. Kabzems. Soil compaction and forest floor removal reduced microbial biomass and enzyme activities in a boreal aspen forest soil. *Biology and Fertility of Soils*, 44(3):471–479, Aug. 2007.

[155] Tapiocozzo. Figure of six centrality measureson same graph (adapted by Sarah Perez), Apr. 2015. Page Version ID: 657599368.

[156] E. Tortoli. Impact of Genotypic Studies on Mycobacterial Taxonomy: the New Mycobacteria of the 1990s. *Clinical Microbiology Reviews*, 16(2): 319–354, Apr. 2003.

[157] E. K. Towlson, P. E. Vrtes, S. E. Ahnert, W. R. Schafer, and E. T. Bullmore. The rich club of the C. elegans neuronal connectome. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 33(15): 6380–6387, Apr. 2013.

[158] A. Valverde, T. P. Makhalanyane, and D. A. Cowan. Contrasting assembly processes in a bacterial metacommunity along a desiccation gradient. *Frontiers in Microbiology*, 5, Dec. 2014.

[159] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural Properties of the Caenorhabditis elegans Neuronal Network. *PLoS Comput Biol*, 7(2):e1001066, Feb. 2011.

[160] A. T. Vincent and S. J. Charette. Freedom in bioinformatics. *Frontiers in Genetics*, 5, July 2014.

[161] W. W. Walthall, L. Li, J. A. Plunkett, and C.-Y. Hsu. Changing synaptic specificities in the nervous system of Caenorhabditis elegans: Differentiation of the DD motoneurons. *Journal of Neurobiology*, 24(12): 1589–1599, Dec. 1993.

[162] S. Wasserman. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Nov. 1994. ISBN 9780521387071.

[163] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998.

[164] J. G. White, D. G. Albertson, and M. a. R. Anness. Connectivity changes in a class of motoneurone during the development of a nematode. *Nature*, 271 (5647):764–766, Feb. 1978.

[165] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. Factors that determine connectivity in the nervous system of Caenorhabditis elegans. *Cold Spring Harbor Symposia on Quantitative Biology*, 48 Pt 2:633–640, 1983.

[166] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The Structure of the Nervous System of the Nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 314(1165):1–340, Nov. 1986.

[167] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12): 6578–6583, June 1998.

[168] C. Will, A. Thrmer, A. Wollherr, H. Nacke, N. Herold, M. Schrumpf, J. Gutknecht, T. Wubet, F. Buscot, and R. Daniel. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16s rRNA genes. *Applied and Environmental Microbiology*, 76(20):6751–6759, Oct. 2010.

[169] R. J. Williams, A. Howe, and K. S. Hofmockel. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Terrestrial Microbiology*, 5:358, 2014.

[170] J. J. Wright. Microbial community structure and ecology of Marine Group A bacteria in the oxygen minimum zone of the Northeast subarctic Pacific Ocean. 2013.

[171] J. J. Wright, K. M. Konwar, and S. J. Hallam. Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology*, 10(6): 381–394, June 2012.

[172] E. Wu, T. Nance, and S. B. Montgomery. SplicePlot: a utility for visualizing splicing quantitative trait loci. *Bioinformatics (Oxford, England)*, 30(7):1025–1026, Apr. 2014.

[173] W. Zachary. Information-Flow Model for Conflict and Fission in Small-Groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. WOS:A1977FG85500006.

[174] A. J. Zehnder and T. D. Brock. Methane formation and methane oxidation by methanogenic bacteria. *Journal of Bacteriology*, 137(1):420–432, Jan. 1979.

[175] A. Zelezniak, S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences*, page 201421834, May 2015.

[176] J. Zhao, Q. Liu, and X. Wang. Competitive Dynamics on Complex Networks. *Scientific Reports*, 4, July 2014.

[177] M. Zhen and A. D. Samuel. C. elegans locomotion: small circuits, complex functions. *Current Opinion in Neurobiology*, 33:117–126, Aug. 2015.

[178] J. Zhou, Y. Deng, F. Luo, Z. He, and Y. Yang. Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO2. *mBio*, 2(4):e00122–11, Sept. 2011.

Appendix A

Chapter 3 supporting material

Table A.1: Number of sequences recovered for samples in ecozone JP with treatment OM0

Sample Id	Number of sequences	Sample Id	Number of sequences
JE122	9437	JS083	9810
JE123	3711	JS084	7643
JE124	7607	JW013	5730
JE125	7943	JW014	5765
JE126	11042	JW019	15763
JS079	16247	JW020	9750
JS080	9802	JW026	6080

Table A.2: Number of sequences recovered for samples in ecozone JP with treatment OM1

Sample Id	Number of sequences	Sample Id	Number of sequences
JE086	5351	JS066	6669
JE087	6722	JS075	11716
JE088	8898	JS076	3465
JE105	7970	JS077	7166
JE106	5012	JS078	10627
JE107	7717	JW005	6010
JE108	5046	JW006	10136
JE117	6965	JW007	7390
JE118	16540	JW008	5974
JE119	7537	JW021	8619
JE120	8787	JW022	20003
JS043	7756	JW023	14810
JS044	7295	JW024	3854
JS045	11382	JW027	10694
JS046	8793	JW028	10651
JS063	3647	JW029	7678
JS064	8167	JW030	7358

Table A.3: Number of sequences recovered for samples in ecozone JP with treatment OM2

Sample Id	Number of sequences	Sample Id	Number of sequences
JE094	1964	JS060	10059
JE095	12584	JS062	12919
JE096	8922	JS068	8416
JE101	8847	JW015	6425
JE102	6212	JW016	7745
JE103	4257	JW017	8166
JE104	4355	JW018	7403
JE113	6989	JW031	7802
JE114	5888	JW032	8521
JE115	7187	JW033	8590
JE116	8624	JW034	10237
JS051	4689	JW035	7383
JS052	6653	JW036	4180
JS053	16082	JW037	11295
JS054	10421	JW038	6181

Table A.4: Number of sequences recovered for samples in ecozone JP with treatment OM3

Sample Id	Number of sequences	Sample Id	Number of sequences
JE092	6223	JS074	5337
JE098	11849	JW002	5945
JE100	4294	JW003	6755
JE110	8403	JW004	9522
JE112	7751	JW010	6313
JS048	10142	JW012	8096
JS050	9501	JW040	6700
JS056	12999	JW042	7214
JS058	6284		

Table A.5: Number of sequences recovered for samples in ecozone MD with treatment OM0

Sample Id	Number of sequences	Sample Id	Number of sequences
BL044	8237	BR071	6132
BL045	9041	BR072	8839
BL046	9110	LH019	6389
BL047	5374	LH020	4973
BL048	7757	LH021	6698
BR067	10270	LH022	3870
BR068	10720	LH023	5030
BR069	8476	LH024	5034

Table A.6: Number of sequences recovered for samples in ecozone MD with treatment OM1

Sample Id	Number of sequences	Sample Id	Number of sequences
BL026	12668	BR053	8763
BL027	5182	BR054	3884
BL028	3677	LH001	10345
BL029	5411	LH002	6087
BL030	11238	LH003	6025
BR049	4815	LH004	7804
BR050	2532	LH005	6764
BR051	5121	LH006	6655

Table A.7: Number of sequences recovered for samples in ecozone MD with treatment OM2

Sample Id	Number of sequences	Sample Id	Number of sequences
BL032	3921	BR059	4798
BL033	5451	BR060	7948
BL034	3279	LH007	9937
BL035	3456	LH008	1883
BL036	12921	LH009	9625
BR055	10010	LH010	7537
BR056	9974	LH011	8605
BR057	6757	LH012	6000

Table A.8: Number of sequences recovered for samples in ecozone MD with treatment OM3

Sample Id	Number of sequences	Sample Id	Number of sequences
BL038	7839	BR065	9514
BL039	10362	BR066	6010
BL040	12021	LH013	12495
BL041	9506	LH014	4967
BL042	6956	LH015	6512
BR061	6097	LH016	8432
BR062	7366	LH017	8522
BR063	8168	LH018	6185

Table A.9: Number of sequences recovered for samples in ecozone SBS with treatment OM0

Sample Id	Number of sequences	Sample Id	Number of sequences
LL056	3716	SL180	2169
LL057	3025	TO115	4482
LL058	2052	TO116	3711
LL059	3518	TO117	3956
LL060	3699	TO118	2654
SL175	2899	TO119	1973
SL176	1722	TO120	2308
SL177	2213		

Table A.10: Number of sequences recovered for samples in ecozone SBS with treatment OM1

Sample Id	Number of sequences	Sample Id	Number of sequences
LL002	4313	SL131	3430
LL003	3970	SL132	2267
LL004	4455	SL133	2516
LL005	3305	SL134	1797
LL006	4047	SL135	3095
LL019	4154	SL136	1992
LL020	3120	SL137	1721
LL021	4452	SL138	1853
LL022	3843	TO061	4232
LL023	5124	TO062	3376
LL024	2904	TO063	5373
LL037	4235	TO064	3065
LL038	4419	TO065	2151
LL039	3444	TO066	2541
LL040	3308	TO079	3779
LL041	3283	TO080	4407
LL042	3151	TO081	3676
SL121	2687	TO082	3101
SL122	3171	TO083	4309
SL123	4593	TO084	2730
SL124	1934	TO097	3543
SL125	2757	TO098	3677
SL126	2151	TO099	3669
SL127	2552	TO100	3187
SL128	2795	TO101	2945
SL129	2056	TO102	2771

Table A.11: Number of sequences recovered for samples in ecozone SBS with treatment OM2

Sample Id	Number of sequences	Sample Id	Number of sequences
LL008	3972	SL149	1366
LL009	2628	SL150	2017
LL010	2429	SL151	2583
LL011	3965	SL152	2776
LL012	4701	SL153	1551
LL025	3768	SL154	1832
LL026	3765	SL155	2773
LL027	3943	SL156	2016
LL028	2926	TO067	4564
LL029	3450	TO068	4706
LL030	4188	TO069	4850
LL043	2986	TO070	3494
LL044	3243	TO071	3367
LL045	4308	TO072	3623
LL046	4368	TO085	7649
LL047	3183	TO086	4748
LL048	2947	TO087	5291
SL139	2535	TO089	3925
SL140	1398	TO090	4111
SL141	1713	TO103	2660
SL142	1739	TO104	4959
SL143	1217	TO105	4536
SL144	1669	TO106	3515
SL145	1326	TO107	2696
SL146	1571	TO108	3407
SL147	2091		

Table A.12: Number of sequences recovered for samples in ecozone SBS with treatment OM3

Sample Id	Number of sequences	Sample Id	Number of sequences
LL017	2346	SL172	4276
LL018	2632	SL173	2159
LL034	4184	SL174	2833
LL035	4230	TO076	2140
LL036	4693	TO077	3942
LL052	2025	TO078	4621
LL053	3428	TO094	4063
LL054	2075	TO095	3604
SL160	2020	TO096	4094
SL161	1976	TO112	4045
SL162	2755	TO113	2569
SL166	2339	TO114	4131
SL167	1789		

Table A.13: Summary of samples numbers in each ecozone for each treatment level

Treatment	OM0	OM1	OM2	OM3	Ecozone total
SBS	17	54	53	27	151
MD	18	18	18	18	72
JP	16	36	32	19	103
Treatment total	51	108	103	64	326

Table A.14: Shannon's entropy of LTSP samples grouped by ecozone and treatment

	OM0	OM1	OM2	OM3
SBS	4.1	4.3	4.3	4.6
MD	5.8	5.7	6.0	6.1
JP	5.6	5.5	5.5	5.7

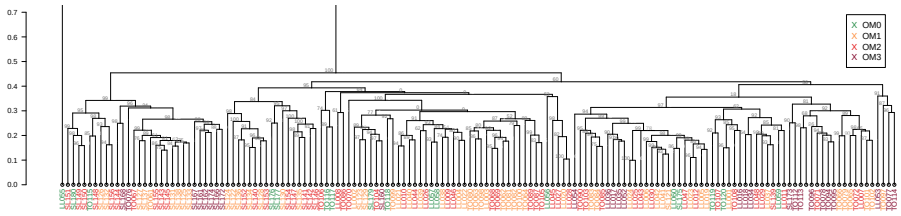

Figure A.1: Hierarchical clustering of SBS samples colored by treatment

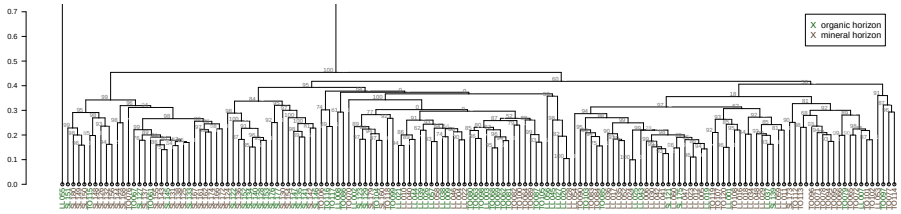

Figure A.2: Hierarchical clustering of SBS samples colored by horizon

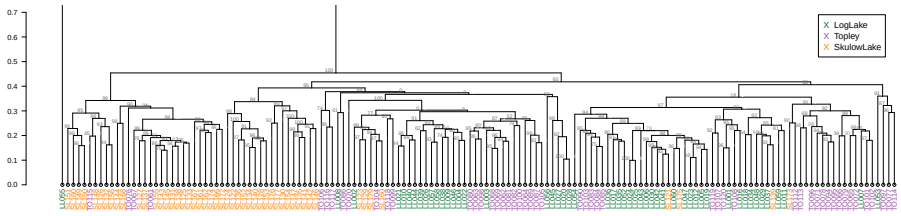

Figure A.3: Hierarchical clustering of SBS samples colored by sample site

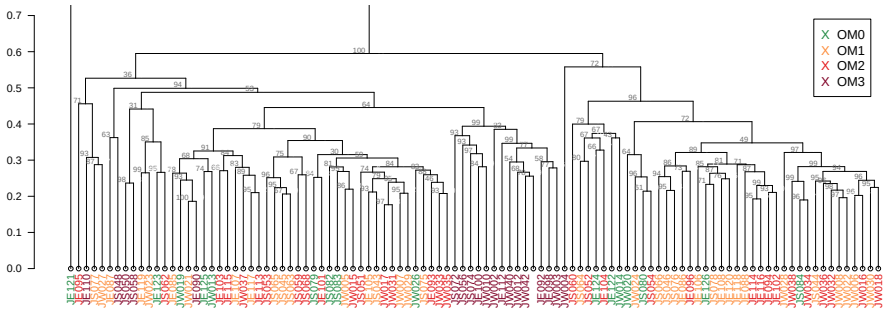

Figure A.4: Hierarchical clustering of JP samples colored by treatment

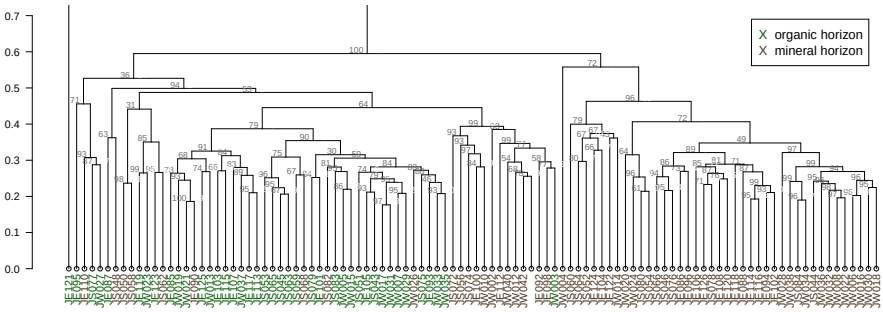

Figure A.5: Hierarchical clustering of JP samples colored by horizon

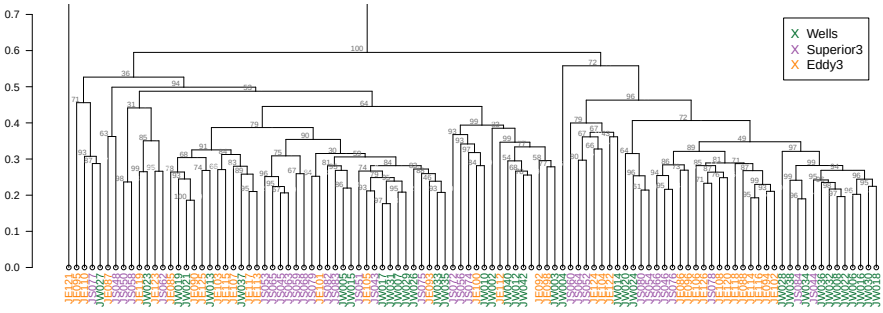

Figure A.6: Hierarchical clustering of JP samples colored by sample site

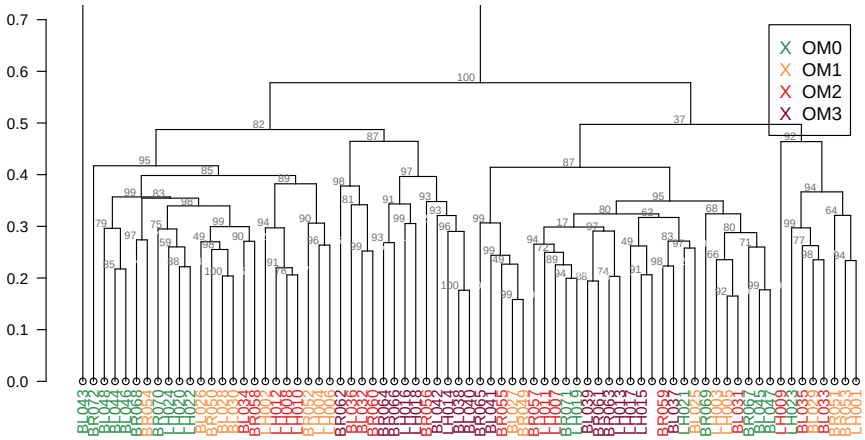

Figure A.7: Hierarchical clustering of MD samples colored by treatment

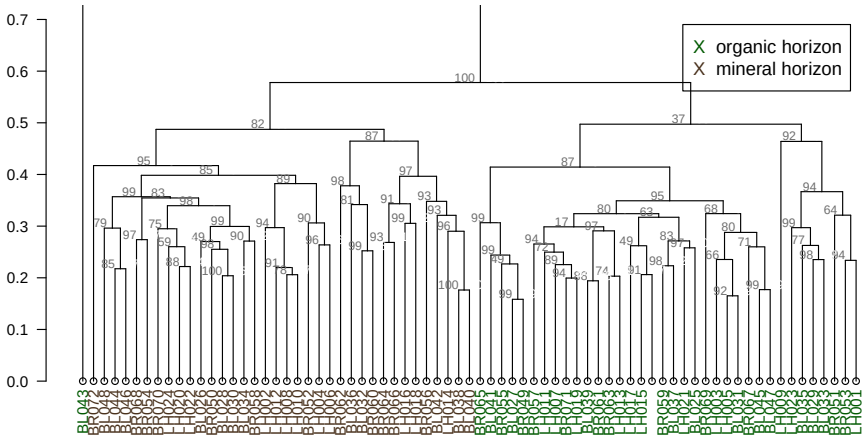

Figure A.8: Hierarchical clustering of MD samples colored by horizon

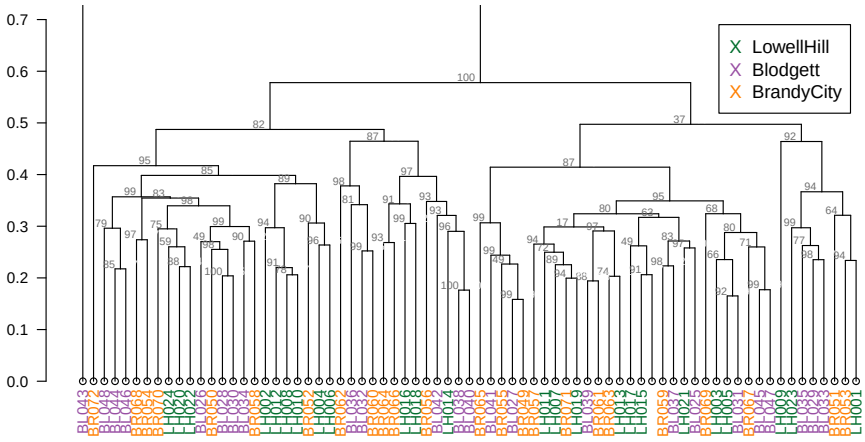

Figure A.9: Hierarchical clustering of MD samples colored by sample site

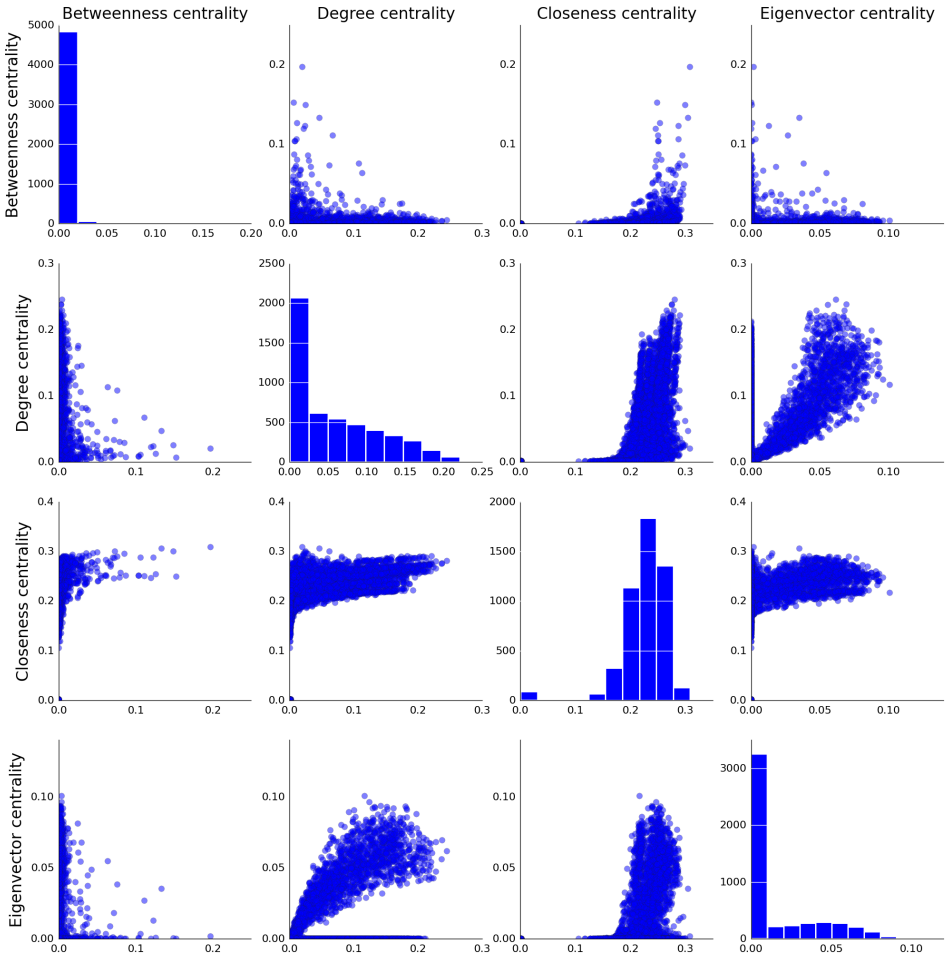

Figure A.10: Scatter matrix plot of four centrality measures in the MD networks. Histograms of each centrality measure is also shown. The different centrality values of OTUs for each treatment network was pooled to produce these plots.

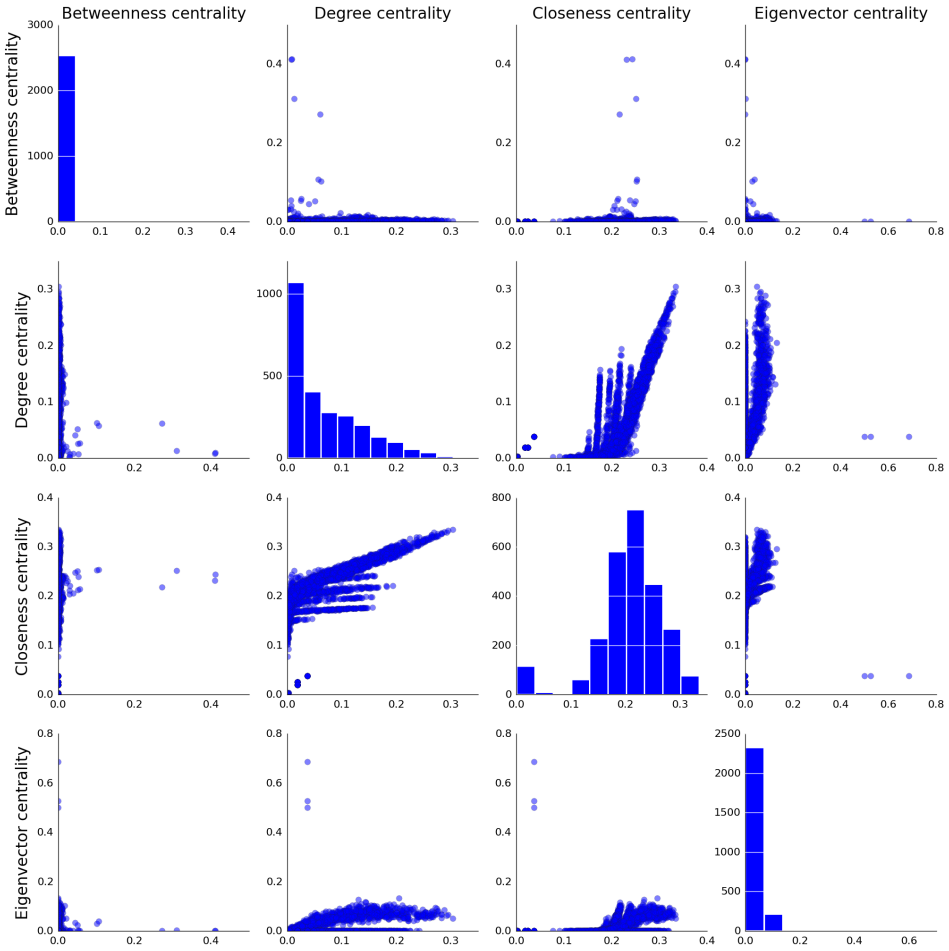

Figure A.11: Scatter matrix plot of four centrality measures in the JP networks. Histograms of each centrality measure is also shown. The different centrality values of OTUs for each treatment network was pooled to produce these plots.

Table A.15: Representation of phyla in central taxa of JP networks

Phylum	Number of taxa	Number of central taxa
Acidobacteria	395	4
Actinobacteria	425	8
Bacteroidetes	36	1
Candidate division OP10	2	0
Candidate division TG-1	3	0
Candidate division TM6	1	0
Candidate division TM7	17	0
Cyanobacteria	50	0
Firmicutes	26	0
Gemmatimonadetes	26	0
Planctomycetes	110	0
Proteobacteria	1028	17
Verrucomicrobia	25	0
WCHB1-60	2	0

Table A.16: Representation of classes in central taxa of JP networks

Class	Number of taxa	Number of central taxa
Acidobacteria	379	4
Actinobacteria	425	8
Alphaproteobacteria	814	14
Bacilli	23	0
Betaproteobacteria	63	0
Chloroplast	1	0
Deltaproteobacteria	77	0
Gammaproteobacteria	71	3
Gemmatimonadetes	26	0
Holophagae	13	0
Lineage IV	3	0
MLE1-12	7	0
Opitutae	19	0
Phycisphaerae	25	0
Planctomycetacia	76	0
Spartobacteria	6	0
Sphingobacteria	36	1
WD272	42	0

Table A.17: Representation of orders in central taxa of JP networks

Order	Number of taxa	Number of central taxa
32-20	10	0
Acidimicrobidae	107	0
Acidobacteriales	379	4
Actinobacteridae	229	8
Bacillales	23	0
Burkholderiales	42	0
Candidatus Xiphinematobacter	2	0
Caulobacterales	37	0
GR-WP33-30	27	0
Gemmatimonadales	26	0
Legionellales	1	0
Myxococcales	49	0
Nitrosomonadales	7	0
Opitutales	19	0
Planctomycetales	76	0
Rhizobiales	376	10
Rhodospirillales	377	4
Rubrobacteridae	81	0
SC-I-84	9	0
Sphingobacteriales	36	1
TRA3-20	1	0
WD2101	25	0
Xanthomonadales	62	3
iii1-8	3	0

Table A.18: Representation of phyla in central taxa of MD networks

Phylum	Number of taxa	Number of central taxa
Acidobacteria	684	17
Actinobacteria	1097	45
Bacteroidetes	271	4
Candidate division OP10	9	0
Candidate division TM7	16	0
Candidate division WS3	7	0
Chloroflexi	89	2
Cyanobacteria	22	2
Fibrobacteres	3	0
Firmicutes	37	1
Gemmatimonadetes	86	5
Nitrospirae	4	1
Planctomycetes	115	1
Proteobacteria	1942	68
Verrucomicrobia	51	0
WCHB1-60	6	0

Table A.19: Representation of classes in central taxa of MD networks

Class	Number of taxa	Number of central taxa
Acidobacteria	659	17
Actinobacteria	1097	45
Alphaproteobacteria	1376	48
Anaerolineae	1	1
Bacilli	25	1
Betaproteobacteria	291	8
Chloroflexi	5	0
Chloroplast	3	0
Clostridia	1	0
Deltaproteobacteria	164	6
Fibrobacteria	3	0
Flavobacteria	1	0
Gammaproteobacteria	92	5
Gemmatimonadetes	86	5
Holophagae	18	0
KD4-96	34	1
MLE1-12	2	0
Nitrospira	4	1
OPB35	1	0
Opitutae	47	0
Phycisphaerae	45	0
Planctomycetacia	60	1
S085	9	0
SHA-109	1	0
Spartobacteria	3	0
Sphingobacteria	268	4
TK10	1	0
Thermomicrobia	1	0
WD272	16	2

Table A.20: Representation of orders in central taxa of MD networks

Order	Number of taxa	Number of central taxa
32-20	11	0
Acidimicrobidae	136	7
Acidobacteriales	659	17
Actinobacteridae	633	25
Anaerolineales	1	1
Bacillales	25	1
Burkholderiales	164	6
Candidatus Xiphinematobacter	1	0
Caulobacterales	100	2
Chloroflexales	5	0
Clostridiales	1	0
DA101	2	0
Fibrobacterales	3	0
Flavobacteriales	1	0
GR-WP33-30	24	1
Gemmatimonadales	86	5
MB-A2-108	2	0
Methylophilales	1	0
Myxococcales	126	4
Nitrosomonadales	50	2
Nitrospirales	4	1
Opitutales	47	0
Planctomycetales	60	1
Pseudomonadales	12	0
Rhizobiales	704	27
Rhodobacterales	3	0
Rhodospirillales	494	17
Rubrobacteridae	295	12
SC-I-84	26	0
SJA-36	1	0
Sphingobacteriales	268	4
Sphingomonadales	41	1
TRA3-20	19	0
WD2101	39	0
Xanthomonadales	76	5
iii1-8	4	0

Table A.21: Representation of phyla in central taxa of SBS networks

Phylum	Number of taxa	Number of central taxa
Acidobacteria	156	24
Actinobacteria	169	36
Bacteroidetes	8	1
Candidate division TM7	1	0
Chloroflexi	35	11
Cyanobacteria	5	1
Firmicutes	2	0
Gemmatimonadetes	10	2
Planctomycetes	1	0
Proteobacteria	366	54
Verrucomicrobia	1	0
WCHB1-60	1	0

Table A.22: Representation of classes in central taxa of SBS networks

Class	Number of taxa	Number of central taxa
Acidobacteria	138	22
Actinobacteria	169	36
Alphaproteobacteria	261	39
Bacilli	2	0
Betaproteobacteria	70	10
Chloroplast	1	0
Deltaproteobacteria	11	3
Gammaproteobacteria	23	2
Gemmatimonadetes	10	2
Holophagae	15	2
KD4-96	34	11
Opitutae	1	0
Phycisphaerae	1	0
RB25	3	0
Sphingobacteria	8	1
WD272	4	1

Table A.23: Representation of orders in central taxa of SBS networks

Order	Number of taxa	Number of central taxa
32-20	15	2
Acidimicrobidae	32	3
Acidobacteriales	138	22
Actinobacteridae	92	22
Bacillales	2	0
Burkholderiales	31	5
Caulobacterales	12	1
Desulfuromonadales	2	0
GR-WP33-30	8	3
Gemmatimonadales	10	2
MB-A2-108	1	1
Myxococcales	1	0
Nitrosomonadales	12	2
Opitutales	1	0
Rhizobiales	134	20
Rhodospirillales	103	15
Rubrobacteridae	42	10
SC-I-84	14	3
Sphingobacteriales	8	1
WD2101	1	0
Xanthomonadales	23	2