Report: bipartite clustering analysis for SSWD using hierarchical SBMs

B. K. M. Case

May 2022

- We may view the abundance of taxa within samples as a weighted bipartite network, where samples are connected to taxa. A natural goal is to find higher level patterns, or *communities*, within this network.
- There have been several applications of bipartite community detection in microbial data. Each of these has sought to identify clusters of highly connected sample-taxa pairs, typically by attempting to maximizing the network's modularity [6, 1].
- We take a different approach, which finds high-level patterns of abundance based only on their ability to replicate the data. This avoids assuming modularity exists *a priori*, while finding richer community structure on empirical datasets [**?**].
- This approach allows us to find clusters of OTUs which behave similarly, i.e. have similar patterns of abundance among the corresponding clusters of samples.

1 Methods

To identify taxonomic communities and other high-level patterns of abundance, we used Bayesian stochastic blockmodeling, a principled approach to network clustering which finds statistically significant partitions in a network. First, a bipartite network was constructed by assigning seastar samples and OTUs to separate node groups. Samples and taxa were then connected with an edge if the taxa was present, with an edge weight of $\log(a_{ij})+1$, where a_{ij} is abundance of taxa j in sample i. OTU abundance was aggregated to the species level, and taxa with total abundance less than 100 across all samples were removed.

Samples and OTUs were assigned to hierarchical groups following the hierarchical stochastic block model (hSBM), with the maximum *a posteriori* estimate found using a specialized Markov chain Monte Carlo algorithm [5]. We used the "degree-corrected" variant of the model [3], since it had a smaller minimum description length than the simpler non-corrected model (Figure S TODO). Analyses were performed using `graph-tool` version 2.44 [2].

2 Results

The Bayesian stochastic block model analysis revealed a heterogeneous network structure,

3 Discussion

4 Supplemental Methods

Given an abundance matrix $\boldsymbol{A}$, where a_{ij} is the abundance of taxa j in sample i, consider a partition of $\boldsymbol{b} = \{b_i\}$ blocks, where node i is assigned a block $b_i \in \{1, \ldots, B\}$. The probability of observing $\boldsymbol{A}$ is written

$$P(\boldsymbol{A} \mid \boldsymbol{\theta}, \boldsymbol{b}), \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of additional parameters controlling how the partition forms the network structure [4].

We are interested in inverting this process, i.e. finding the most likely partitions given our abundance data. Using Bayes' rule, the evidence for a particular partition $\boldsymbol{b}$ is the posterior probability

$$P(\boldsymbol{b} \mid \boldsymbol{A}) \propto P(\boldsymbol{A} \mid \boldsymbol{b}) P(\boldsymbol{b}), \tag{2}$$

where

$$P(\boldsymbol{A} \mid \boldsymbol{b}) = \int P(\boldsymbol{A} \mid \boldsymbol{\theta}, \boldsymbol{b}) P(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3}$$

is the marginal likelihood.

The posterior probability Eq. (2) can be used to identify partitions which accurately describe the abundance data. For example, the most likely partition is the one that maximizes Eq. (2), i.e.

$$\boldsymbol{b}^* = \text{argmax}_{\boldsymbol{b}} P(\boldsymbol{b} \mid \boldsymbol{A}). \tag{4}$$

Because

References

[1] F. Massol, E. Macke, M. Callens, and E. Decaestecker. A methodological framework to analyse determinants of host–microbiota networks, with an application to the relationships between Daphnia magna's gut microbiota and bacterioplankton. *Journal of Animal Ecology*, 90(1):102–119, 2021.

[2] T. P. Peixoto. The graph-tool python library. *figshare*, 2014.

[3] T. P. Peixoto. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, Jan. 2017.

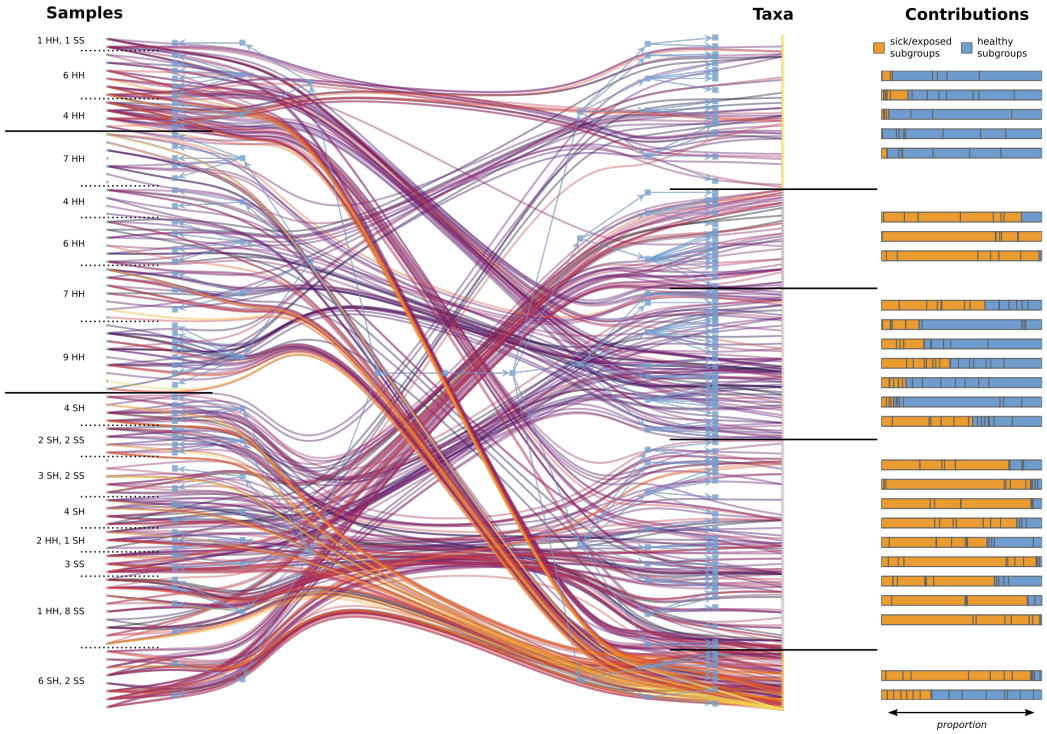

Figure 1: Caption

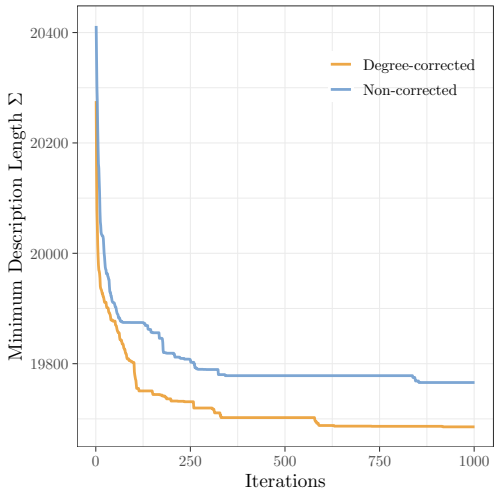

Figure 2: Minimum description length (nats) of two proposed models, for each iteration during MCMC fitting.

[4] T. P. Peixoto. Bayesian Stochastic Blockmodeling. In *Advances in Network Clustering and Blockmodeling*, chapter 11, pages 289–332. John Wiley & Sons, Ltd, 2019.

[5] T. P. Peixoto. Merge-split Markov chain Monte Carlo for community detection. *Physical Review E*, 102(1):012305, July 2020.

[6] C. Zhang and L. Deng. Microbial Community Analysis based on Bipartite Graph Clustering of Metabolic Network. *Journal of Physics: Conference Series*, 1828(1):012092, Feb. 2021.