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 maximize the network's modularity [8, 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 [3].
- 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 [5]. 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 [6]. We used the "degree-corrected" variant of the model [4], since it had a smaller minimum description length than the simpler non-corrected model (Figure 3). Additional

details on the model may be found in the Supplemental Methods. Analyses were performed using `graph-tool` version 2.44 [2].

2 Results

The best fit partition from the Bayesian hSBM analysis is shown in Figure 1, demonstrating a highly heterogeneous community structure. At the lowest level in the hierarchy, there were 103 groups for the 264 OTUs, and 56 groups for 85 samples, which shows a high level of model complexity was required to explain the data.

The group hierarchy clearly distinguished the three health conditions of the samples. At the highest level there were three sample groups, one of which contained all the impacted samples except one, and only three healthy individuals. The next level then tended to separate exposed and sick samples, with 50% of sick individuals falling into a single group (Figure 4).

The hierarchical partitions also allowed us to identify regions of high and low diversity in the data, which helps clarify communities driving the trends in α and β diversity observed in Figures [currently Figs 1D and 1E in the Google Doc][1]. Figure 2 shows the Shannon-Weaver diversity of each sample/OTU, grouped according to their level 1 block membership. Sample blocks containing only healthy individuals tended to have lower entropy, although the three blocks

We found the average diversity between blocks was significantly different than random (Welch's one-way test; $F = 17.3$, $p < 0.001$ for sample blocks; $F = 29.7$, $p < 0.001$ for OTU blocks). Among sample blocks, those belonging to the middle level 2 block of Figure 1 had notably lower diversity compared to the other blocks.

3 Discussion

- The two blocks in the lowest-right corner of the Figure 1 contain taxa which were highly present in all samples (lowest block) and sick/exposed samples (block just above). The taxa in these blocks probably come up elsewhere in the Discussion, e.g. in the paragraphs on *Spirochaetaceae* and *Vibrionaceae*, and can be discussed there.

- Figure 5 shows the evenness, a normalized form of diversity [7]. Interestingly, while the evenness among sample blocks remains significant ($p < 0.001$, it is no longer so for OTU blocks ($p = 0.13$), although there is high variance in evenness within blocks. This suggests that a lack of diversity in some OTUs can be attributed to appearing in fewer samples than others, rather than heterogeneity in levels of abundance for the samples in which they appear.

[1] the relationship to β diversity isn't exact since this is based only on abundance and not phylogenetic distance

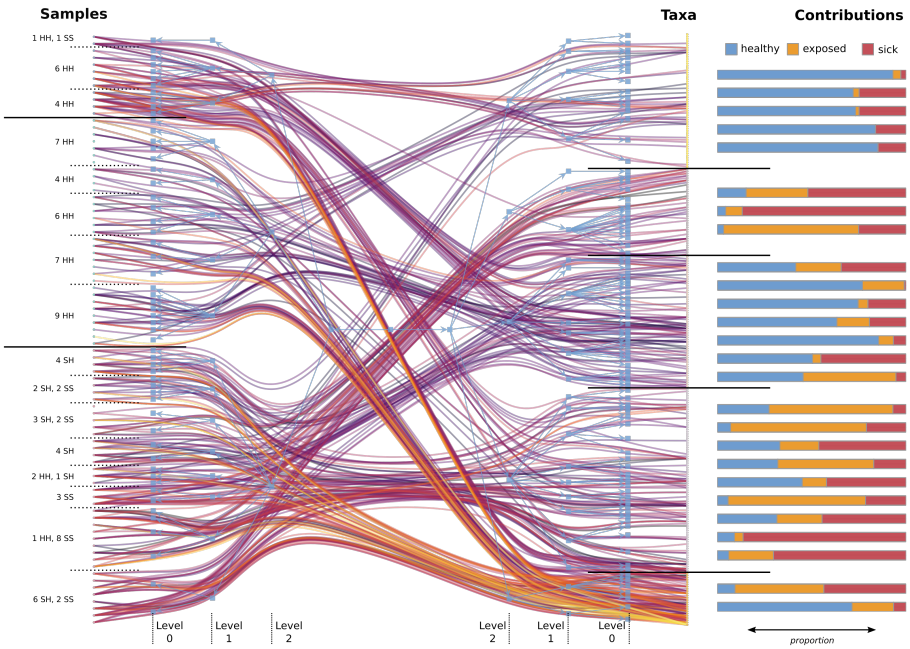

Figure 1: Maximum *a posteriori* partition found in the bipartite clustering analysis. The hierarchical partition is shown in blue, while a random sample of 400 links from the original abundance network is shown ranging from purple (lowest abundance) to orange (highest abundance). Levels of the partition tree have been labelled as they are referenced in the text. Far right: the proportion of abundance contributed from samples of each health status, for each level 1 taxa group. HH=healthy, SH=exposed, SS=sick/wasting.

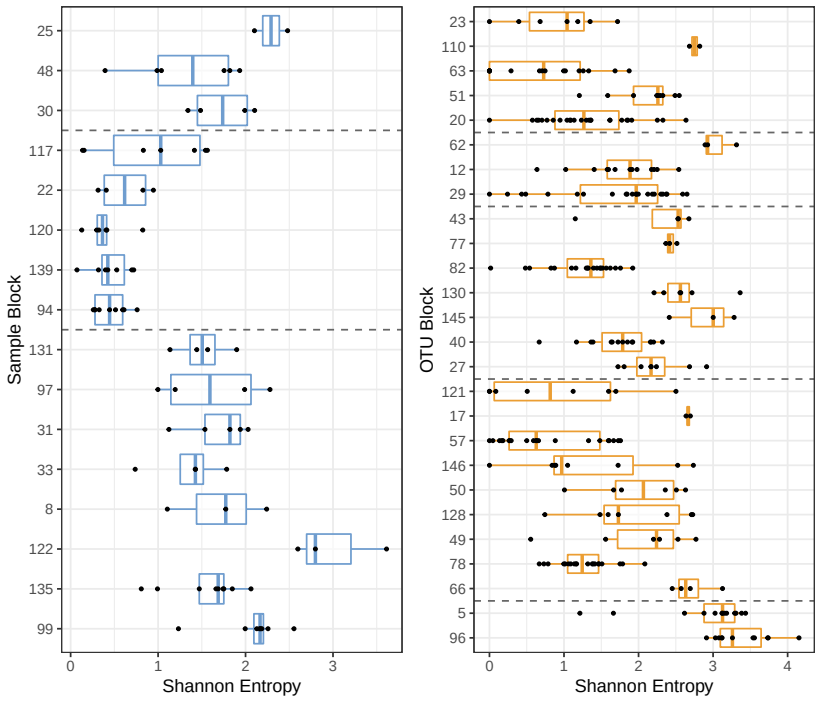

Figure 2: Shannon–Weaver diversity within blocks found in the bipartite clustering analysis. Each point is a sample/OTU, and the y-axis indicates the level 1 block membership of each point. Dashed lines indicate membership in the highest level (level 2) blocks.

- The three samples in block had a higher diversity than all other samples, and all three were sick. They may be interesting to look into these three samples further.
- The OTUs with the lowest non-zero evenness are the most "complex" in the sense that when they do appear, their relative abundance can vary greatly. We could pick the top ten of these, compare their abundance among health groups, see if they appear elsewhere in the manuscript, etc.
- A limitation in this analysis is that the abundances are taken "as-is," rather than corrected for bias like in the differential abundance and correlation analyses. This would be non-trivial to implement but an opportunity for future work.

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 [5].

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.

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

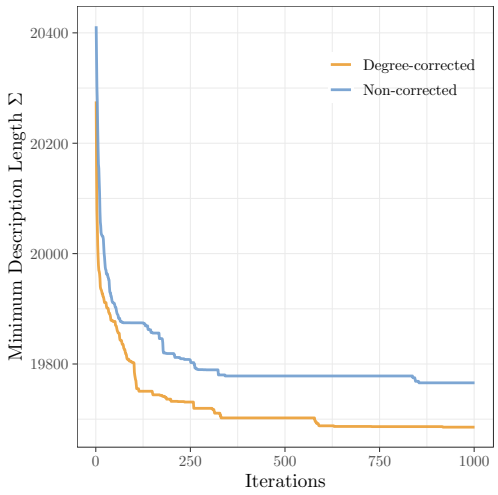

Figure 3: Minimum description length (nats) of two proposed models, for each iteration during MCMC fitting.

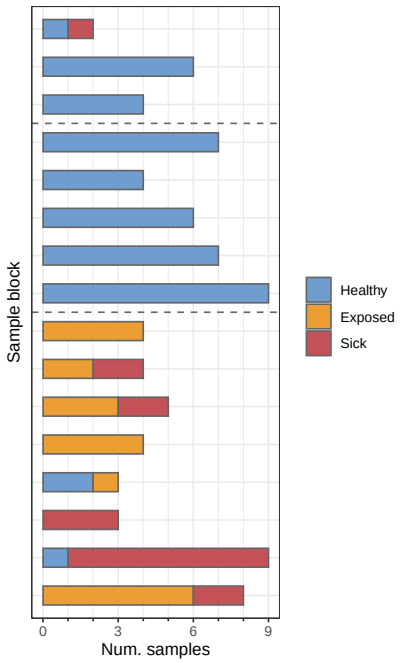

Figure 4: Number of samples within each sample block from the bipartite clustering analysis, at grouping level 1. Counts are separated by the health status of the sample. Sample blocks are ordered corresponding to the order on the left side of Figure 1. Dashed lines indicate membership in the three highest level (level 2) sample groups.

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. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047, Mar. 2014.

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

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

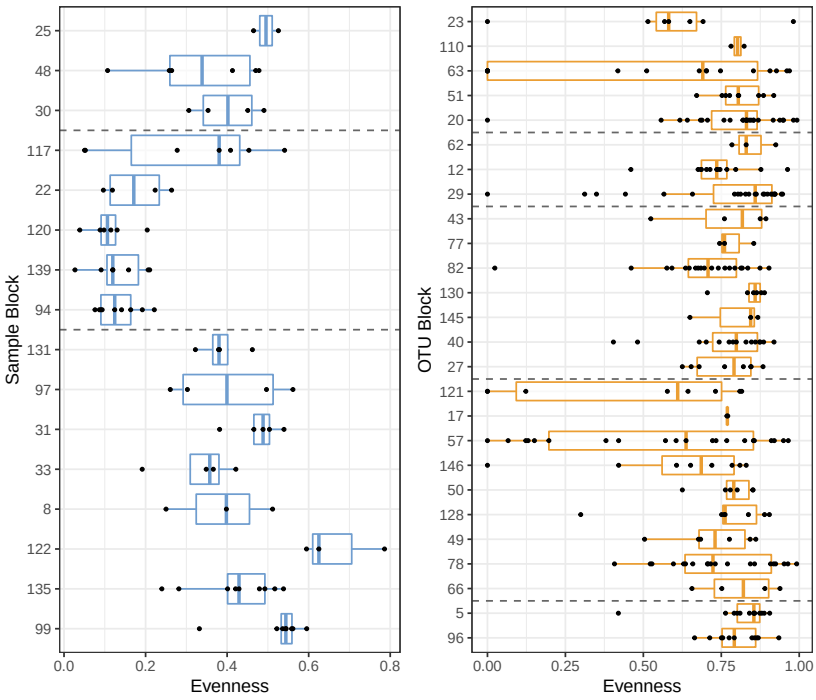

Figure 5: Pielou's evenness index [7] within blocks found in the bipartite clustering analysis. Each point is a sample/OTU, and the y-axis indicates the level 1 block membership of each point. Dashed lines indicate membership in the highest level (level 2) blocks. Evenness is defined $H/\log S$, where S is the number of OTUs/samples present in the row/column and H is the Shannon-Weaver diversity.

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

[7] E. C. Pielou. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144, 1966.

[8] 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.