SpeSpeNet 1.0 manual

Table of Contents

[Abstract 1](#_Toc160105958)

[Dependencies 1](#_Toc160105959)

[1. The navigation bar 2](#_Toc160105960)

[2. Uploading and processing data 2](#_Toc160105961)

[3. Coloring and changing aesthetics of the network 5](#_Toc160105962)

[4. Downloading and interactivity 7](#_Toc160105963)

[5. The summary tab 8](#_Toc160105964)

[5. The raw data tab 11](#_Toc160105965)

## Abstract

Correlation networks are a commonly used method for untargeted exploration of microbiome data. In these networks, nodes are taxa and edges represent correlations between the relative abundances of taxa higher than a certain cutoff. Clusters of correlating taxa (co-response groups) are often indicative of a shared response to environmental drivers. Currently, most tools for network-visualizations of microbiome data either require the researcher to have coding skills or are not user-friendly in terms of the time expenditure and customizability of the networks. Furthermore, existing tools lack focus on the relationship between environmental drivers and the structure of the microbiome, even though many edges in correlation networks can be understood through a shared relationships of two taxa with the environment. For these reasons we developed SpeSpeNet (Species-Species Network), a practical and user-friendly R-shiny tool to construct correlation-networks from count tables. The details of data preprocessing, construction of the network and the network visualization are automated, require no programming ability for the web version, and are highly customizable. Furthermore, SpeSpeNet detects clusters of co-occurring microbes and visualizes the relationship between these clusters or individual taxa and environmental metadata. This document outlines the functionalities of the SpeSpeNet tool.

## Dependencies

Using the web-version of SpeSpeNet does not require any programming on the user’s side. If the user wants to use the development version, they should have R and Rstudio installed. The following R libraries need to be installed:

1. shiny
2. shinythemes
3. shinycssloaders
4. tidyverse
5. tidygraph
6. igraph
7. ggraph
8. ggdark
9. RColorBrewer
10. scales
11. plotly
12. phyloseq
13. shinyBS
14. MGnifyR
15. tools
16. pals
17. ggiraph
18. htmlwidgets
19. reshape2

Running the Launch\_tool.R script will launch SpeSpeNet.

## 1. The navigation bar

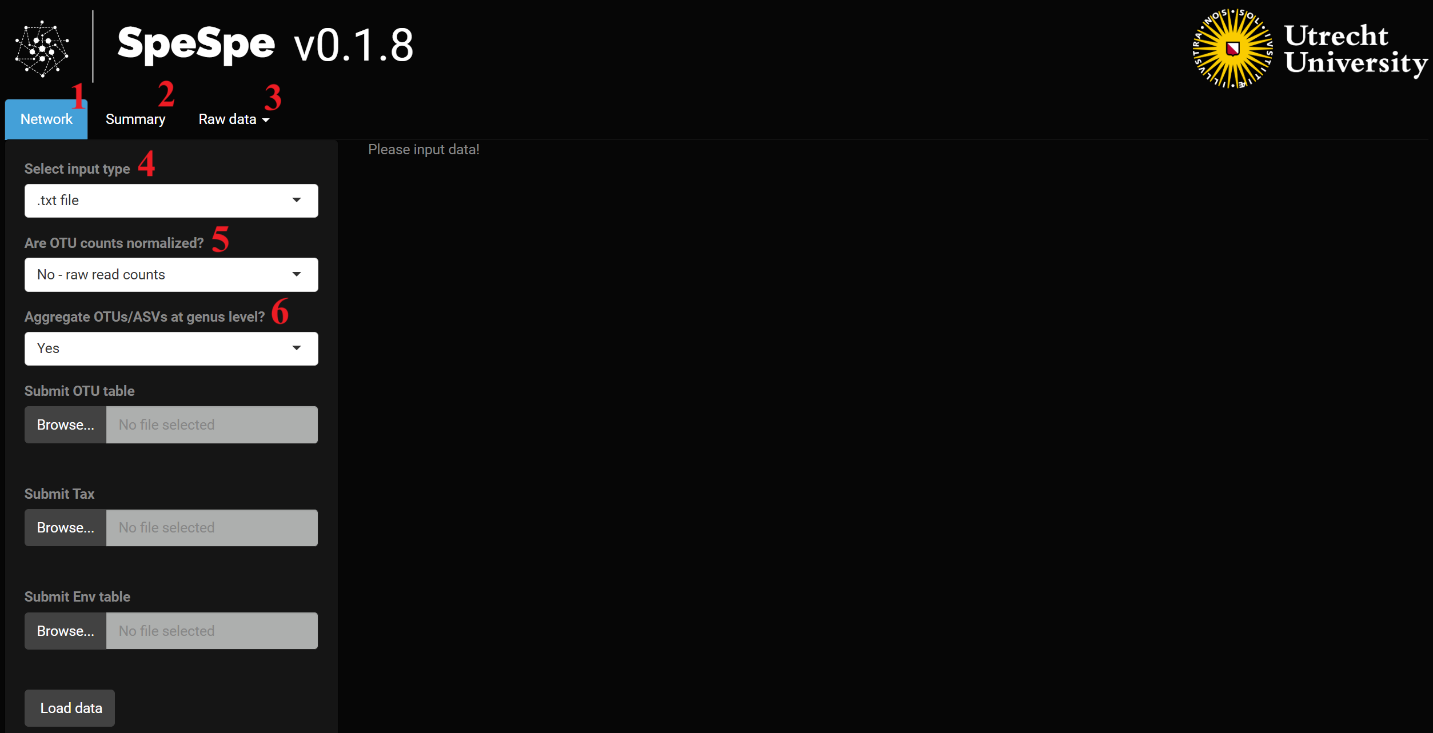
The navigation bar is found at the top of the screen (panels 1,2 and 3 in **Figure 1**). The navigation bar can be used to switch between the three tabs in SpeSpeNet. The “Network” tab is where data is uploaded and the network visualization is shown. The “Summary” tab shows a summary of the network in terms of the taxonomic distributions and the relationship between co-response clusters/environmental categories and environmental drivers (see section 5: The summary tab). Finally, the user can download .txt or HTML files of the network data in the “Raw data” tab.

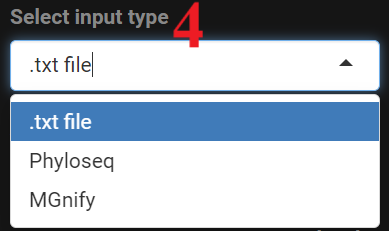
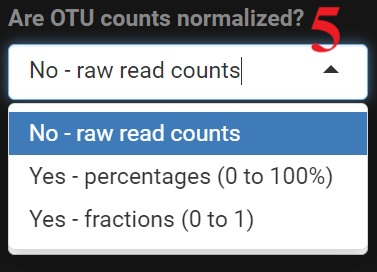
## 2. Uploading and processing data

After loading SpeSpeNet the network tab is shown by default. Here, the user can input data, either as phyloseq objects (in a single .rds file) or as three separate .txt files that contain the abundance table, the taxonomy table, and the environmental table (menu 4 in **Figure 1**). The data can be input as raw read counts, percentages, or fractions (menu 5 in **Figure 1**). The .txt files need to be formatted in a standardized way: They should be tab-separated, the OTU table should have operational taxonomic units as rows and sample identifiers as columns, and the metadata file should have samples as rows and variables as columns. Furthermore, the row names (containing OTU/ASV IDs) should match between the abundance table and the taxonomy table and be in the same order. Finally, the column names of the abundance table should match the row names in the environmental data and be in the same order. The column names of the taxonomic data should be taxonomic levels (such as phylum, class, order, family, genus and species). To write row name and column names to a .txt files, the following R code can be used:

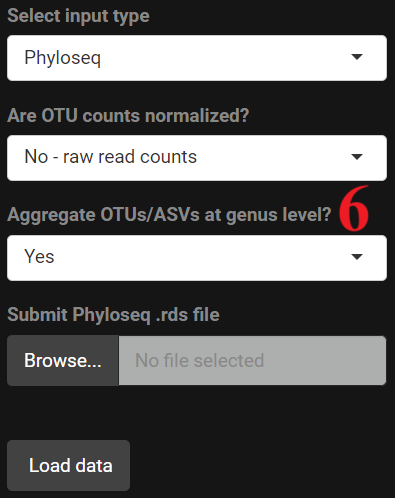
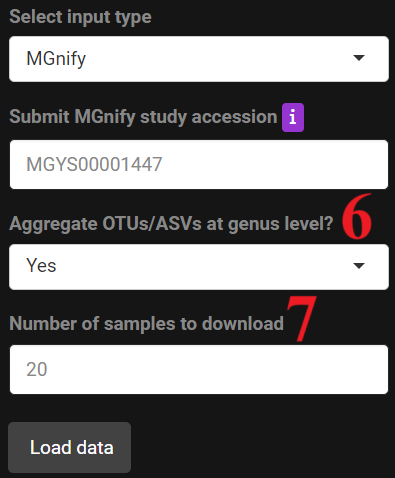
write.table(matrix, "file.txt", sep = "\t", col.names = T ,row.names=T, quote = T).

Phyloseq objects must adhere to the same standards as the .txt files. SpeSpeNet can also download studies from the MGnify database if provided with a MGnify study accession number. The number of samples that will be downloaded from MGnify can be limited (menu 7 in **Figure 2**) in case downloading the whole study takes too long. Finally, the user can choose to aggregate the OTUs/ASVs at the genus level (menu 6 in **Figures 1 and 2**). This is recommended for 16S datasets from diverse habitats like soil to constrain the number of nodes.



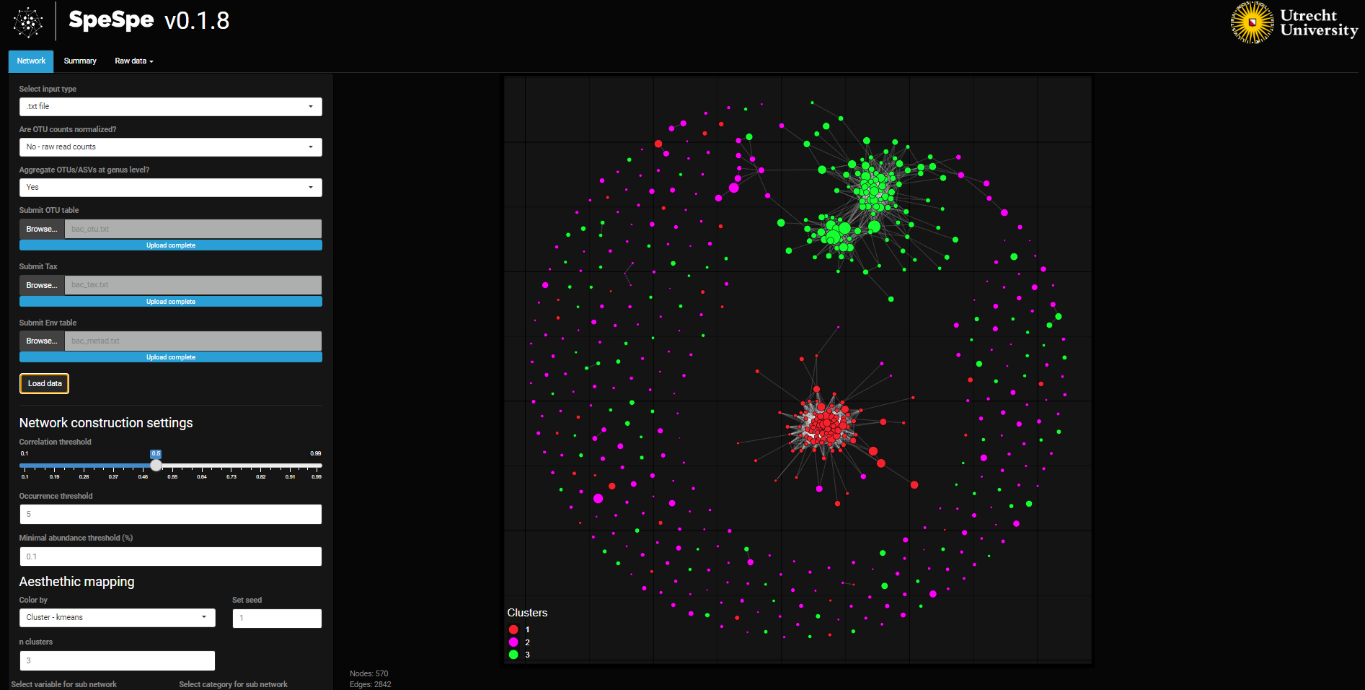
 

**Figure 1**. SpeSpeNet input options for .txt files

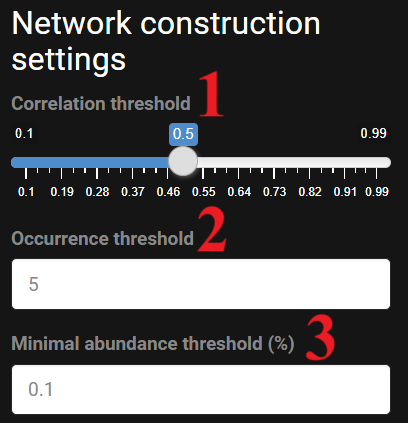
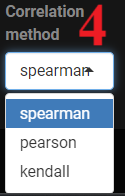
 

**Figure 2**. Input options for phyloseq objects and MGnify study accession numbers.

After the data upload is finished SpeSpeNet constructs a co-occurrence network based on the default settings (**Figure 3**). We recommend the user change the default settings to suit their dataset.These include a threshold on the minimum number of samples a taxon should occur in to be included in the network (box 2 in **Figure 4**), and a threshold on the minimum relative abundance a taxon should have in at least one of the samples (box 3 in **Figure 4**). After filtering, the correlation matrix of the relative abundances of the taxa is calculated and used for network construction. The correlation method (Pearson, Spearman or Kendall) can be chosen by the user (menu 4 in **Figure 4**). We recommend a rank-based correlation due to the compositional nature of amplicon data. The correlation cutoff used to define edges can also easily be changed by the user (slider 1 in **Figure 4**).



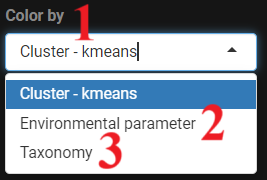
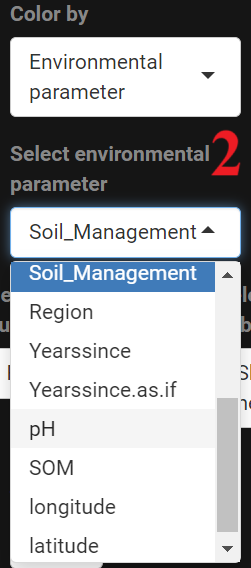
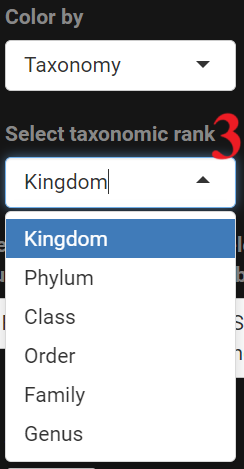
**Figure 3**. A co-occurrence network based on the default settings (data from Rijssel et al (https://doi.org/10.3389/fmicb.2018.00294, Study S1 in the original paper)

**Figure 4**. Network construction settings and correlation method options.

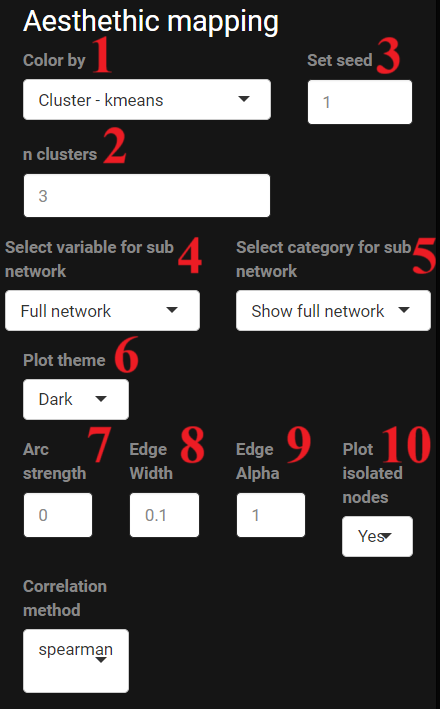
## 3. Coloring and changing aesthetics of the network

The network can be colored with three different variables: 1) k-means clustering, 2) taxonomy and 3) relationship with an environmental variable (menu 1 in **Figure 5**). In case 1) the k-means algorithm is applied on the correlation matrix of dimensions taxa x taxa. The user can specify the number of clusters made by the k-means algorithm (box 2 in **Figure 5**). The network is then colored according to the clustering. In case 2) the network can be colored on taxonomy at a chosen taxonomic level (menu 3 in **Figure 5**). SpeSpeNet will show the 11 most abundant clades and lump the others together as “Other” to constrain the size of the legend. Finally, in case 3) the network is colored on an environmental variable in the metadata (menu 2 in **Figure 5**). If a numeric environmental variable is selected (e.g. pH), the color of the nodes represents the correlation of the relative abundance of the taxa to the environmental variable. If a categorical variable is selected (e.g. healthy/diseased or sand/clay), SpeSpeNet will color on the category in which the taxon has the highest mean relative abundance.

**Figure 5.** Options to color the network.

The aesthetic parameters of the network can be tweaked according to the user’s preference (**Figure 6**). The different options are found underneath the network construction options. Aesthetic parameters of the network that can be altered are the background theme, arc strength of edges, edge width, edge transparency, and whether isolated nodes should be plotted (boxes 6-10 in **Figure 6**). The random seed (box 3, **Figure 5**) changes the layout of the network and is useful e.g. if the network overlaps with the legend. The seed will not change the nodes, edges or clustering of the network.

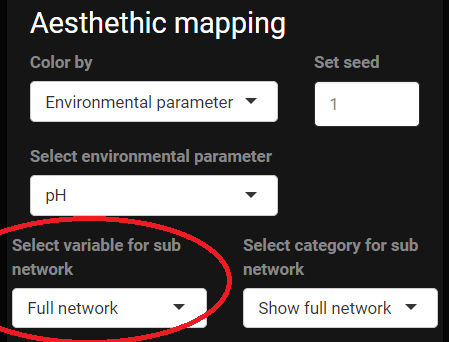
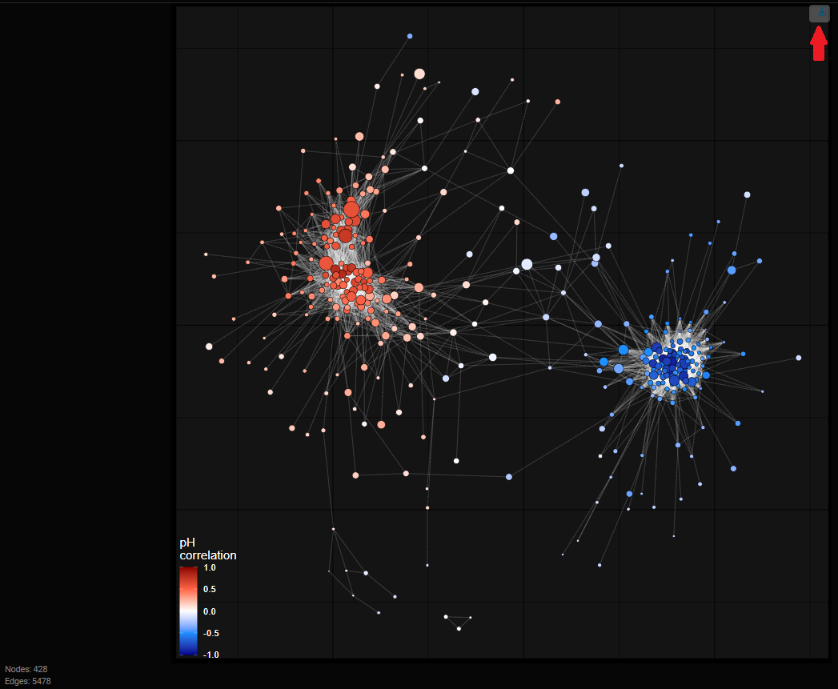


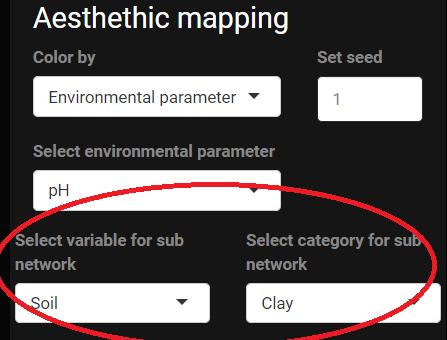
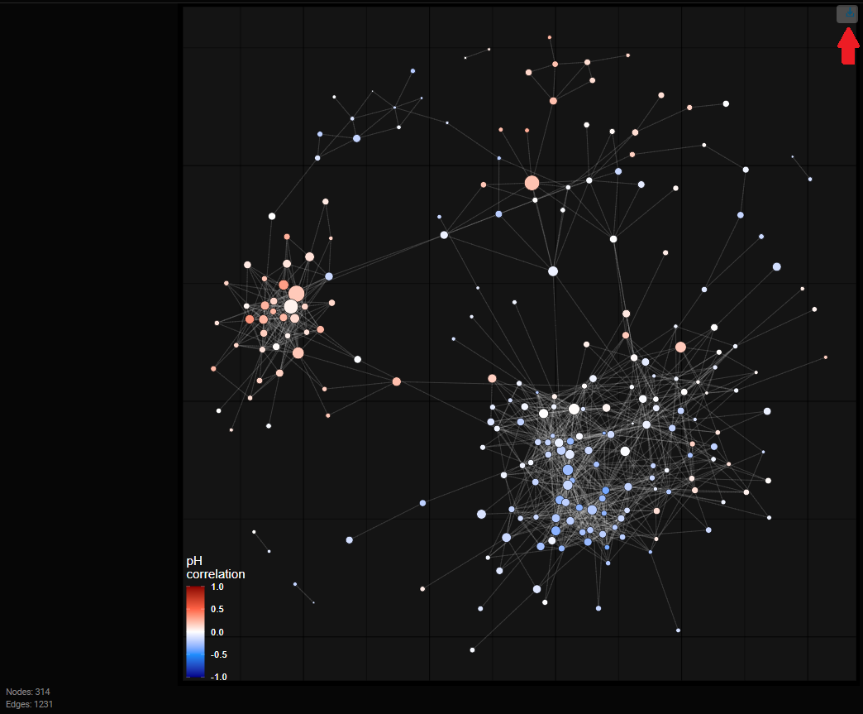
**Figure 6.** Aesthetic mapping options

If there are categorical environmental variables, the user can zoom in on a specific category (menu 4 in **Figure 6**). This is only possible for categories with at least eight samples. A new network is then constructed using only the samples in the chosen category. As an example, we show from Rijssel et al (https://doi.org/10.3389/fmicb.2018.00294, Study S1 in the original paper) a network constructed using the entire dataset and a network constructed using marine clay samples only (**Figure 7**).

## 4. Downloading and interactivity

The network can be downloaded as a png with the button in the top right corner (upwards pointing arrow in **Figure 7**). The network is interactive, meaning that hovering the pointer over a node will highlight the taxonomy (**Figure 8**)**.**

**Figure 7**. Contrast between the full dataset network and the subnetwork constructed using only samples from clay soils.

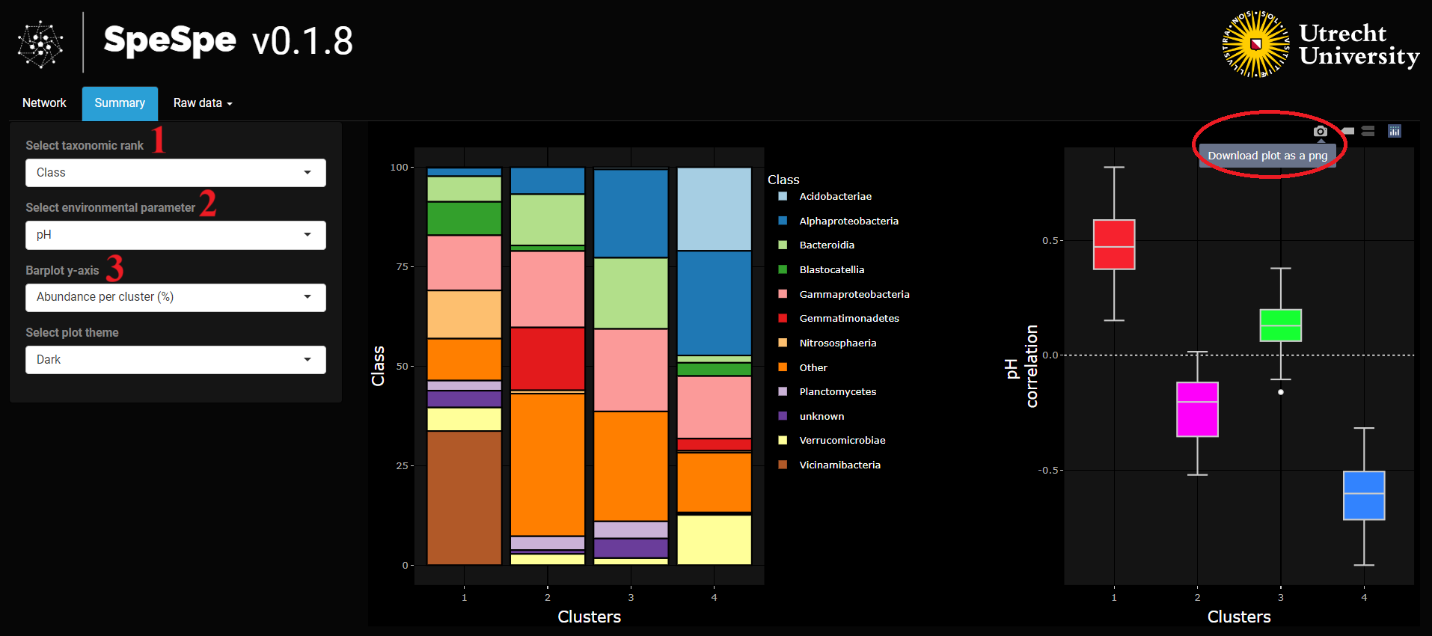
A black background with white text and red and blue dots

Description automatically generated

**Figure 8.** Hovering the pointer over a node shows the taxonomic classification.

## 5. The summary tab

The second tab right next to the “Network” tab is the “Summary” tab. This tab contains a bar plot and a box plot (**Figure 9**).

**Figure 9.** Overview of the Summary tab

If the network is colored on k-means clustering or a categorical environmental variable, the bar plot shows the taxonomic composition of each individual cluster or category (increasing or decreasing the number of clusters in the network panel has an immediate effect on the summary tab too). If the network is colored on taxonomy or a numerical environmental variable, it shows the overall taxonomic composition of the entire dataset. The user can choose at which taxonomic rank they want the distribution plotted (menu 1 in **Figure 9**) and whether the y-axis shows the mean relative abundance over all samples or per cluster (menu 3 in **Figure 9**) (Compare top panel with bottom panel in **Figure 10**). The bar plot shows the mean relative abundance as a percentage when hovering the mouse over the bar plot (**Figure 10**).

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Figure 10.** The interactivity of the bar plot.

The box plot shows the correlations of the taxa (nodes in the network) per cluster or over the entire dataset with a numeric environmental variable in the metadata (menu 2 in **Figure 9**). Hovering the mouse over a box shows summary statistics of the distributions (minimum and maximum value, median and the two remaining quartiles). The two plots can be downloaded together as a png with the button in the top right (circled in red in **Figure 9**).

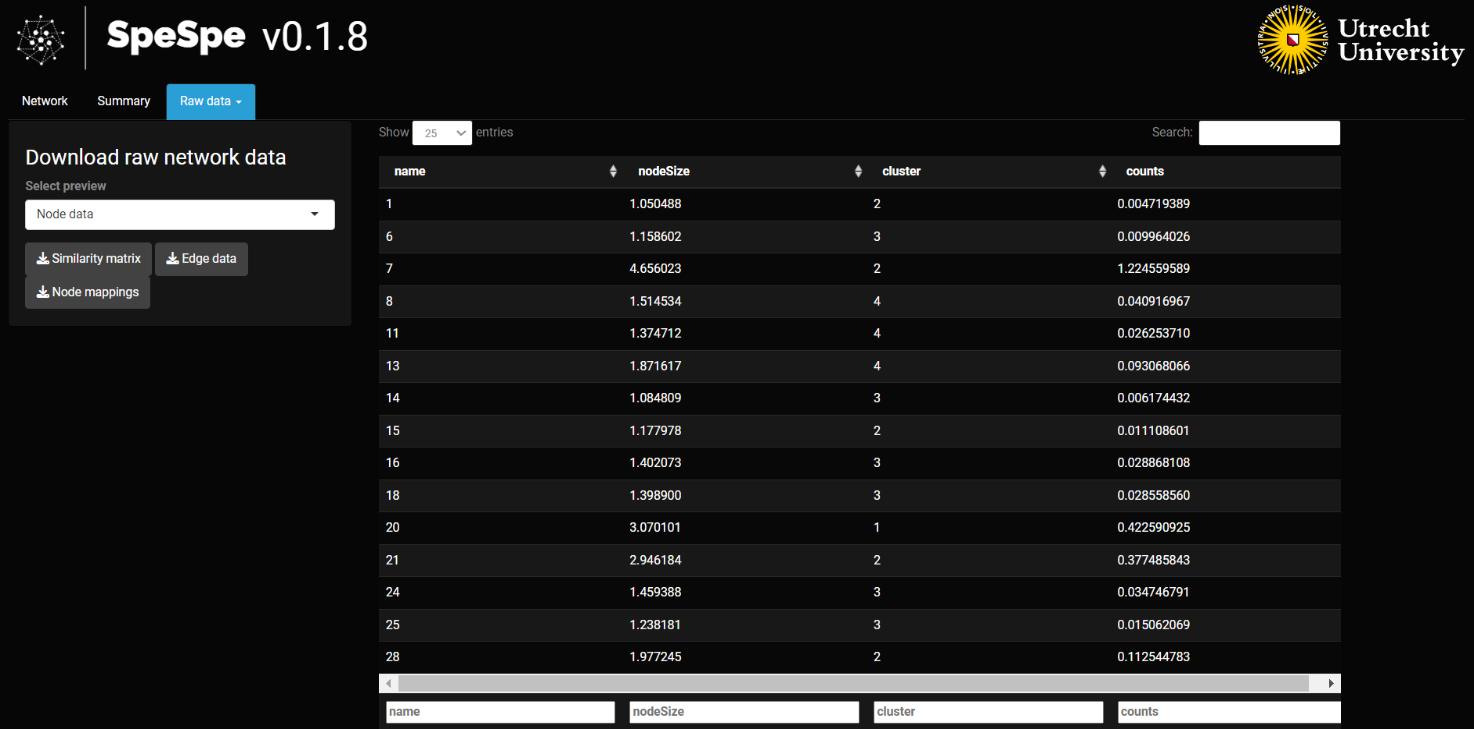
A screenshot of a graph

Description automatically generated

**Figure 11.** The interactivity of the boxplot.

## 5. The raw data tab

The third tab is the “Raw data” tab, from network data can be previewed and downloaded. The user can download the correlation matrix in the form of a .txt file and the node and edge data as html files.

**Figure 12.** Overview of the Raw data tab.