Thresholded Boolean Co-Abundance (ESABO) analysis for continuous-valued datasets

By

Devi Chandran

Student ID: 2430977

Supervisor: Dr. Jens Christian Claussen

A thesis submitted to the University of Birmingham

For the degree of MSc in Data Science

School of Computer Science

University of Birmingham, Birmingham, UK

September 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Communities proliferate in the world of microorganisms and constitute essential components of the environments they inhabit. It is of the utmost importance to comprehend these societies and the complex relationships that link them. Using the well-known Entropy shifts of Abundance Vectors under Boolean Operations(ESABO) approach, our effort goes deep into the investigation of these interactions. To reveal the nature of interactions between microorganisms, this approach uses entropy fluctuations in data on microbial abundance. Our main aim behind the project lies in the generalisation of the ESABO method to account for continuous valued datasets using thresholding of the dataset. We are able to determine the best thresholding settings for various subsets of a microbiome dataset(959 species) owing to this adaption and also examine the effect of thresholding on various data compositionality. Furthermore, we extend the ESABO approach to a novel dataset of abundance of 354 plant species, in order to discover its viability in extracting interactions from ecological data. In order to shed insight on the complex web of interactions between these microbial and ecological communities, we also analyse the network topologies created by the ESABO technique. We get a better understanding of microbial and ecological dynamics through our thorough investigation, opening the door for useful applications and insights in these areas. Our study shows that adapting the abundance threshold opens a way for a better understanding of microbial and ecological dynamics, opening the door for useful applications and insights in these areas.

Keywords: Microbial communities; ESABO; Thresholding; Network topology.

Acknowledgment

I would like to offer my sincere thanks to my supervisor Dr. Jens Christian Claussen for his remarkable support and guidance throughout the period of my dissertation. The endeavour would not have been possible without the constant feedback and insightful comments provided by him. His interest towards science and research was highly motivating to witness and it was a pleasure to work under him.

Next, I would like to extend my special thanks to my parents, sister and my dear friends for their unwavering support, prayer and understanding while undertaking my research. Their constant motivation is what led to the accomplishment of the project.

Above all, I would like to thank the Almighty for always being a guiding light in times of stress and difficulties. You are the one who made me believe in myself and led to the completion of this project.

Chapter 1: Overview

1.1 Introduction

Microscopic life forms, including bacteria, fungi, viruses, and other microorganisms, have a notable impact on diverse ecosystems [1][2]. In recent years, the advancement of high-throughput sequencing technologies has revolutionized our understanding of microbial communities and their role in various ecosystems. Microbial abundance networks have emerged as a powerful tool to elucidate the complex and intertwined interactions among microorganisms within these communities. Experimental and computational methodologies to comprehend interactions in microbial communities and forecast their reaction to disturbances, have also been developed[3].

The clinical relevance of microbial analysis is gaining importance [4][6]. One breakthrough in the study of microorganisms has been in the researching of co-abundance patterns among different species. A study indicated that there were variations in co-abundance patterns among members of oral, intestinal and pancreatic bacterial microbiomes among individuals with pancreatic cancer and other gastrointestinal disorders[5]. Many such studies pertained to the substantial correlations between microbiome associations and diseases such as cancer and autism spectrum disorder[[8]-[13]]. Thus, studying microbiome associations can be of high significance for human health.

A number of researchers have tried studying microbial communities in the past. But, recent advancements in 16S rRNA sequencing technologies have made it easier for the analyses of these communities through recognition of individual species and its taxonomical classification. In order to dive deeper into the recognition of the structure and function of microbial communities, it is vital to identify the network of interactions between these communities[15]. Scientists have used the information of co-occurrence of these species to draw a parallel with the complex communities and to study the ecological relationships among them.

The abundance information of the microbiomes can help recognise the species present in the community. The difficulty lies in analysing how the interactions take place with the help of the abundance information. It is quite understandable that species interactions impact their corresponding abundances. Correspondingly, there were several studies which involved inferring interactions among the microbes through the help of high throughput genome sequencing.

In the light of this context, there is a specific study[16] which engaged in network inference by studying the entropy shifts of abundance vectors under Boolean operations(ESABO) approach to human gut microbiome abundance data. ESABO approach deals with the application of Boolean AND operation to binary abundance vectors for calculating the shift in the entropy of these abundance vectors under random permutations of the second vector. This method has been evolutionary in comparison with other microbial interaction network approaches due to its success in finding the associativity of low abundant microbes of the human gut. In this way, it unveils the concealed networks in the dataset which would go unnoticed under other methods.

ESABO stands as a significant advancement within the realm of computer science, offering a unique fusion of concepts from Boolean Algebra, logical operators, and Shannon's information entropy[38]. This innovative approach expands the horizons of data analysis, enabling researchers to delve deeper into the intricate web of relationships inherent within datasets and also the topology of the data.

According to [17], the randomness of the ESABO approach’s results can be reduced to increase the computational efficiency of the approach with an analytical formula as will be discussed in the background.

1.2 Motivation

Our project will involve the ESABO approach[16] and modified ESABO approach[17] in a systematic manner. Both the approaches input binary abundance vectors for co-abundance analysis. Below are the motivations of the current study:

- Extending the ESABO approach to continuous valued abundance datasets. The current ESABO approach assumes a threshold of 1(i.e, absence-0, presence-1). We are interested in investigating whether this is the optimal threshold for the data. This investigation further motivates us to binarize the data using an optimal threshold value to attain maximum information from the data. As a result, we are able to capture the threshold which can lead to larger number of links in the data.

- The ESABO approach with the analytical formula has been stated to be more computationally efficient compared to the original ESABO in [16]. We are interested in investigating this to a specific case scenario and comparing between the methods to find the best one for application in our project.

- Thirdly, we apply the ESABO approach to a different scenario of application in ecological framework by applying it to an abundance data of the plants. In this context, we aim to uncover the valuable the insights that can inform ecological research and conservation efforts.

- We are also interested in analysing the complex networks attained as a result of the ESABO approaches. The analysis of these inferred networks help in its evaluation through network complexity measures and also comprehends whether the optimal threshold value can actually result in inference of denser networks in the data.

1.3 Research questions

Our study poses certain research questions which can be critical for our understanding. Here are a few research questions that needs to be addressed:

- Which model(ESABO or modified ESABO) leads to a better network creation? Are there significant differences between the networks attained by both?
- How do the differences in data compositionality lead to a difference in threshold values?
- Does the increase in sample size lead to better performance and more significant links?
- Which are the significant links derived through the network inference methods?
- Which are the keystone species driving the community?

1.4 Methodology

The research has been broadly divided into two subsections where we will looking extensively into two different datasets of interest: (i) microbiome dataset and, (ii) plant species abundance dataset. The study initiates with a comparison of the original and modified ESABO approach in order to adopt the best approach as the primary method for our investigation. It proceeds with the thresholding of different subsets of the microbiome data and analysis of the research questions on the same. All key insights regarding the project will be derived in this section. The application of ESABO approach is extended to the plant abundance dataset in the later part of the study.

This chapter introduces the motivations behind the project by providing a small background of the entire setting of the project. It also lays down several research questions which are crucial to be answered for our project. The rest of the project is outlined as follows:
Chapter II will encompass the literature review of the project, diving deeper into the other researches in the field and giving summarised review of the approaches along with outlining the importance of the ESABO approach. Chapter III introduces the Background of the project and details the key methodologies, definitions and equations used in the projects' building. Chapter IV discusses the application of ESABO method to microbial abundance data and contains two subsections: Section 1 introduces the methods used for application on microbial abundance data, including the methods introduced for answering the research questions and the criteria used for testing the methods and Section II introduces the results obtained after the application of the methods and consequent evaluation of the results. Similarly, Chapter V discusses the application of ESABO on plant species abundance data and contains two subsections: Section I describes the different methods on application and testing on the plant abundance data and Section II outlines the results of the Section I and its evaluation. Chapter VI contains the discussion on the key results obtained from the analysis on Chapter IV and Chapter V and also outlines the limitations and future possibilities of work. Chapter VI also concludes the research with a glimpse of the results obtained.

Chapter 2: Literature Review

Introduction

Recent years have seen an increase in interest in microbial interaction networks, and numerous studies have been carried out to comprehend the intricate patterns of microbial interactions. Different methods including correlation-based and other techniques have been developed to study these interactions. In this section, we will be discussing in detail about the different network inference available similar to the ESABO method.

2.1 IMPARO: Inferring microbial interactions through parameter optimisation

IMPARO[20] method introduces a shift in the usage of statistical methods to usage of biological model, generalised Lotka Volterra Model(GLV) and others for the inference of microbial interactions through parameter optimisation. One advantage of this method is that it widens the possibility of creating multiple microbial interaction networks from similar abundance profiles, by arguing that a unique solution is not always satisfactory(Figure 1).

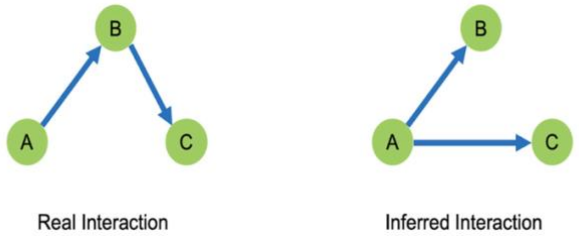

Figure 1: Inferred interaction captures the link A -> B -> C in real interaction as link A -> C[20].

IMPARO method also has an added advantage of involving the rarer OTUs through Monte Carlo approach[21] into the Microbial Interaction Networks(MIN) when other statistical models which fails to capture the effects of such OTUs. However, IMPARO lags at inferring these rarer OTUs because of a lower prediction quality in case of rarer OTUs, thus indicating that, the rarer OTUs cannot well predict the interactions between the microbial species based on microbial abundance profile. ESABO method outperforms IMPARO method in this respect, by focusing on the rarer OTUs, i.e. the low abundance profiles and finding associations among them.

2.2 MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles

MetaMIS [22] is a simulator tool based on the Lotka Volterra Model which can tolerate a large amount of missing data to estimate time series data of microbial community profiles. To infer microbial interactions, it employs a discrete-time Lotka-Volterra model in conjunction with partial least square regression for the estimation of parameters.

The challenging task of extracting interaction networks is made simpler through this method and it has succeeded in reconstructing 27 interaction networks using a human male intestinal microbiota dataset. It can also organise multiple interaction networks into a consensus, as proven in male and female faecal microbiomes.

Although MetaMIS and ESABO approach are carried out using different model assumptions and may be applicable to specific case scenarios, ESABO method is specifically designed to capture the non-linear interactions between microbes which may be ignored by MetaMIS. Also, owing to the simplicity of ESABO approach in using a direct logical AND gate to study the interactions, ESABO method might be more computationally efficient than MetaMIS.

2.3 RMN: Rule-based Microbial Network

RMN [23] introduces its own model where three regulatory OTUs(Ordered Taxonomical Unit) which interact with each other are grouped together across different time points to describe the network. The regulatory OTUs act as the nodes in the network and their interactions between other OTUs are represented by the edges in the network. The algorithm utilises a parametric weighting function for allocating weights to OTU interactions.

It is worth noting that correlations between microbial abundances occasionally fail to suggest microbial interactions. The RMN method aids in the inference of regulatory links between microbial pairs, revealing the true nature of microbial interactions. RMN algorithm has performed exceptionally well in discovering interactions in the gut microbiome and across a range of other environments including infant gut and Antarctic seawater. While the method works in such case scenarios, it might not be suitable to the design case of binary data where only the presence and absence of microbes are taken into account for relatively simpler formation of network structure.

2.4 SPIEC-EASI

SPIEC-EASI[24] makes use of a statistical method that combined with StARS(Stability Approach to Regularization Selection) to form weighted undirected graphical framework that can describe the correlational associations in the data. It exploits StARS method for reducing overfitting of the network and assumes that there are sparse ecological relations in the data.

A sparse network assumption only takes the most important interactions in the network, which helps improve the accuracy of the network and disregards the spurious networks. This is contradictory to the state-of-the-art correlation based techniques such as SparCC(Sparse Correlations for Compositional data) [25] and CCREPE(Compositionally Corrected by REnormalization and PErmutation) [26] which finds both direct and indirect edges or links. SPIEC-EASI also outperforms SparCC and CCREPRE by revealing more consistent and sparser interaction networks in real American Gut Project(AGP) unrecognised by the two methods.

The method gives an idea about the closeness of the OTUs but does not indicate the type of interaction among them. Interestingly, the ESABO method outlines the nature of the interactions between the OTUs and also recognizes whether the type of interaction is synergistic or competitive.

2.5 LIMITS: Landscape of Inferred Microbial Interaction from Time Series

LIMITS [27] method uses the time series of microbial communities to identify abundance patterns of different microbes over time. Leveraging these patterns, a mathematical model is fabricated to represent these interactions, which is then compared with the observed data to predict the nature of the interactions.

Revolving around the central approach of discrete-time Lotka Volterra equations, the algorithm provides intricate dynamics of microbial interactions such as competition or cooperation. These interaction dynamics aid in distinguishing intricate details of the microbial communities. The LIMITS algorithm in majorly used in deducing keystone species in the community and has been applied to simulated data and two human gut samples for testing.

While the LIMITS approach focuses on time series data of microbial abundance for inference of microbial interaction networks, ESABO method is more versatile to the types of data it can analyse by providing insights into the entropy shifts in the abundance vectors.

2.6 Gao et al. [28] method for longitudinal metagenomic data

Gao et al. [28] uses a Lotka Volterra model approach, coupled with a forward step-wise[1] regression with bootstrap aggregation[2] and a simple t-test to select candidate models[3]. A Bayesian information criterion(BIC)[29] filters the candidate models resulting in multiple models which are consolidation to form a resulting model. The procedure has been validated on cheese microbial community data and performs well in describing the dynamics of relative abundances and individual growth.

It has proven to be a versatile model for hypothesis testing and highlights that knowing specific microbial relationships can be vital for microbial studies. The method is also tested on gut microbiota of healthy and type 1 diabetes infants. As a result, a specific conclusion made was that there were more interactions derived in healthy infant gut microbiota compared to type 1 diabetes infants.

This method's known advantage over prior methods is that it has demonstrated to infer biologically confirmed microbial interactions from relative abundance data, something other methods have yet to prove. The method is used for longitudinal metagenomic data and however, ESABO method is much more focused on compositional data and inferring the interactions from them.

2.7 Maximal Information Coefficient(MIC)

MIC [45] is a criterion used for measuring the strength of linear or non-linear relationship between two variables. MIC describes a mathematical function that quantifies the strength of the relationship between two variables or species. It is a statistical measure which requires discrete binning for maximising the mutual information between two variables.

MIC gives a quantitative degree of association with a normalized score between 0 and 1, with 0 indicating no relationship and 1 indicating a functional relationship. The advantage of ESABO method over MIC is that, ESABO method not only determines the functional relationship between two variable but also provides the sign of the interaction between them. A positive z-score indicates mutualistic interaction and a negative z-score is symbolic of a negative interaction.

[1] Forward step-wise regression: Method which iteratively adds the independent variable to create a regression model

[2] Bootstrap aggregation: Forms multiple data subsets through random selection of data points with replacement. The selected points are further trained using regression.

[3] Candidate models: Models created using forward step-wise regression which will act as candidates for model selection based on an evaluation criterion.

2.8 Boolean dynamic model

The Boolean Dynamic Model [30] posits a binary relationship between OTUs and uses k-means binarization [31] to find a threshold value for the abundance data. The method does not use any biological model like the other approaches but rather uses a "recapitulating approach" which updates and maintain binary links between the OTUs. The study is aimed at assessing *Clostridium difficile* infection and clindamycin antibiotic treatment through perturbation analysis, that is, the analysis through removal or addition of certain entities.

In order to tailor medicines to avoid microbiome disruption or prevent the cause of infection, it is essential to uncover what contributes to the stability of the microbial community. Traditional microbiology approaches do not discover the dynamics or properties of the large community. Even though earlier attempts were able to describe the basic dynamics of the gut, it uses many parametrisations which leads to overfitting. Boolean dynamic models solve this problem by reducing the parametrizing features.

The results are aligned to show how the difference in the dynamic network analysis can impact the healthy state of an organism. Through the metabolic reconstructions of the taxa it is possible to understand the mechanistic basis of the interactions.

This research is revolutionary in terms of being the first ever Boolean dynamic model from metagenomic sequence information and the very first integration to microbial analysis. It also sheds light onto how the microbial communities can contribute to the host metabolism and regarding what are its functional capabilities. The study is aimed at assessing *Clostridium difficile* infection and clindamycin antibiotic treatment through perturbation analysis. *B. intestinihomins* is proved to decreases the growth of *C. difficile* according to the invitro analysis.

2.9 Summarized review of the approaches:

As evident from the literatures above, the microbial interaction network inference methods have evolved from time to time. A major difference in the methodology is the adaptation of models for inference. Previous methods like Pearson Correlation method, SparCC [25] and CCREPE [26] have used pure statistical approaches as core methodology. In spite of the fact that SparCC has shown a considerable improvement from Pearson Correlation method in attaining a root mean squared error(RMSE) as low as 0.02 and a real life application to the Human Microbiome Project, similar to other correlation based methods, it faces criticisms like false associations in the inferred network and failure in accounting the compositionality or relative abundance structure of the community [32] and inference of false positives. Other potential disadvantages are the limitation of attaining only linear relationships and its difficulty in handling time series data. An evolution in this direction is the SPIEC-EASI [24] method which uses a correlation based approach with a graphical approach and deals with overfitting caused by spurious correlations by the identification of important associations in the network. However, the method has disadvantages like the requirement of parameter tuning which affects the network structure, non-directionality of the links similar to other correlation based methods and sparsity of the networks by the penalization of small associations.

To account for the intrusiveness of the microbial population, there has been an switch of purely statistical models like Pearson Correlation coefficient to model based approaches such as IMPARO [20], MetaMIS [22], RMN [23], LIMITS [27], Gao et al. [28] and Boolean dynamic model [30]. Approaches such as these delve into the biological aspect and capture the behavioural dynamics of the community. It is clear that the Lotka Volterra model has served as the primary model in the majority of these model-based techniques. There several reasons for it outlined as follows. Lotka-Volterra model

has (i)explored the fluctuations of the microbial populations over time within diverse ecological settings, and (ii)simulated microbial interactions in several environments such as on the gut microbial systems of humans and mice [33][34], aquatic ecosystems [35] and the microbial ecosystem developed on the onset of cheese ripening [36]. Simulating the relationships of microbes in an environment can help foresee the overall stability of a microbial setup.

Due to their ability to reflect the dynamic nature of microbial communities, Boolean models have been effectively used to infer microbial interaction networks like the Boolean dynamic model [30] already discussed above. Boolean models are able to capture the non-linearity of microbial interactions which the correlation based and other model based including the Lotka-Volterra models fail to capture [37]. It is also notable that Boolean models are more computationally efficient than the Lotka-Volterra based models such as IMPARO [20], MetaMIS [22], etc. due to its non-requirement of estimating parameters. Hence, this paves the way for more Boolean models for the capturing of microbial interaction patterns.

2.10 Importance of ESABO approach

ESABO, a Boolean model-centric approach pioneered by Claussen et al. [16], distinguishes itself through its exhaustive utilization of the statistical potential within the binary state space. This unique framework is used to analyse data on microbial abundance and reveal co-abundance patterns in the frequently bypassed low-abundance region of the human gut microbiome [16]. The discovery of highly significant interactions among low-abundant microorganisms, which charts a distinct direction in scientific investigation, is what distinguishes ESABO. This specific focus represents a paradigm shift in microbial research as it differs from past studies that mostly investigated interactions among high-abundance species.

ESABO's applications extend beyond microbiology into the area of gene regulatory networks. The precise distinctions between positive and negative interactions among genes are provided by ESABO scores, which are produced by the rigorous evaluation of gene states (active and inactive) [39]. This application enhances our knowledge of gene regulation and demonstrates ESABO's adaptability across a range of scientific fields.

The study also brings a new direction to the scientific domain of microbial studies by incorporating microbial interactions to shifts in the entropy. This approach is further enhanced by Shannon's information entropy theory, which adds rigour to the research by offering a quantitative measurement of uncertainty and information content. As a result, ESABO provides a multidimensional viewpoint that goes beyond conventional data analysis methods, making it a crucial tool for exploring complex systems and comprehending the nuanced interactions between data points. By harnessing the principles of Boolean Algebra and logical operators, ESABO empowers researches to navigate the complexity of datasets with greater precision and clarity. The method will be discussed in detail in the **Section(3.2).**

In summary, by enabling academics with a potent set of tools and approaches to unearth the hidden insights buried within the data, ESABO constitutes a cutting-edge addition to computer science. Its novel methodology has the potential to fundamentally alter how we analyse and interpret datasets, illuminating unexplored areas and opening the door to ground-breaking findings across a range of study fields.

Chapter 3: Background

Introduction

This section outlines the different approaches that have been adopted for implementation in our project. The main sections include 3.2 and 3.3 where the original ESABO and modified ESABO have been explained in detail. Section 3.4 will detail on the different network analysis methods used for analysis of complex networks in our project.

3.1 Information entropy

The concept of information entropy enables us to quantify the uncertainty or randomness present in a given set of data. Several types of entropy functions like Tsallis and Renyi entropy are popular. Nevertheless, when it comes to information theory, Shannon entropy [38] is extremely important. Frequently referred to as Shannon Information Entropy, it is extensively used for anticipating the value of information inside a discrete distribution.

The project revolves around the concept of information entropy and how it accelerates our approach of finding the relationship among microbiomes. Information entropy was a concept developed by Shannon et al. [38] to describe the measure of uncertainty or randomness of a random variable or a data source. In his paper, Shannon quantifies the thermodynamical aspect of "entropy" to the realm of communication and information.

The information entropy of a discrete random variable X with n possible outcomes $\{x_1, x_2, ..., x_n\}$ and corresponding probabilities $\{P(x_1), P(x_2), ..., P(x_n)\}$ is given by the formula:

$$H(X) = -P(x_i) * log_2(P(x_i)) \quad (3.1)$$

from i =1 to n

where:
- $H(X)$ represents the information entropy of the random variable X.
- $P(x_n)$ is the probability of the random variable X taking on the value x_i.

The higher the entropy, the more unpredictable the message and the more information it contains. The concept of information entropy is valuable in our study as it provides a quantitative measure of the uncertainty of interactions, i.e., when a species has very predictable and consistent interactions within an ecosystem, it signifies that it fills a specific niche or has a defined role.

3.2 ESABO(Entropy Shifts of Abundance Vectors under Boolean Operations)

ESABO method was originally introduced by Claussen et al. [16] for the inference of associations among microbial communities. It inputs binary abundance data for the study of abundance patterns and extraction of useful correlations in the data. Due to this interesting feature, it has proven to be beneficial for deriving the interactions among the low abundance segment of the human gut microbiome [16].

The method has original used a threshold of 1 for the creation of binary data from abundance information. Hence, it only focuses on the presence(1) or absence(0) of a microbial data. The ESABO

method works in two steps: (i) it uses the binary information from the microbial abundance data to find the entropy of two vectors and, (ii) then creates a null model of entropy shifts, where the second vector is randomly permuted.

The method is described more detailed as below:

Input:

The ESABO takes binary data $A \in B^{N_A \times N}$, where $B = \{0,1\}$which is a set of 0s and 1s. Here,

Row vectors: Samples of the data
Column vectors: Abundances of species in each specific sample

Steps of the approach:

For each species i and j, below are the steps for the calculation of ESABO score or z-score.

I. **Application of logical AND operation to vectors $\vec{b}_i$ and $\vec{b}_j$**:

Logical AND operation is applied to the abundance vectors of $\vec{b}_i$ and $\vec{b}_j$ of species *i* and *j* respectively. The element-wise application of logical AND operation yields the vector $\left(\vec{x}_{ij}^{AND}\right)$.

$$\left(\vec{x}_{ij}^{AND}\right)_k = \left(\vec{b}_i\right)_k \ \ AND \ \left(\vec{b}_j\right)_k \tag{3.2}$$

II. **Calculation of entropy of initial vectors $\vec{b}_i$ and $\vec{b}_j$:**

$$H\left(\vec{x}_{ij}^{AND}\right) = \ - \sum_{l \in \{0,1\}} \ \ p_l\left(\vec{x}_{ij}^{AND}\right) \log\left(p_l\left(\vec{x}_{ij}^{AND}\right)\right) \tag{3.3}$$

Here, $p_l\left(\vec{x}_{ij}^{AND}\right)$ is the probability of l in the vector $\vec{x}_{ij}^{AND}$, where $l = \{0,1\}$.

III. **Creation of null model using $\vec{b}_i$ and $\vec{b}_j^*$:**

Now, we take random number of shuffling of the second vector $\vec{b}_j$ and create a set of vectors $\vec{b}_j^*$, where step I and step II are repeated for each of these vectors $\vec{b}_j^*$ with the first vector $\vec{b}_i$(equations 3.4 and 3.5)

$$\left(\vec{x}_{ij^*}^{AND}\right)_k = \left(\vec{b}_i\right)_k \ \ AND \ \left(\vec{b}_j^*\right)_k \tag{3.4}$$

$$H\left(\vec{x}_{ij^*}^{AND}\right) = \ - \sum_{l \in \{0,1\}} \ \ p_l\left(\vec{x}_{ij^*}^{AND}\right) \log\left(p_l\left(\vec{x}_{ij^*}^{AND}\right)\right) \tag{3.5}$$

Hence, a null model of the entropy values (3.5) is created.

IV. **Calculation of z-score:**

Now, in-order to calculate the shift in the entropy or the ESABO score between species i and j, we calculate the z-score using the following equation:

$$Z_{ij} = \ \frac{H\left(\vec{x}_{ij}^{AND}\right) - \mu}{\sigma} \tag{3.6}$$

where, μ – mean of the null model of entropy distributions, σ – standard deviation of the null model of entropy distributions.

Interpretation of z-score values

According to [16], the positive z-scores between two species would imply that there is a positive or synergistic link between the two and a negative z-score is symbolic of a negative or competitive interaction among the two species. However, it is worth noting that only the z-scores above a threshold value of 1 is marked as a significant interaction.

3.3 Analytical formula for the calculation of mean and standard deviation

In order to account for the randomness of the ESABO method and for more computational efficiency, Mender et al. [17] introduced a new approach for the creation of null model entropy distribution. Here, the step III of the ESABO method is replaced by an analytical formula which will speed up the computation of the permutations of the second vector $\vec{b}_j$ and thereby a faster and efficient calculation of z-score.

The steps for the calculation of the mean and standard deviation are as follows:

I. Calculation of n and m:

$$n = p_1\left(\vec{b}_i\right) \cdot N_A \tag{3.7}$$

$$m = p_1\left(\vec{b}_j\right) \cdot N_A \tag{3.8}$$

Here, n is the number of ones in abundance vector $\vec{b}_i$ and m is the number of ones in abundance vector $\vec{b}_j$ and N_A is the number of samples in the dataset.

II. Calculation of z:

Let P be the set of all feasible permutations of the vector $\vec{b}_j$ and $\pi(j)$ be one such permutation from the set *P.* Now, z is the number of ones in the vector $\vec{x}_{ij}^{AND}$. Hence for $j = \pi(j)$, $z\left(\pi(j)\right)$ is the number of ones in $\vec{x}_{i\pi(j)}^{AND}$.

III. Calculation of $w(z)$:

Now we iterate over each of the values of z the interval $z \in [\mathrm{Max}(0,\ n + m - N_A),\ \mathrm{Min}(n,\ m)]$ and calculate a list of z values. According to each z values in the interval, we calculate $w(z)$ using the following formula:

$$w(z) = \frac{n!}{(n-z)!}\,\frac{(N_A - n)!}{(N_A + z - n - m)!}\binom{m}{z}(N_A - n)! \tag{3.9}$$

IV. Calculation of mean μ and standard deviation σ:

a) Calculation of mean μ:

$$\mu = \frac{1}{N_A!} \sum_Z H_z w(z) \tag{3.10}$$

where, $H\left(\vec{x}_{ij}^{AND}\right)$ is calculated using equation (3.3).

b) Calculation of standard deviation σ:

$$\langle H^2 \rangle = \frac{1}{N_A!} \sum_Z H_z^2 w(z) \tag{3.11}$$

$$\sigma = \sqrt{\langle H^2 \rangle - \mu^2} \tag{3.12}$$

where, $\langle H^2 \rangle$ is the squared entropy and the mean is subtracted from this in order to calculate the standard deviation.

3.4 Network science

3.4.1 Edge density

Edge density [42] determines the interlinkage or density of an undirected weighted graph(as in our project). It is derived by dividing the graph's actual edge count by the greatest number of edges that might be present in a graph with an identical number of vertices. The edge density (D) formula is as follows:

$$D = \frac{2E}{(V(V-1))} \tag{3.13}$$

where: D represents edge density, E is the number of edges in the graph and V is the number of vertices in the graph.

A node's connections offer important information about its function inside the network. This goes beyond only taking into account a node's degree or the degrees of every node in the network. Examining edge density is another method for learning more about both individual nodes and the entire network.

A network's density can be used to gauge how linked it is. It is a statistical measurement that, to be more specific, contrasts the actual number of edges existing in a network with the maximum number of edges that might possibly exist. Take two networks as an illustration, each with the same number of nodes. While there are roughly the same number of nodes in both networks (a and b), network (b) has a significant number of edges linking those nodes. Network (b) thus displays a higher level of density [42].

3.4.2 Local clustering coefficient

Local clustering coefficient [44] is a network metric which captures the extent to which the nodes in a network tend to link to each other or form triangles with other nodes. It is a measurement which details on a network's local structure and measure's its local link density at each node. For a node i with degree k_i, the local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{k_i(k_i-1)} \tag{3.14}$$

where L_i represents the number of links between the k_i neighbors of node i. The local clustering coefficient C_i is always a value between 0 and 1, with 0 indicating no linkage between the nodes, 0.5 indicating 50% probability of linkage of two nodes and 1 symbolising a complete network. The local clustering coefficient of node i is higher the more densely connected the area around it is.

3.4.3 Average clustering coefficient

Average clustering coefficient[44] offers a general assessment of the network's likelihood to organise into clusters or strong subgroups. It is a certainty that two randomly chosen nodes in a network will link to one another. The average clustering coefficient of a whole network is given by,

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{3.15}$$

where, C_i is the local clustering coefficients of the nodes $i = 1, \dots, N$.

3.4.4 Degree

For a node in a network, the degree of a node is represented as the number of connections that particular node has to other nodes in the network. [44]

3.4.5 Average Degree

The average degree [44] is a network metric which quantifies the average number of links that a node has in a network. For an undirected network, the average degree of the network is given by,

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{3.15}$$

where, k_i is the degree of the nodes $i = 1, \dots, N$.

3.4.6 Degree distribution

Individual nodes are described by their degree[44]. We may construct the degree distribution to gain a general understanding of the degree for each node in the network. This displays the number of nodes with each potential degree. Below are the steps for calculating the degree distribution:

- Calculate the degree for each network node to get a degree distribution.
- The next step is to count how many nodes have each degree.
- Plot the degree on the X-axis and number of nodes to the Y-axis.

3.4.7 Degree centrality

A node's degree, or the sheer number of edges it possesses, determines its degree centrality. The node is more centrally located the higher the degree. For calculating, the central nodes using degree centrality measure, we simply calculate the degree of each node and choose the node with the highest degree as the central node [42].

Chapter 4: Application to microbial abundance data

Introduction

The section 4.1 introduces the different methodologies that will shape the analysis of the microbial abundance data. This section addresses most of the research questions outlined in the Introduction and provides an organised approach of perceiving the problem. The section 4.2 will present the results of the analysis made in section 4.1 and provide evaluation of the results.

4.1 Methodology and Testing

4.1.1 Dataset Description

The microbiome dataset for our study is obtained from The Inflammatory Bowel Disease Multi'omics Database (IBDMDB)[4] website within the framework of the Integrative Human Microbiome Project. The IBDMDB encompasses different types of temporal data for analysing the gut microbial system in order to draw an investigation on the intricate connections between Inflammatory Bowel Disease(IBD) and disturbances in gut microbial ecosystems and diversity. Among this variety of longitudinal data, we are particularly interested in the taxonomic data, which is a dataset for the 16S ribosomal RNA gene that provides information on the taxonomic makeup.
The detailed procedure behind the taxonomic profiling of microbial communities in 178 biopsy samples using 16S rRNA gene sequencing is described in [40]. In short, the 178 biopsy samples were collected, DNAs were extracted from them, 16S rRNA gene was amplified using Polymerase Chain Reaction(PCR)[5] and the amplified DNA was sequenced and grouped into Operational Taxonomic Units(OTUs)[6], by mapping them to SILVA database [41]. We have selected one specific 16S rRNA dataset for our study from a collection of them in the website which contained these 178 samples and 982 species.

4.1.2 Data Pre-processing

The microbiome dataset underwent the following pre-processing steps to prepare it for analyses:

- The dataset was reindexed using OTU(Operational Taxonomic Unit) names replacing the conventional number indices for a better understanding of the microbial species involved in interactions. As a result, the dataset became more amenable to comprehensive analysis, with entries represented as numeric values for enhanced clarity.
- Another critical step was data cleaning. Any columns containing missing or "NaN" values were meticulously reviewed and removed from the dataset.
- The dataset was initially containing 982 species. But, we removed the 23 unwanted rows containing a 'zero' sum of abundances. This transformed the dataset to contain 959 species and 178 samples.
- For a more streamlined dataset, certain columns that were unnecessary like the "taxonomy" column were removed.

[4] https://ibdmdb.org
[5] Polymerase Chain Reaction(PCR) - Polymerase Chain Reaction (PCR) is a laboratory technique that makes copies of a specific segment of DNA, allowing scientists to study and analyze DNA more easily.
[6] Operational Taxonomic Unit(OTU)- Operational Taxonomic Units (OTUs) are a way to group and classify similar microorganisms, like bacteria or fungi, based on their genetic sequences, helping scientists identify and study different species in microbial communities.

4.1.3 Analysis and specification

In this section, we will explore the data in detail and discuss more on how the problem was analysed and what steps were taken to address the issue. It will include the steps taken to preliminary analysis of the data and forming of the research questions along with it.

4.1.3.1 Exploratory data analysis

One of the primary investigations of the data begins with exploring the underlying aspects of the data. Hence, we resort to analyse the data in the following manner.

Data structure

The rows of the dataset will represent the species and the columns will represent the different samples in the data. There are a total of 959 species and 178 samples in the current dataset.

i. Bar Plot: Exploration of the species and relative abundances

To assess the samples that had the most species in comparison to the other samples, we had generated a stacked bar chart. It is evident that samples 18 and 167 have the highest species abundance in the dataset (see Appendix B). This bar graph provides a thorough overview of species that may have high abundance in the dataset by additionally displaying the relative abundance of each species within each sample.

ii. Relative abundance calculation

Calculations of relative abundance were used to assess the species diversity in each sample. With the use of this analytical method, we were able to identify the range of species, from those with great prevalence to those with low representation, giving us a more nuanced view of the ecological dynamics of the dataset. These are the following steps taken for the calculation:

- **Initial Data Frame**: Beginning with the initial data frame containing species abundance data where each column represents the samples, and the values represent the abundance of different species across different samples.
- **Transformation**: We transform the data frame for the calculation of relative abundance. The aim is to create a new data frame where each entry indicates the relative abundance of different species across different samples. Thus, each individual columns(samples) will sum up to a total of 1, and will symbolise the proportion of each species within that sample.

iii. High and low abundance species

After the calculation of relative abundance data frame, it is made suitable for sorting the species into high and low abundance species. Following are the steps for the calculation:

- **Average relative abundance column:** Within the relative abundance data frame, a column for average relative abundance was created. This column showcases the mean of relative abundance of each species across different samples.
- **Sorting:** After the average relative abundance column was created, the data frame was sorted to give the 20 most highly abundant and 20 most lowly abundant species. Table 1 indicates the top five highly abundant species and Table 2 indicates the top five lowly abundant species.

Table 1: Top five highly abundant species

OTU ID	Average relative abundance
Unc05bd1	0.147306
Unc64172	0.128497
Unc054vi	0.095067
UncG3786	0.061076
Unc91005	0.042241

Table 2: Top five low abundant species

OTU ID	Average relative abundance
GVMFal10	2.060207e-07
Unc00psp	2.186920e-07
SxiAlask	2.679310e-07
Unc00lld	2.849885e-07
Unc014fz	3.921526e-07

The reason for choosing average relative abundance for sorting instead of sum of abundances is because it gives a normalised result. In this way, not only the total abundance of species is taken into account but also its population distribution across all samples.

4.1.4 Defining of threshold function

The ESABO method only inputs binary abundance vectors in its algorithm. Hence, it is important to binarize the dataset with a threshold value, As a preliminary step for thresholding of the data frame, we define a threshold function which inputs the data frame and the threshold value that thresholds the data. Any entry on the data frame equal to or above this value is indicated as 1 and below is indicated as 0. We will discuss each threshold value used for each step and also measure the optimal threshold value for different case scenarios in detail in the sections below.

4.1.5 Comparison between ESABO and MODIFIED ESABO

In this section we specify the approaches that we take to compare between the ESABO [16] and modified ESABO [17]. Below are the steps that we take in this direction:

Parameters, Input and Cases: We have assumed a threshold value of 1 for applying the original and modified version of the ESABO approach. Firstly, we take a small subset of 25 species(rows) and 178 samples(columns) from the dataset. This subset will serve as the test subset for comparison. For the effective comparison of original and modified ESABO methods, we have taken three case scenarios:

- **Case 1**: First, using a data frame of z-scores of links obtained from original ESABO using a randomisation factor of 50 (i.e., 50 random permutation calculations for the creation of null model, refer section 3.3 step III) appended by a column of z-scores obtained from modified ESABO.
- **Case 2:** Second, repeating the same scenario but, due to the randomness of permutations there will be a slight changes in the original ESABO z-score results. As a result, the second data frame will have slightly different but almost identical z-scores and a similar data frame as that of Case I is created by appending with a column of z-scores obtained from modified ESABO which will remain the same for any iterations due to a constant value.

- **Case 3:** Now, we create a data frame same as in Case 1 except now replacing with a randomisation factor of 1000 permutations (i.e., 1000 random permutation calculations for the creation of null model, refer section 3.2 step III) as mentioned in the original ESABO method.

Procedure: For each of the cases, we two columns within the data frame which will encompass the absolutive change and relative percentage change . Now, these results are compared for testing the similarity.

Implementation and Testing: For a comprehensive evaluation of the similarity and differences in both methods, we draw conclusions from the column of relative percentage change in z-scores. From this, we examine the number of rows(species) that have a relative percentage change exceeding 50%. This helps us to critically analyse the proportion of links that has a massive difference in values when produced by both methods. Calculating the ratio of these links to the total number of links can help us evaluate the approximate proportion of results which would differ when using both methods for analysing. Also, to check whether there is much difference between the significant links (i.e., links with an absolute z-score greater than or equal to 1) we check for the number of significant links and for the indices with z-scores relative percentage change of less than 30%. Finally, the networks are generated using Case 3 and are analysed using network complexity measures such as edge density, average degree, average clustering coefficient and degree distribution for similarity testing.

4.1.6 Analysing the impact of sample size on z-scores

Our next step of analysis is concentrated on finding whether there is a boost in the z-scores when different sample sizes are taken into consideration. We are interested in this investigation as this will aid us in getting the right sample size that can maximise the number of significant links by increasing the number of absolute z-scores above a threshold of 1. We also analyse whether there is a general trend of increase or decrease of z-scores with varied sample sizes. These are the different sample sizes we deeply investigated by keeping the number of species intact.

- **Case I**: 20 Samples and 50 species
- **Case II**: 50 Samples and 50 species
- **Case III**: 100 Samples and 50 species
- **Case IV**: 178 Samples(all samples) and 50 species

The results of this investigation are outlined in Section 4.2.

4.1.7 Optimal thresholding of the data using MODIFIED ESABO approach

The primary investigation behind our project is to find an optimal threshold value which will binarize the dataset and lead to more information gain when ESABO approach is applied. In this problem, we are investigating which is the optimal threshold value that can lead to a boost in the z-scores and hence, an increased number of links. We will be finding the optimal thresholding for three specific cases:

- **Case I**: 20 most highly abundant species to check the optimal threshold in the high abundance scenario. Here, we will also be seeing as to how to test whether it is the optimal threshold using network analysis parameters such as edge density by comparing this with other threshold values. This result will serve as the key for opting the optimal threshold value.
- **Case II:** Two cases of 200 and 300 most low abundant species are considered for optimal thresholding.
- **Case III:** A mixture of high and low abundance species**.** This is for analysing the optimal threshold value for different compositionality and size of the datasets.

Procedure for Case I and Case II: Threshold values are selected randomly from the linear space of values between the maximum and minimum values of the dataset. We find the best threshold for high abundance and low abundance microbiomes by finding the absolute sum of significant z-scores (greater than 1) for each of the threshold values taken. A line plot is made denoting the different threshold values and the different summations in z-scores obtained.

Testing for Case I and Case II: This parameter for estimating the best threshold is used as the main testing parameter for optimal thresholding. Case I is also analysed with an additional network analysis parameter of edge density and average degree to check the performance of the threshold values and the optimal threshold for the creation of better networks.

Procedure for Case III: In Case III, we analyse the optimal thresholding of the datasets using the summation in z-scores for each of the datasets. For Case III, we analyse three sets of data. First, using a mixture of 50 high and low abundant species. Second, for a mixture of 100 high and low abundant species. Thirdly, for a mixture of 150 high and low abundant species.

4.1.8 Network creation and analysis using the optimal threshold value

In this section, we explore the different networks formed using the optimal thresholding in case of 150 high and low abundance species. The networks are then analysed using all the network complexity measures as mentioned on Section 3.4, including edge density, average degree, clustering coefficient, degree distribution, and centrality measures for finding the central nodes in the network.

4.2 Results and evaluation

4.2.1 Comparison of ESABO and MODIFIED ESABO

Each of the different cases according to section 4.4.3 were examined.

- Case I resulted in 94 links which had a relative percentage change in z-scores from both methods to be greater than 50% out of 300 total links, reaching a ratio of approximately 3:10.
- Case II also witnessed a similarity to Case I wherein only 92 out of 300 links had a relative percentage change in z-scores exceeding 50%.
- Case III, on the other hand, made a count of 98 out of 300 links forming a proportion of $1/3^{rd}$ of the links to have a significant change in z-score values.

Further analysis of Case III data frame revealed that out of 50 significant links, 42 of them had a relative percentage change of less than 30% from both the methods.

Table 3 suggests that both original ESABO and MODIFIED ESABO has a very small difference in the network analysis parameter results and yields almost similar results. The degree distribution graphs(Figure 3) from both networks also portrays similarity in values and appearance. Although, it is worth noting that original ESABO led to a higher run time compared to modified ESABO method especially in Case III and modified ESABO according to [17] can give a faster computation of z-scores.

Thus, comparing the results we are able to arrive at the conclusion that both the methods yield similar results in terms of z-scores with only $3/10^{th}$ of links with a significant relative percentage change of exceeding in Cases I, II and $1/3^{rd}$ in case of Case III. This along with the results from network analyses tells us that both methods lead to similar results and that the analytical formula used in modified ESABO

approach gives a good approximation of permutations for creation of null model(section 3.4 step IV). Thus, this motivates us to use modified ESABO approach for the rest of our project as the primary method.

Table 3: Network analysis results for both methods

PARAMETER	ORIGINAL ESABO	MODIFIED ESABO
Edge density	0.08333	0.09
Average degree	4.0	4.30
Average clustering coefficient	0.25026	0.31292

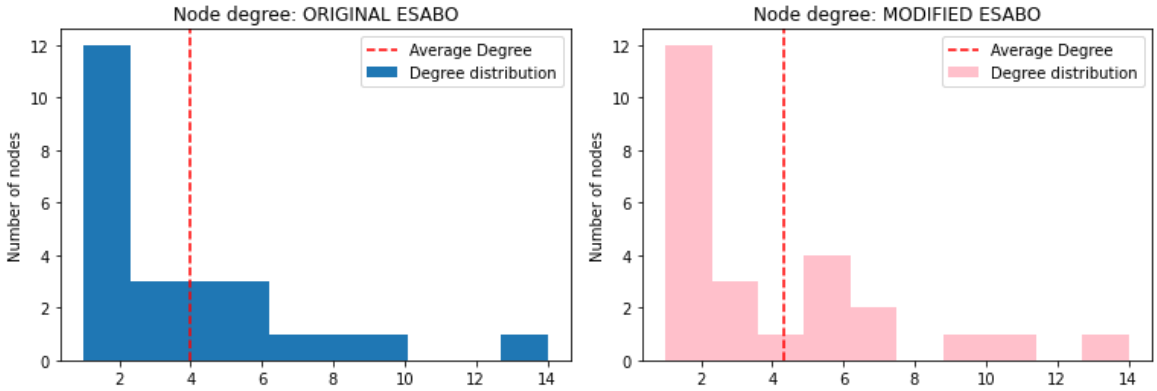

Figure 2: Degree distribution obtained from a network of 25 species using original ESABO method(left) and modified ESABO method(right)

4.2.2 Analyzing the impact of sample size on z-scores

The table below illustrates the varying number of samples with the same count of species. It depicts that the z-scores increase with an increase in the number of samples. Thus, there is a general trend of increase in the z-scores on average.

A possible explanation for this is the gain in more information when more number of samples are taken into account. An expansion of sample pool indicate that there might be more significant links due to an added granularity on the abundance of microbiomes in that sample. Consequently, our analysis benefits from a more comprehensive aspect potentially revealing more insights within the data. The exploration is vital in underscoring the validity of sample size in the interpretation of z-scores.

Table 4: Varying sample sizes and their information gain

Count of Species	Count of Samples	Summation of z-scores

50	20	192.015625
50	50	311.78125
50	100	521.765625
50	**178**	**660.625**

4.2.3 Optimal thresholding for high abundance dataset

As noticed, from the Table 5, the high abundance organisms containing 20 species have an optimal threshold value of 30. This threshold value corresponds to the highest information gain or absolute summation of significant z-scores equal to 158.75, i.e., the highest number of significant links in the dataset. The testing parameters used, including the edge density and average degree is also the highest for this optimal threshold valuing to up to 0.37894 and 14.4 respectively**.** Thus, it indicates that we are able to use information gain, i.e., the absolute sum of significant z-scores as testing parameter for optimal threshold value. This is clearly an indication that the threshold value is not necessarily 1 for all cases as stated in the original paper.

Table 5: Threshold values and corresponding information gain for high abundance microorganisms

Index	Threshold values	Information Gain	Edge Density	Average Degree
1	1	77.968750	0.2	7.6
2	4	121.437500	0.30526	11.6
3	20	151.437500	0.36842	14
4	25	155.375000	0.37368	14.2
5	27	155.750000	0.37368	14.2
6	**30**	**158.750000**	**0.37894**	**14.4**
7	2174.67	54.921875	0.153846	4
8	4349.34	27.203125	0.112	2
9	6524	3.000000	0.3334	1.3334

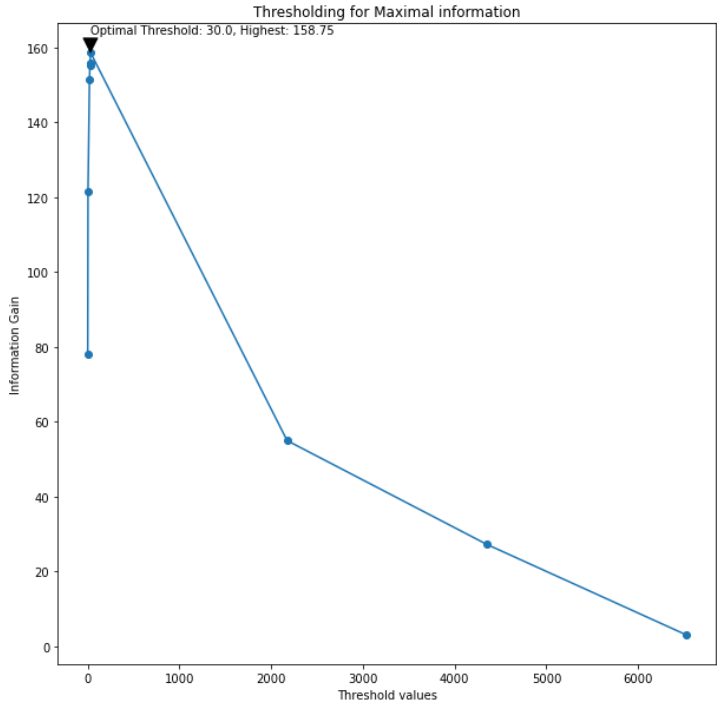

Figure 3: Line plot indicating the optimal threshold value for high abundance microorganisms.

4.2.4 Optimal thresholding for low abundance dataset

Table 6 clearly shows the difference in values of information gain when the low abundance dataset is posed with different threshold values. It is noticeable that for a number of 200 and 300 low abundance species, the information gain is the maximum in the case of threshold value of 1 which is 9336.2187 and 19438.62500 respectively. Also, a look into the edge density and average degree of networks formed from 200 and 300 low abundances species is highest in the case of threshold value of 1. This is also demonstrated by the line graphs from Figure 4**.** Even though the values range between 0-7 in the case of 200 species and between 0-11 in the case of 300 species, the optimal threshold value is still 1.

Table 6: Threshold values and corresponding information gain for low abundance microorganisms under different species sizes

Index	Number of species	Threshold values	Information gain	Edge Density	Average Degree
1	**200**	**1**	**9336.2187**	**0.02555**	**8.943**
2	200	2	8612.31250	0.02803	8.5228
3	200	3	2744.62500	0.04825	6.2727
4	**300**	**1**	**19438.62500**	**0.02339**	**13.1943**
5	300	2	17723.59375	0.02283	11.5098

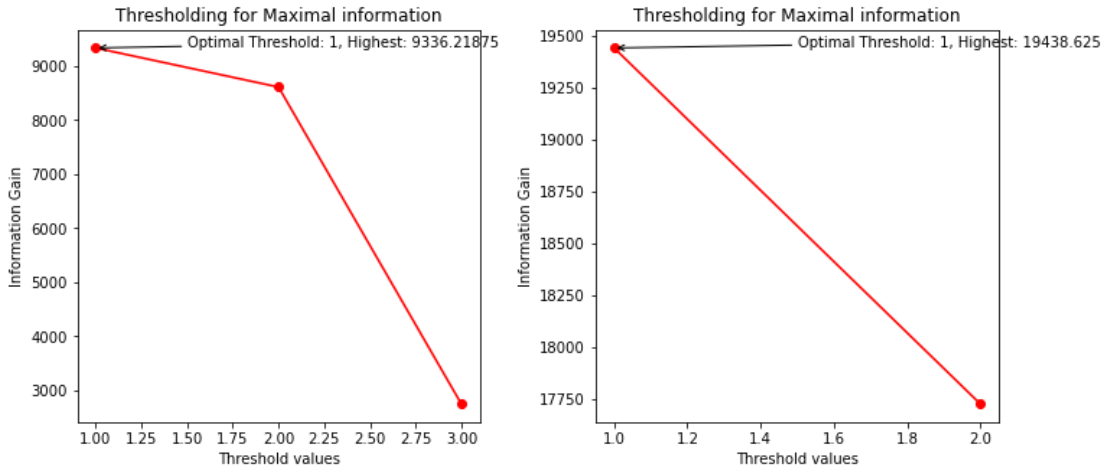

Figure 4: Line Plot indicating the optimal threshold value for low abundant microorganisms with 200 low abundant species(left) and 300 low abundant species(right)

4.2.5 Optimal thresholding for a mixture of high and low abundance dataset

[1] Results for 50 high and low abundance species:

From Table 7, we are able to figure out that for a composition of 50 high and low abundant species, the optimal threshold value is 50 with a maximal information gain of 1083.484375**.** This observation is also evident from line graph from Figure 5(left). However, one particular remark from the graph is that, for the threshold of 70 and 100, the surge in the information gain is not sudden even with a gap of 30 points between the threshold values.

[2] Results for 100 high and low abundance species:

Table 8 indicates that for a mixture of 100 high and low abundant species(with an increase in the low abundant species), the optimal threshold value has now become 40 with an information gain of 4424.968750, which is a bit lower compared to the first case. The line graph in Figure 5(right) also denotes that, even though the threshold has become 40, there is a possibility of variable threshold. This is showcased with a rise in the information gain to 4399.140625 at the threshold of 15, which then considerably reduces at the threshold of 30 and again rises to give an optimal threshold value at 40.

Table 7: Threshold values and corresponding information gain for a mixture of 50 high and low abundance microorganisms.

Index	Number of species	Threshold values	Information gain
1	50	9	975.593750
2	50	15	1008.484375
3	**50**	**50**	**1083.484375**
4	50	70	1060.406250
5	50	100	1054.937500
6	50	1000	500.562500

Table 8: Threshold values and corresponding information gain for a mixture of 100 high and low abundance microorganisms.

Index	Number of species	Threshold values	Information gain
1	100	10	4318.890625
2	**100**	**15**	**4399.140625**
3	100	30	4373.968750
4	**100**	**40**	**4424.968750**
5	100	50	4422.468750
6	100	100	4138.546875

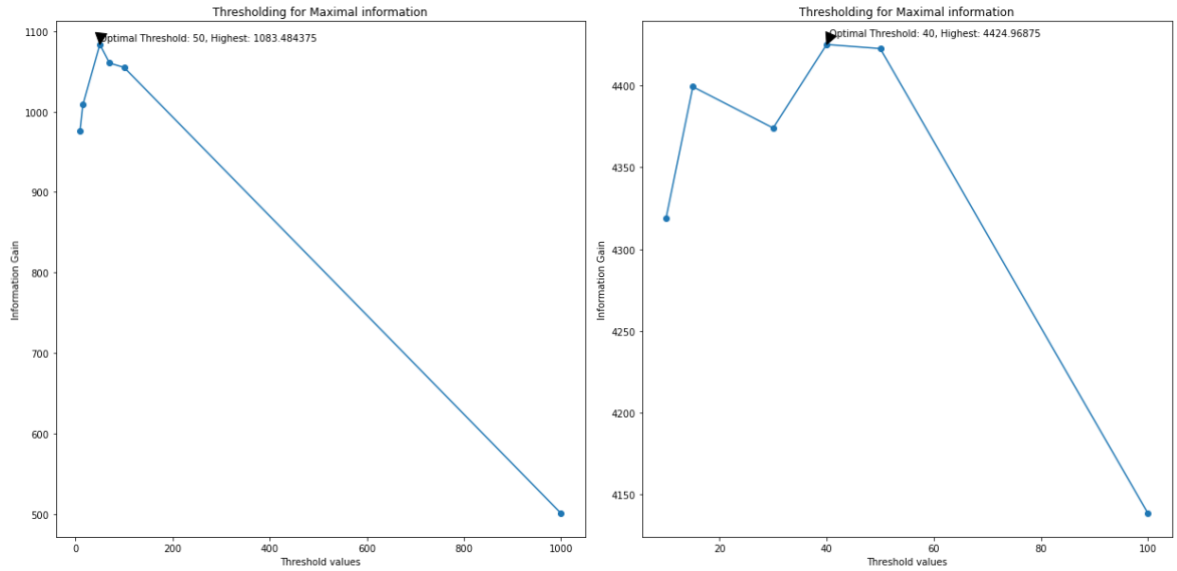

Figure 5: Line Plot indicating the optimal threshold value for a mixture of high and low abundant microorganisms with 50 species(left) and 100 species(right)

[3] Results for 150 high and low abundance species:

As evident from the Figure 6 and Table 9, 5 is the optimal threshold value for the data containing a mixture of 150 high and low abundant micro-organisms, with an information gain of 9539.531250. Hence, the introduction of microorganisms with lower abundances has led to a notable adjustment in the threshold value, reducing it to a lower threshold of 5.

Table 9: Threshold values and corresponding information gain for a mixture of 150 high and low abundance microorganisms.

Index	**Number of species**	**Threshold values**	**Information gain**
1	150	1	9512.468750
2	**150**	**5**	**9539.531250**
3	150	10	9158.062500
4	150	20	8982.781250
5	150	30	8817.500000
6	150	50	8775.500000
7	150	70	8088.953125

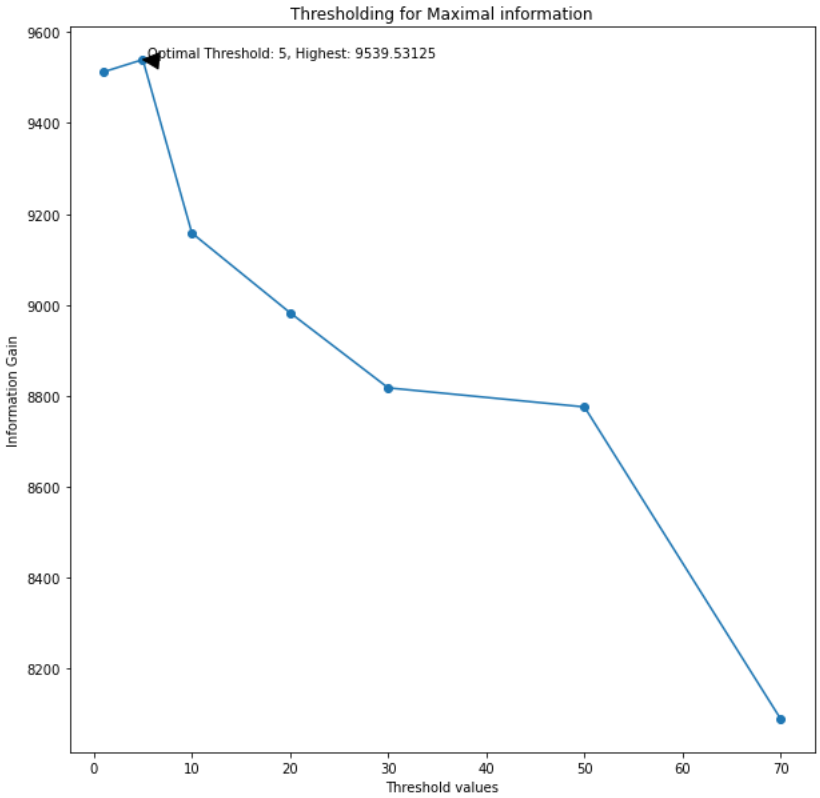

Figure 6 : Line Plot indicating the optimal threshold value for a mixture of 150 high and low abundant microorganisms

4.2.6 Network creation using the optimal threshold value for a mixture of 150 species

Figure 7 indicates the network formed by the mixture of 150 high and low abundant species. It is visible from the figure that there were more positive interactions(pink edges) compared to negative interactions(grey). Positive interactions predominate, which may indicate that species in this network prefer cooperative or mutually beneficial interactions over antagonistic or competitive ones. This is because not all the low abundant microbiomes were taken into consideration and only a small mixture of both compositionality were used for analysis.

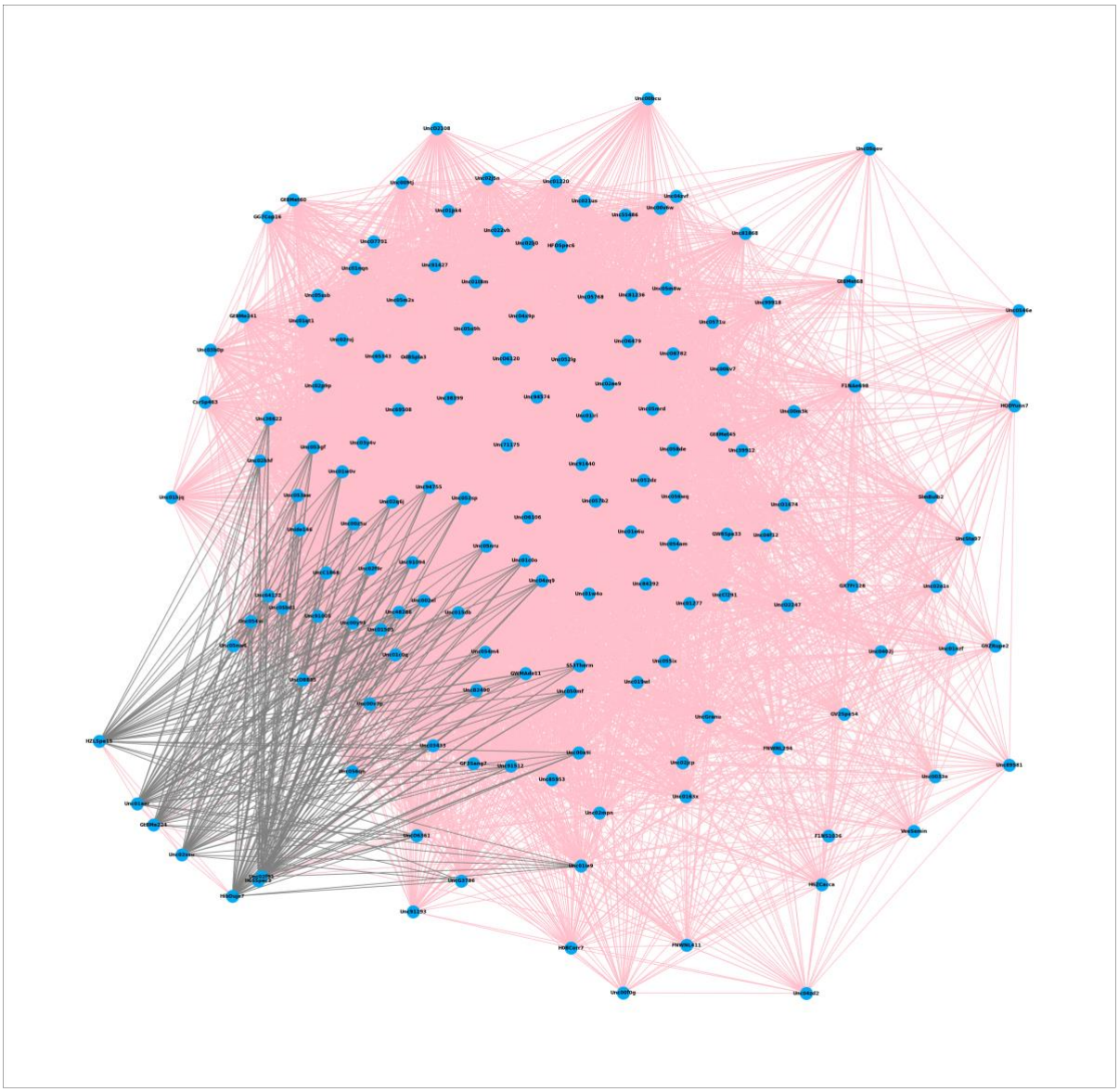

Figure 7: Network formed by the mixture of 150 species with threshold of 5- positive links(pink), negative links(grey)

Table 10: Species with most significant negative interactions from Figure 7.

Source	Target	Weights
UncO8895	HZLSpe15	-2.59375
Unc05bd1	HibDuja7	-2.59375
UncO8895	Unc02f91	-2.59375
Unc054vi	Unc02f91	-2.59375
UncO8895	H6SSpec3	-2.59375

Table 11: Species with most positive interactions from Figure 7.

Source	Target	Weights
HZLSpe15	Unc01aer	13.37500
HZLSpe15	Unc02xsu	13.37500
HZLSpe15	HibDuja7	13.37500
Unc01aer	Unc02xsu	13.37500
Unc01aer	HibDuja7	13.37500

4.2.7 Network analysis using the optimal threshold value for a mixture of 150 species(Table 12)

- The network indicates an edge density of 0.311409, which tells us that there is a possibility of 31.14% connections between the nodes in the network.
- An average degree of 92.8 denotes that a node in the network is typically connected to an average of 92.8 of other nodes.
- Average clustering coefficient of 0.376717 signifies that on an average each node tends to form triangles with their neighbors with a probability of 37.67%.
- The table also indicates that there are three communities within the network, indicating a group of three densely connected clusters in the graph.
- A central node of 'Unc01c0o' using degree centrality measure is representative of the fact that the microorganism 'Unc01c0o' is well connected to other nodes in the network.

Table 12: Network analysis results of a mixture of 150 species

Measure	Value
Edge Density	0.311409
Average Degree	92.8
Average Clustering coefficient	0.376717
Number of communities	3
Central Node(s)	Unc01c0o

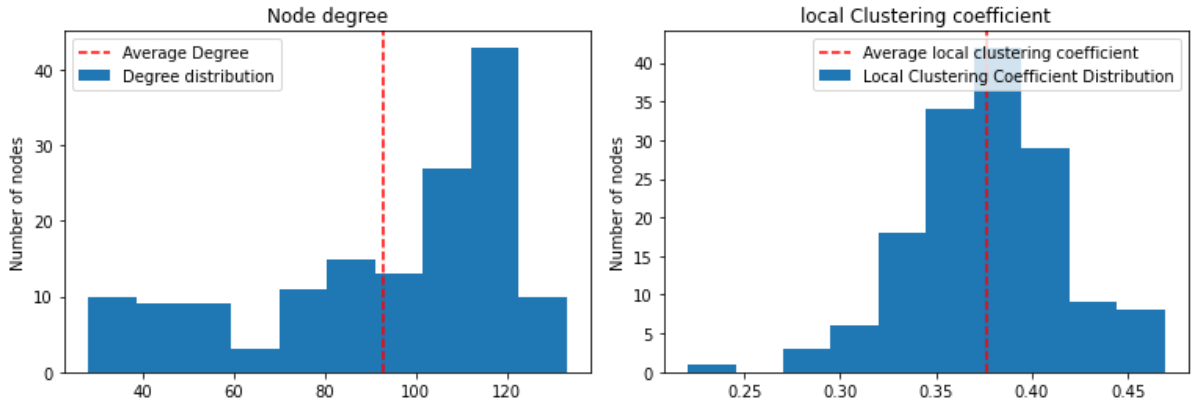

Figure 8: Degree Distribution(left) and Local Clustering Coefficient Distribution(right) of 150 species

- Degree Distribution - Figure 7 shows the positive and negative links in the network. Table 10 has specified the top 5 most significant negative and positive interactions in the network. Figure 8(left) denotes that almost 45 nodes have a high degree of around 110 which signifies that they are connected to 110 other nodes in the network.

- Local Clustering coefficient distribution - Figure 8(right) denotes that around 40 have a clustering coefficient of 0.37 and this indicates that these nodes are 37% likely to form tight knit clusters within the network. There is also a proportion of around 8 nodes which reaches a clustering coefficient of nearly 0.5. No nodes around the range of 0.23 to 0.27 and low clustering coefficient values below that range is indicative that there is a possible hole in the network because those nodes are not able to cluster or connect with the other nodes.

Chapter 5: Application to plant abundance dataset

Introduction

The section 5.1 introduces the different methodologies that will shape the analysis of the plant species abundance data. This section pertains to the application of ESABO approach to a plant species abundance data and thus will prove the viability of the method to ecological datasets. The section 5.2 will present the results of the analysis made in section 5.1 and provide evaluation of the results.

5.1 Methodology and Testing

5.1.1 Dataset Description

The plant abundance dataset was publicly open for use and obtained from the New South Wales(NSW) government website[7]. The dataset was collected for the purpose of establishing the effect of stock grazing on various sites. The study's objective was to provide information for future choices on Red gum and Cypress pine grazing permits and licences. In our project, we specifically look into the River Red Gum tree sites and the plant abundances in that region and derive the interactions of the plant species in that region, describing how they impact each other.

5.1.2 Data Pre-processing

Below are some of the pre-processing steps taken before the data was ready for analysis.

- **Data cleaning**: The dataset initially contained extraneous information like plant cover and area, which necessitated a thorough data cleaning process. Additionally, there were a lots of redundant information, with instances such as the site 'RRG_001'(First River Red Gum site) containing the same species 'Hordeum marinum' in repetition, which meant that the they had to be amalgamated together to give the sum of abundance of that species. Furthermore, we required a transformation of the data in a more structured manner pertaining to species and their abundance across different sites. Thus, a pivot table was created in excel which selected only the specific labels required and grouped the abundances for same species by summing them up.
- After the data was loaded, the data had to be cleaned in python. Several unwanted rows and columns were removed. The 'NaN' values were replaced with 0s.
- The rows of dataset as in the case of microbiome dataset was reindexed to the species names and the columns of the dataset was reindexed to the site names.

5.1.3 Exploratory data analysis

In order, to dive deeper into the intricacies of the dataset, we explored some key insights from the data. This included finding the relative abundances of the different plant species similar to section 4.1.3.1 ii.

[7] https://datasets.seed.nsw.gov.au/dataset/nsw-grazing-studyd114

Highly and low abundance species

After the calculation of relative abundance data frame similar to section 4.1.3.1 ii, it is made suitable for sorting the species into high and low abundance species. Following are the steps for the calculation:

- **Average relative abundance column:** Within the relative abundance data frame, a column for average relative abundance was created. This column showcases the mean of relative abundance of each species across different samples.
- **Sorting:** After the average relative abundance column was created, the data frame was sorted to give the 50 most highly abundant and 200 most lowly abundant species. Table 13 indicates the top five highly abundant species and Table 14 indicates the top five lowly abundant species.

Table 13: Top 5 Highly abundant plant species

Genus species	Average relative abundance
Lolium rigidum	0.14315
Bromus diandrus	0.074052
Eleocharis acuta	0.073124
Eleocharis pusilla	0.061192
Marsilea drummondii	0.060051

Table 14: Top 5 lowly abundant plant species

Genus species	Average relative abundance
Schoenoplectus validus	3.196982e-07
Arthropodium sp. 3 (aff. Strictum)	3.263335e-07
Rosa sp.	4.027467e-07
Melia azedarach	4.535765e-07
Dichelachne crinite	4.728803e-07

Data structure

The rows represent different plant species and the columns represent the different plant sites(‘RRG_001’, ‘RRG_002’, etc.). After the pre-processing steps discussed before, there are a total of 354 plant species and 150 sites in the current dataset.

5.1.4 Optimal thresholding using MODIFIED ESABO for highly and lowly abundant plant species

Our first stage of investigation on the plant abundance data lies in finding an optimal threshold value for a set of 50 highly abundant plant species and 200 low abundant species. In the case of the latter, our aim is to encompass a more diverse range of values beyond the recurring 0s and 1s, thus necessitating a larger dataset size. Within this new subset of low abundant species, the values span from 0 to 122, providing a more fertile ground for the analysis of an optimal threshold. We now employ the threshold function discussed in section 4.1.4 to determine the best threshold value for both the datasets.

Procedure: The primary stage behind finding the optimal threshold value for the dataset lies in opting the different threshold values for the data. Threshold values are selected randomly from the linear space

of values between the maximum and minimum values of the datasets. For each of the threshold value selected, we calculate the information gain or the summation of significant absolute z-scores using ESABO method. A line plot indicating the different threshold values and their corresponding information gain is also obtained.

Testing: We are inclined to use the sum of absolute significant z-scores (greater than 1) as a metric for assessing the optimality threshold values based on the findings in Chapter IV. Thus, a larger sum of absolute z-scores will indicate a higher number of links in the dataset with that particular threshold value. The network created with the ideal threshold value will become less sensitive to entropy loss as a result of the growth in the number of links, making it a better parameter for data splitting.

5.1.5 Network creation and analysis using the optimal threshold value

In this section, we explore the different networks formed in case of 50 high and 200 low abundance species from the section before using the optimal threshold value for both. The networks are then analysed using all the network complexity measures as mentioned on section 3.4, including edge density, average degree, clustering coefficient and degree distribution.

5.1.6 Network creation and analysis of the entire dataset using the threshold value of 1

The major focus of the plant abundance dataset was to successfully implement the ESABO method on the complete dataset containing 354 species and 150 sites. The ESABO method is applied on the dataset using a threshold value of 1, because as seen in the case of microbiome dataset in chapter IV, a single optimal threshold for the entire dataset is not possible due to differing compositionality in various data subsets.

After the application of ESABO method on the entire dataset, the large network is analysed using all the network analysis parameters mentioned on section 3.4, such as edge density, average degree, clustering coefficient, degree distribution, community detection and central nodes.

All the results obtained are detailed in the next section.

5.2 Results and evaluation

5.2.1 Optimal thresholding using MODIFIED ESABO for highly abundant plant species

It is observable from Table 15 that the optimal threshold value for the 50 high abundant plants is 10 and the maximal information gain obtained was 731.62500. This observation is reinstated by the line graph Figure 9 (left). Accordingly, when a threshold of 10 is applied to the information pertaining to these abundantly found plants, it produces the greatest decrease in uncertainty or increase in knowledge, as measured by the information gain.

5.2.2 Optimal thresholding using MODIFIED ESABO for lowly abundant plant species

It is observable from Table 16 that the optimal threshold value for the 200 low abundant plants is 1 and the maximal information gain obtained was 5073.453125. This observation is reinstated by the line graph Figure 9 (right). Thus, even though a set of 200 low abundant species was selected for optimal thresholding in the case of low abundant species, there is still a shift in the threshold value towards the value of 1. A possible reason might be prevalence of small values compared to the high ones in the dataset.

Table 15: Threshold values and information gain for 50 high abundant plants

Index	Threshold values	Information gain
1	5	704.375000
2	**10**	**731.625000**
3	15	711.546875
4	30	710.078125
5	40	692.781250

Table 16: Threshold values and information gain for 200 low abundant plants

Index	Threshold values	Information gain
1	**1**	**5073.453125**
2	2	2656.578125
3	4	1324.875000

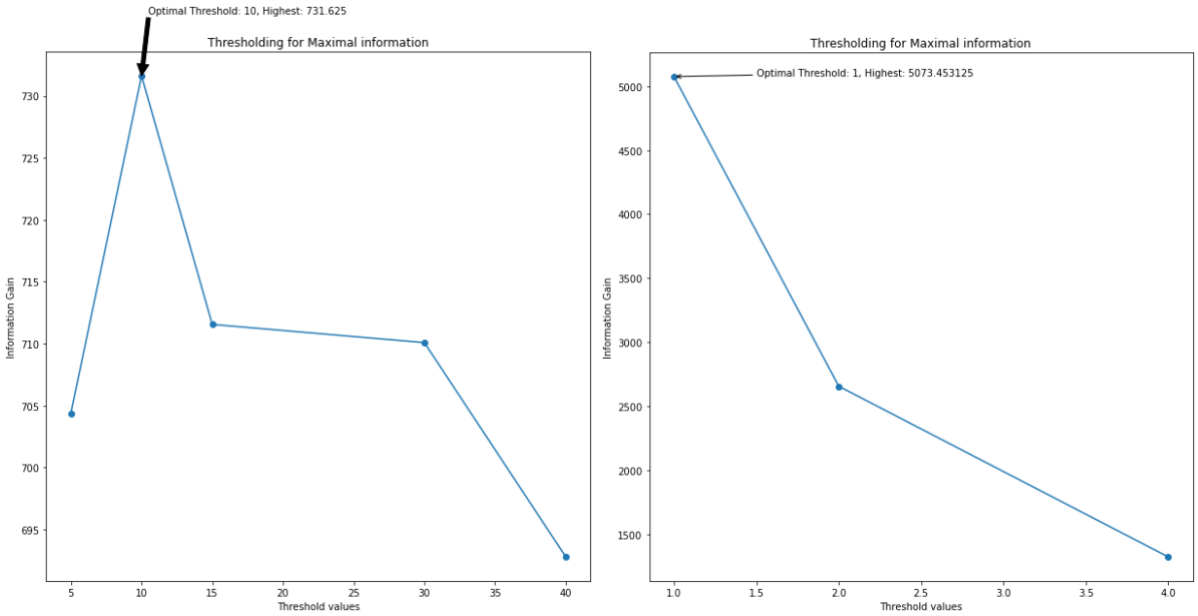

Figure 9: Line Plot indicating the optimal threshold value for 50 high abundant microorganisms (left) and 200 low abundant microorganisms(right)

5.2.3 Network creation using the optimal threshold value for 50 highly abundant plant species

Figure 10 demonstrates the entire network of highly abundant plant species. There are a lot of significant positive links within the network but only one significant negative link (in red) is observed in this dataset. This is because of the high abundance in the plants that most negative links are weak in nature. Since, the plants are highly abundant it indicates that they naturally are co-existing together and do not face any inhibitory or competitive interaction from the other species in the area.

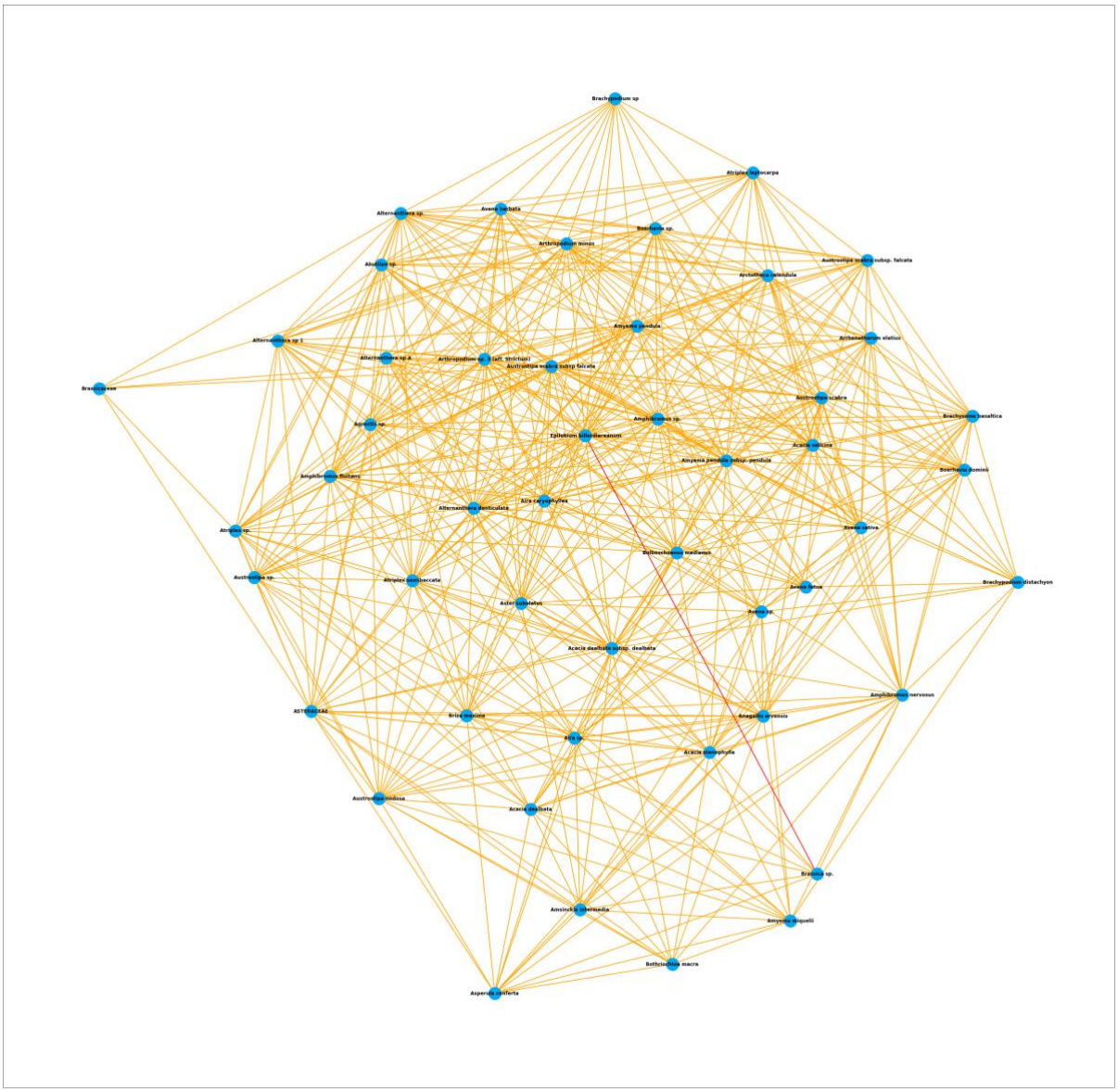

Figure 10: Network generated for 50 highly abundant plants with threshold of 10, positive links(orange), negative links(red)

5.2.4 Network analysis results for the network of 50 highly abundant plants(Table 17)

- The network indicates an edge density of 0.168571, which tells us that there is a possibility of 16.8 % connections between the nodes in the network.
- An average degree of 16.52 denotes that a node in the network is typically connected to an average of 16.52 of other nodes.
- Average clustering coefficient of 0.26996 signifies that on an average each node tends to form triangles with their neighbors with a probability of 26.996 %.
- The table also indicates that there are three communities within the network, indicating a group of three densely connected clusters in the graph.

Table 17: Network analysis results of 50 high abundant species

Measure	Value
Edge Density	0.168571
Average Degree	16.52
Average Clustering coefficient	0.26996
Number of communities	3

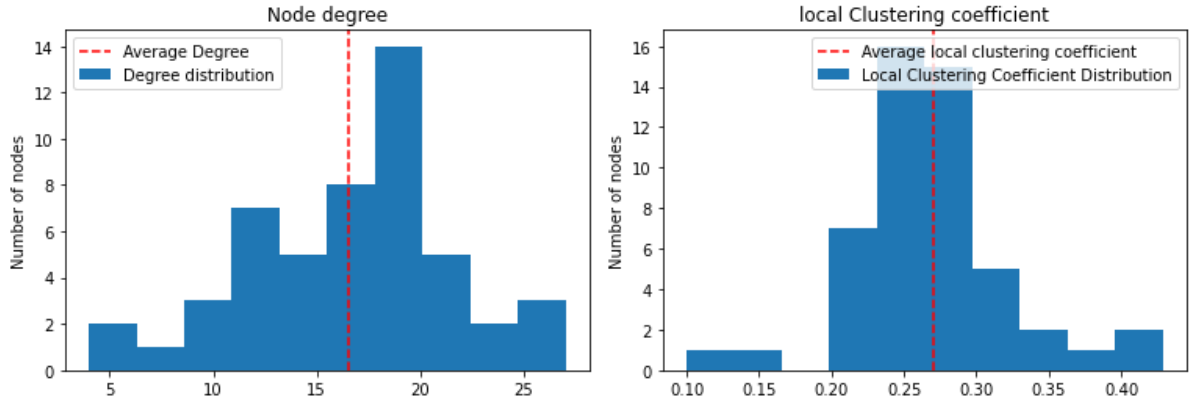

Figure 11: Degree Distribution(left) and Local Clustering Coefficient Distribution(right) of 50 highly abundant species network

- Degree Distribution- Figure 11(left) denotes that almost 14 nodes have a high degree of around 18 which signifies that they are connected to 18 other nodes in the network.

- Local Clustering coefficient distribution-Figure 11(right) denotes that around 16 have a clustering coefficient of 0.27 and this indicates that these nodes are 27% likely to form tight knit clusters within the network. There is also a proportion of around 2 nodes which reaches a clustering coefficient of nearly 0.45. No nodes in the range of 0.17 to 0.20 possibly denotes that there is a hole in the network and low clustering coefficient values below that range is indicative that there is a possible hole in the network because those nodes are not able to cluster or connect with the other nodes.

5.2.5 Network creation using the optimal threshold value for 200 lowly abundant plant species

Figure 12 demonstrates the entire network of lowly abundant plant species. There are a lot of significant positive links within the network and no single negative links. This is because all the plants considered are low abundant that most negative links are weak in nature. Due to the absence of any high abundant plant species, the low abundant species can co-exist mutually without any inhibition.

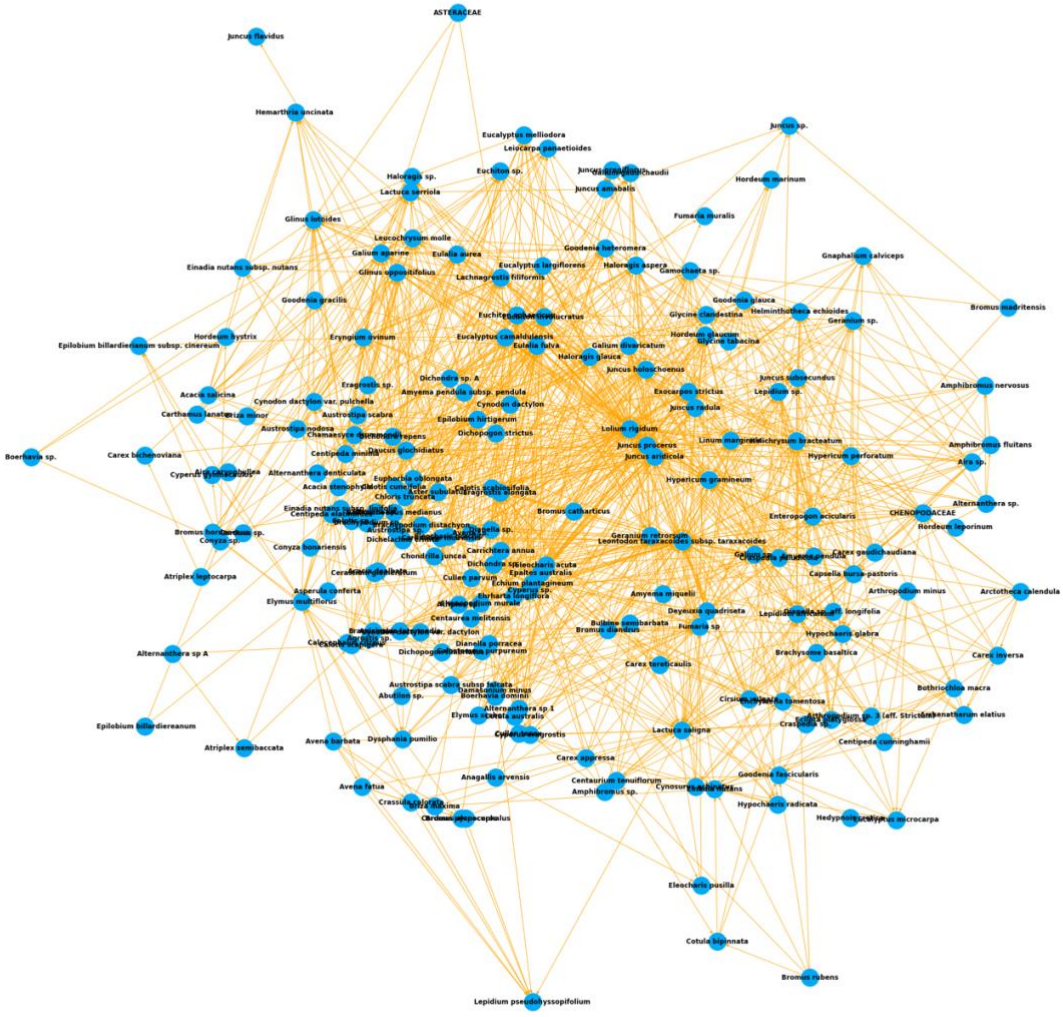

Figure 12: Network generated for 200 lowly abundant plants with threshold of 1, positive links (orange)

5.2.6 Network analysis results for the network of 200 lowly abundant plants(Table 18)

- The network indicates an edge density of 0.032783, which tells us that there is a possibility of 3.27 % connections between the nodes in the network.
- An average degree of 12.78571 denotes that a node in the network is typically connected to an average of 12.8 of other nodes.
- Average clustering coefficient of 0.33254 signifies that on an average each node tends to form triangles with their neighbors with a probability of 33.254 %.
- The table also indicates that there are six communities within the network, indicating a group of six densely connected clusters in the graph.

Table 18:Network analysis results of 200 low abundant species

Measure	Value
Edge Density	0.032783
Average Degree	12.78571
Average Clustering coefficient	0.33254
Number of communities	6

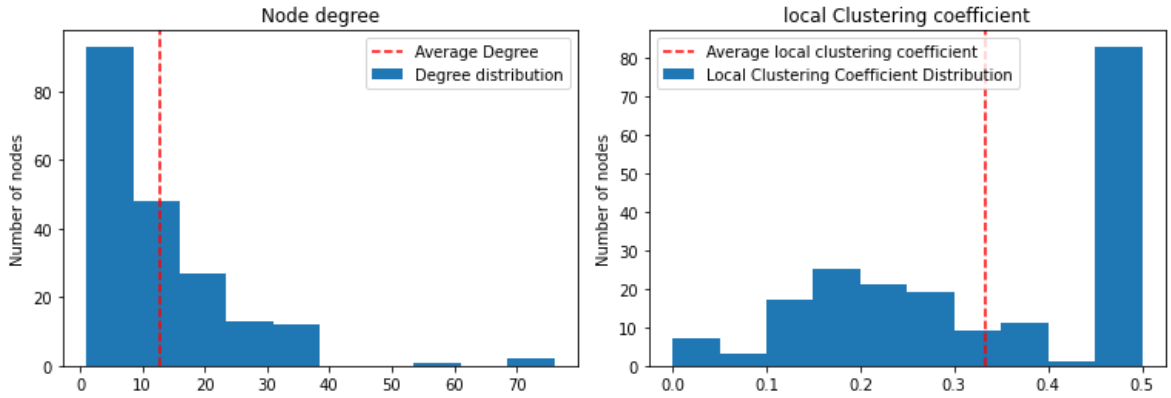

Figure 13: Degree Distribution(left) and Local Clustering Coefficient Distribution(right) of 200 lowly abundant species network

- Degree Distribution- Figure 13(left) denotes that almost 100 nodes have a degree of around 1 which signifies that they are connected to 1 other node in the network. Only some of the nodes have a high degree of 55 and 70 and the rest of the nodes have very less degree showing that the network is very sparse.

- Local Clustering coefficient distribution-Figure 13(right) denotes that around 80 have a clustering coefficient of 0.45 and this indicates that these nodes are 45% likely to form tight knit clusters within the network. Less number of nodes in the range of 0.05 to 0.1 and 0.4 to 0.45 possibly denotes that there are holes in the network and low clustering coefficient values for most of the nodes further depicts the sparsity of the network.

5.2.7 Network creation for the entire plant species dataset

Figure 14 demonstrates the entire network of plant species dataset. There are a lot of significant positive links and negative links within the network. The highly significant positive and negative links are depicted by Table 19 and Table 20.

Table 19 and Table 20 indicates same amount of z-scores. This clearly represents that the interactions can be of the same type and weight between different species. A lot of them are interconnected depicting that there could be many inferred interactions additionally in the network apart from the direct real interactions.

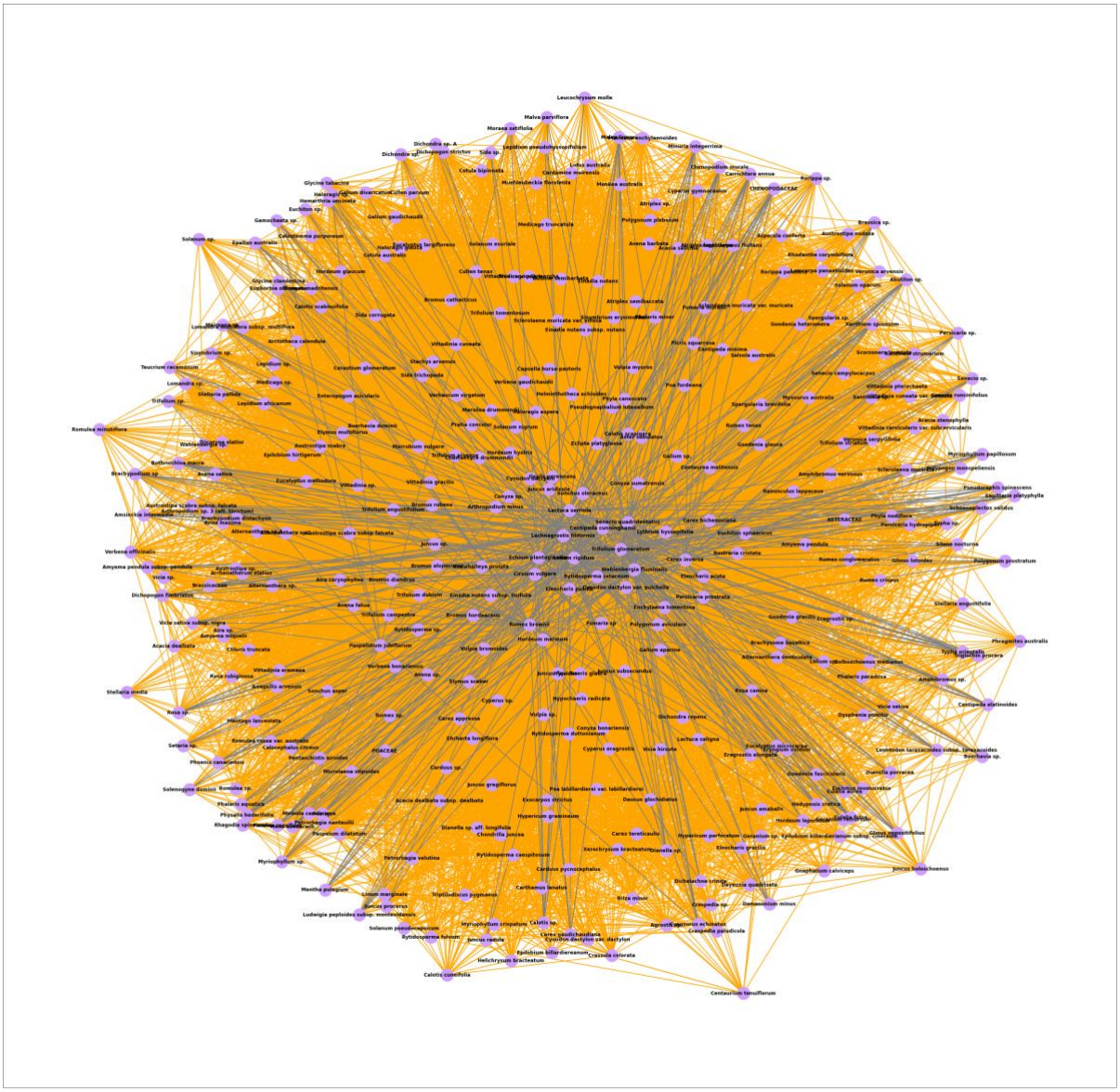

Figure 14: Network generated for entire plants species abundance dataset with threshold of 1, positive links (orange) and negative links (grey)

Table 19: Top 5 most negative links in the plant abundance data network

Source	Target	Weights
Paspalidium jubiflorum	Physalis hederifolia	-1.8125
Cynosurus echinatus	Paspalidium jubiflorum	-1.8125
Melia azedarach	Paspalidium jubiflorum	-1.8125
Paspalidium jubiflorum	Rorippa palustris	-1.8125
CHENOPODACEAE	Paspalidium jubiflorum	-1.8125

Table 20: Top 5 most positive links in the plant abundance data network

Source	Target	Weights
Persicaria sp.	Polypogon monspeliensis	12.2500
Dichopogon strictus	Galium divaricatum	12.2500
Chenopodium murale	Minuria integerrima	12.2500
Brassica sp.	Persicaria sp.	12.2500
Malva linnaei	Minuria integerrima	12.2500

5.2.8 Network analysis results for the entire plant species network(Table 21)

- The network indicates an edge density of 0.10105, which tells us that there is a possibility of 10.105 % connections between the nodes in the network.
- An average degree of 71.14447 denotes that a node in the network is typically connected to an average of 71 of other nodes.
- Average clustering coefficient of 0.33254 signifies that on an average each node tends to form triangles with their neighbors with a probability of 33.254 %.
- The table also indicates that there are three communities within the network, indicating a group of three densely connected clusters in the graph.
- 'Oxalis perennans' is the central node in the network through degree centrality measure, indicating that it is well connected to other nodes in the network.

Table 21:Network analysis results of entire plant abundance data network

Measure	Value
Edge Density	0.10105
Average Degree	71.14447
Average Clustering coefficient	0.332541
Number of communities	3
Central node	Oxalis perennans

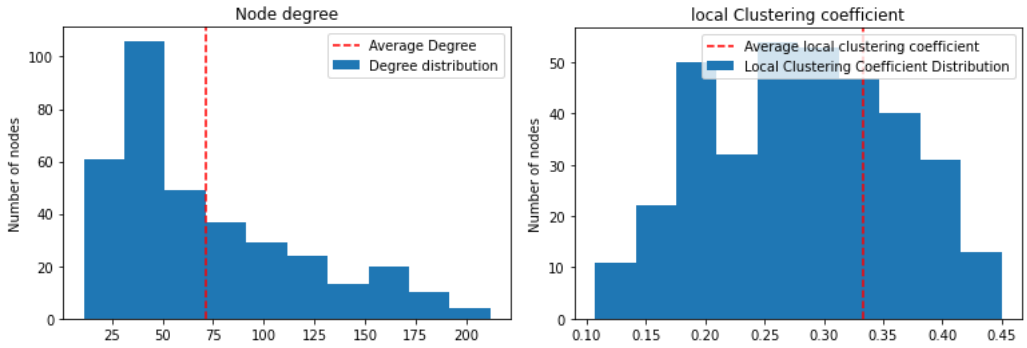

Figure 15: Degree Distribution(left) and Local Clustering Coefficient Distribution(right) of entire plant species network

- Degree Distribution- Figure 15(left) denotes that almost 100 nodes have a degree of around 30 which signifies that they are connected to 30 other nodes in the network. Only some of the nodes have a high degree of above 100.
- Local Clustering coefficient distribution-Figure 13(right) denotes that the network has a good clustering coefficient values. There is no sudden dip in the network which indicates the absence of any holes in the network.

Chapter 6 : Discussion

In this section, we will be discussing about the key findings from the project and the current limitations of the project so as to give a comprehensive evaluation on our progress. The section will also outline some future prospects of the project which could not be achieved in the current time.

One of the primary investigations we focussed on was regarding the comparison of ESABO and modified ESABO results. Through three interesting case investigations we were able to analyse that only one-third of the results showed significant variation of greater than 50 % relative percentage change in z-scores in both the methods. However, it was observable that original ESABO took more computation time than modified ESABO which even took up to half an hour for the calculation of scores in the entire plant abundance dataset. Thus, due to this reason we are interested in modified ESABO since it gives quicker results with a good approximation of the results of original ESABO method.

Secondly, we observed a massive increase in the z-scores when the sample pool was widened. The observed improvement in information gained, brought on by a more in-depth analysis of microbiome abundance, emphasises the significance of sample size in the interpretation of z-scores. This investigation highlights the importance of thorough data analysis and its ability to shed more light on complicated datasets.

Thirdly, an examination into the optimal thresholding for differences in compositionality of the datasets revealed significant results. It was our aim to check whether the optimal threshold value remains 1 as stated in the paper [16] or whether it varies according to data composition. The results for 50 high and low abundance microbiomes with high abundance in majority revealed an optimal threshold value of 50 with maximal information gain of 1083.484375. When 50 more low abundance microorganisms were added to the test dataset, it revealed a reduction in the threshold value to 40 with a higher information gain than before of 4424.968. It is also remarkable to notice from the line graph in Figure 5 (right) that, for this set of 100 high and low abundant species, there was rise in the information gain to 4399.140625 for a threshold of 15 even after the drop in information gain to 4373.968750 at a threshold of 30. This was unusual because the information gain is typically high for the optimal threshold value and does not see an increase once it is past the optimal threshold. This is an important finding from our investigation, which encourages us to adopt adjustable discrete thresholds for various dataset compositionality. This also motivates us to use a variable-threshold for varying data compositions and the usage of discretized Fuzzy logic [43]. Furthermore, this finding is fuelled by a reduction in threshold value to 5 with a much higher information gain of 9539.531250.

Our work recognizes the limitation of not considering variable thresholds and discrete intervals according to abundance strength due to time constraints. However, we have discussed in detail and presented results pertaining to different threshold values or a shift in the threshold value when a mixture of abundances is added in the data subset. Fuzzy logic [43] could be used to extend the work in the future and define discretized fuzzy sets in the data according to varying compositionality. This removes the doubt of a single threshold value for a data subset of diverse compositionality and further grouping or clustering of the data to categorical fuzzy sets like ‘Low’, ‘Medium’, ‘High’, etc. and finding optimal threshold for the same.

Another interesting point was that, even though a large set of low abundance data of 300 species with values ranging from 0 to 11 for microbiomes and 200 species with values ranging from 0 to 122 for plant species were taken, the optimal threshold value for low abundance data still pertained to the value of 1. The dataset's predominance of small numerical values relative to larger ones could be one explanation for this phenomena. The prevalence of small values may have an impact on the choice of threshold because a threshold of 1 efficiently captures these small values and makes ESABO method to be particularly an efficient algorithm for the low abundance segment, leading to a significant information gain.

However, this shift towards a lower threshold value of 1 further highlights the complexity of thresholding for the low abundance microbiomes or plants as in our project. Our study acknowledges the current restrictions in this field and, in light of this complexity, acts as a motivator for additional research. We think more research in this area is necessary to obtain deeper understanding and broaden the usefulness of thresholding techniques for low abundance datasets.

Network creation and analysis proved to be a critical tool in analysing the interactions that formed. Figure 7 enhances the claim that high abundance microbiome species may have a competitive interactions with the low abundance microbiomes which is signified by the network. However, in the case of plant abundance dataset, the absence of high abundance plants allows for more mutualistic interactions between the 200 low abundant species as in Figure 12. Our study is successful in predicting the interactions according to abundance which is backed by the results of network analysis and creation.

Chapter 7 : Conclusion

Our project has effectively demonstrated the application of ESABO into the domain of continuous or real valued datasets by employing optimal thresholding techniques to binarize them. It underscores the fact that 1 is not the optimal thresholding for all sorts of data and that the thresholding will vary according to differences in abundances of species. A high abundance of species shifts the threshold to a higher value and an addition of low abundance shifts the threshold to a lower value for attaining maximal information gain or recognize the significant links within the dataset. The discussion in the previous chapter explores more on the current limitations of our project in dealing with varying data compositionality and discretized fuzzy logic [43].

The modified ESABO method provides a more quicker computation of z-scores with high degree of similarity with respects to the results from original ESABO method. Its performance is well explored in our project in the context of deriving links, optimal thresholding and creation of networks. A substantial observation from our probe into sample sizes and the information gained therefrom is that larger sample sizes both increase the information gained and the ESABO approach's potential to identify significant links.

Furthermore, modified ESABO method has worked smoothly in the entire plant species abundance dataset and derived the significant interactions both positive and negative from the network obtained using z-score interpretation. This enables the application of ESABO approach to a broader perspective and dynamic ecosystems. Thus, ESABO method could be improved to be a vital tool in data science and computer research by enhancing it more into the area of complex datasets. It digs deep into the intricacies in the data and covers every detailed interactions which was ignored by most studies as described in Chapter II. It can be used as an effective tool in computer science to find the intricate web of relations in the data.

Additionally, the network science methods describes in background section, helps in further investigating into the topology and structure of the network formed via ESABO method focusing more on the network’s characteristics and possible conclusions about the network’s essential nature to form linkages, average number of linkages with the nodes, the key species and different communities present in the network. Thus, ESABO method in conjunction with network analysis tools can be a huge success in the field of data science and analysis.

Hence, in conclusion our study enhances the possibility of thresholding in the context of continuous valued datasets and details the impact that thresholding can have on gaining more information from the datasets. It also opens many future studies that can be explored in this direction and can provide meaningful contribution to science.

References

[1] P. Tiwari, S. K. Bose, and H. Bae, “Plant growth-promoting soil microbiomes: Beneficial attributes and potential applications,” *Sustainable Development and Biodiversity*, pp. 1–30, 2021. doi:10.1007/978-3-030-73507-4_1

[2] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight, “The impact of the gut microbiota on human health: An integrative view,” *Cell*, vol. 148, no. 6, pp. 1258-1270, 2012. doi:10.1016/j.cell.2012.01.035

[3] C. Zuñiga, L. Zaramela, and K. Zengler, “Elucidation of complexity and prediction of interactions in microbial communities,” *Microbial Biotechnology*, vol. 10, no. 6, pp. 1500–1522, 2017. doi:10.1111/1751-7915.12855

[4] L. Ghanbari Maman *et al.*, “Co-abundance analysis reveals hidden players associated with high methane yield phenotype in sheep rumen microbiome,” *Scientific Reports*, vol. 10, no. 1, 2020. doi:10.1038/s41598-020-61942-y

[5] M. Chung *et al.*, “Comparisons of oral, intestinal, and pancreatic bacterial microbiomes in patients with pancreatic cancer and other gastrointestinal diseases,” *Journal of Oral Microbiology*, vol. 13, no. 1, 2021. doi:10.1080/20002297.2021.1887680

[6] B. Shi *et al.*, “Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis,” *mBio*, vol. 6, no. 1, 2015. doi:10.1128/mbio.01926-14

[7] J. R. Galloway-Peña *et al.*, “Characterization of oral and gut microbiome temporal variability in hospitalized cancer patients,” *Genome Medicine*, vol. 9, no. 1, 2017. doi:10.1186/s13073-017-0409-1

[8] V. K. Ridaura *et al.*, “Gut microbiota from twins discordant for obesity modulate metabolism in mice,” *Science*, vol. 341, no. 6150, 2013. doi:10.1126/science.1241214

[9] W. E. Moore and L. H. Moore, “Intestinal floras of populations that have a high risk of colon cancer,” *Applied and Environmental Microbiology*, vol. 61, no. 9, pp. 3202–3207, 1995. doi:10.1128/aem.61.9.3202-3207.1995

[10] J. Ahn *et al.*, “Human gut microbiome and risk for colorectal cancer,” *JNCI: Journal of the National Cancer Institute*, vol. 105, no. 24, pp. 1907–1911, 2013. doi:10.1093/jnci/djt300

[11] P. J. Turnbaugh *et al.*, “A core gut microbiome in obese and Lean Twins,” *Nature*, vol. 457, no. 7228, pp. 480–484, 2008. doi:10.1038/nature07540

[12] G. Hajishengallis, R. P. Darveau, and M. A. Curtis, “The keystone-pathogen hypothesis,” *Nature Reviews Microbiology*, vol. 10, no. 10, pp. 717–725, 2012. doi:10.1038/nrmicro2873

[13] E. Y. Hsiao *et al.*, “Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders,” *Cell*, vol. 155, no. 7, pp. 1451–1463, 2013. doi:10.1016/j.cell.2013.11.024

[14] S. Wu *et al.*, “A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses,” *Nature Medicine*, vol. 15, no. 9, pp. 1016–1022, 2009. doi:10.1038/nm.2015

[15] K. Faust and J. Raes, “Microbial interactions: From networks to models,” *Nature Reviews Microbiology*, vol. 10, no. 8, pp. 538–550, 2012. doi:10.1038/nrmicro2832

[16] J. C. Claussen *et al.*, “Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome,” *PLOS Computational Biology*, vol. 13, no. 6, 2017. doi:10.1371/journal.pcbi.1005361

[17] I.-H. Mendler, B. Drossel, and M.-T. Hütt, “Microbiome abundance patterns as attractors and the implications for the inference of Microbial Interaction Networks,” arXiv.org, https://arxiv.org/abs/2306.02100
[18] M. S. Matchado *et al.*, “Network analysis methods for studying Microbial Communities: A mini review,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2687–2698, 2021. doi:10.1016/j.csbj.2021.05.001
[19] R. Vidanaarachchi, M. Shaw, S.-L. Tang, and S. Halgamuge, “Imparo: Inferring microbial interactions through parameter optimisation,” *BMC Molecular and Cell Biology*, vol. 21, no. S1, 2020. doi:10.1186/s12860-020-00269-y
[20] N. Metropolis and S. Ulam, *Monte Carlo method-a popular description*, 1949. doi:10.2172/4427100
[21] R. Vidanaarachchi, M. Shaw, S.-L. Tang, and S. Halgamuge, “Imparo: Inferring microbial interactions through parameter optimisation,” *BMC Molecular and Cell Biology*, vol. 21, no. S1, 2020. doi:10.1186/s12860-020-00269-y
[22] G. T.-W. Shaw, Y.-Y. Pao, and D. Wang, “Metamis: A metagenomic microbial interaction simulator based on Microbial Community Profiles,” *BMC Bioinformatics*, vol. 17, no. 1, 2016. doi:10.1186/s12859-016-1359-0
[23] K.-N. Tsai, S.-H. Lin, W.-C. Liu, and D. Wang, “Inferring microbial interaction network from microbiome data using RMN algorithm,” *BMC Systems Biology*, vol. 9, no. 1, 2015. doi:10.1186/s12918-015-0199-2
[24] Z. D. Kurtz *et al.*, “Sparse and compositionally robust inference of Microbial Ecological Networks,” *PLOS Computational Biology*, vol. 11, no. 5, 2015. doi:10.1371/journal.pcbi.1004226
[25] J. Friedman and E. J. Alm, “Inferring Correlation Networks from Genomic Survey Data,” *PLoS Computational Biology*, vol. 8, no. 9, 2012. doi:10.1371/journal.pcbi.1002687
[26] D. Gevers *et al.*, “The treatment-naive microbiome in new-onset crohn’s disease,” *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, 2014. doi:10.1016/j.chom.2014.02.005
[27] C. K. Fisher and P. Mehta, “Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression,” *PLoS ONE*, vol. 9, no. 7, 2014. doi:10.1371/journal.pone.0102451
[28] X. Gao, B.-T. Huynh, D. Guillemot, P. Glaser, and L. Opatowski, *Inference of significant microbial interactions from longitudinal metagenomics sequencing data*, 2018. doi:10.1101/305326
[29] R. E. Kass and L. Wasserman, “A reference bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 928–934, 1995. doi:10.1080/01621459.1995.10476592
[30] S. N. Steinway, M. B. Biggs, T. P. Loughran, J. A. Papin, and R. Albert, “Inference of network dynamics and metabolic interactions in the gut microbiome,” *PLOS Computational Biology*, vol. 11, no. 6, 2015. doi:10.1371/journal.pcbi.1004338
[31] N. Berestovsky and L. Nakhleh, “An evaluation of methods for inferring boolean networks from time-series data,” *PLoS ONE*, vol. 8, no. 6, 2013. doi:10.1371/journal.pone.0066031
[32] M. S. Matchado *et al.*, “Network analysis methods for studying Microbial Communities: A mini review,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2687–2698, 2021. doi:10.1016/j.csbj.2021.05.001

[33] R. R. Stein *et al.*, “Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota,” *PLoS Computational Biology*, vol. 9, no. 12, 2013. doi:10.1371/journal.pcbi.1003388

[34] S. Marino, N. T. Baxter, G. B. Huffnagle, J. F. Petrosino, and P. D. Schloss, “Mathematical modeling of primary succession of murine intestinal microbiota,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 1, pp. 439–444, 2013. doi:10.1073/pnas.1311322111

[35] P. Dam, L. L. Fonseca, K. T. Konstantinidis, and E. O. Voit, “Dynamic models of the complex Microbial Metapopulation of Lake Mendota,” *npj Systems Biology and Applications*, vol. 2, no. 1, 2016. doi:10.1038/npjsba.2016.7

[36] J. Mounier *et al.*, “Microbial interactions within a cheese microbial community,” *Applied and Environmental Microbiology*, vol. 74, no. 1, pp. 172–181, 2008. doi:10.1128/aem.01338-07

[37] D. Vo, S. C. Singh, S. Safa, and D. Sahoo, “Boolean implication analysis unveils candidate universal relationships in Microbiome Data,” *BMC Bioinformatics*, vol. 22, no. 1, 2021. doi:10.1186/s12859-020-03941-4

[38] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communicaton*. Urbana: University of Illinois Press, 1949.

[39] Hütt M-T, C, Lesne A. Gene regulatory network. Dissecting structure and dynamics. In: Wolkenhauer O, editor. Systems Medicine: Integrative, Qualitative and Computational Approaches. Reference Modules in Biomedical Sciences (peer-reviewed, accepted 20/06/2019). Amsterdam: Elsevier; 2020. p. 77-85.

[40] J. Lloyd-Price *et al.*, “Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases,” *Nature*, vol. 569, no. 7758, pp. 655–662, 2019. doi:10.1038/s41586-019-1237-9

[41] E. Pruesse *et al.*, “Silva: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB,” *Nucleic Acids Research*, vol. 35, no. 21, pp. 7188–7196, 2007. doi:10.1093/nar/gkm864

[42] J. Golbeck, “Network structure and measures,” *Analyzing the Social Web*, pp. 25–44, 2013. doi:10.1016/b978-0-12-405531-5.00003-1

[43] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. doi:10.1016/s0019-9958(65)90241-x

[44] A.-L. Barabási, “Network science,” BarabásiLab, http://networksciencebook.com/

[45] D. N. Reshef *et al.*, “Detecting novel associations in large data sets,” *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011. doi:10.1126/science.1205438

Appendix A: GitLab Repository

The project is accessible through the GitLab repository link below:

https://git.cs.bham.ac.uk/projects-2022-23/dxc287

There are three files on the repository for the project.

1) The python file under the name- 'Dissertation_finalcode.ipynb'
2) The first dataset for use under the name- 'taxonomic_profiles.xlsx'
3) The second dataset for the project under the name – 'plant_species.xlsx'

The code must be run on a python 3 environment and a GPU access would speed up the computations to some extent.

Appendix B: Figures

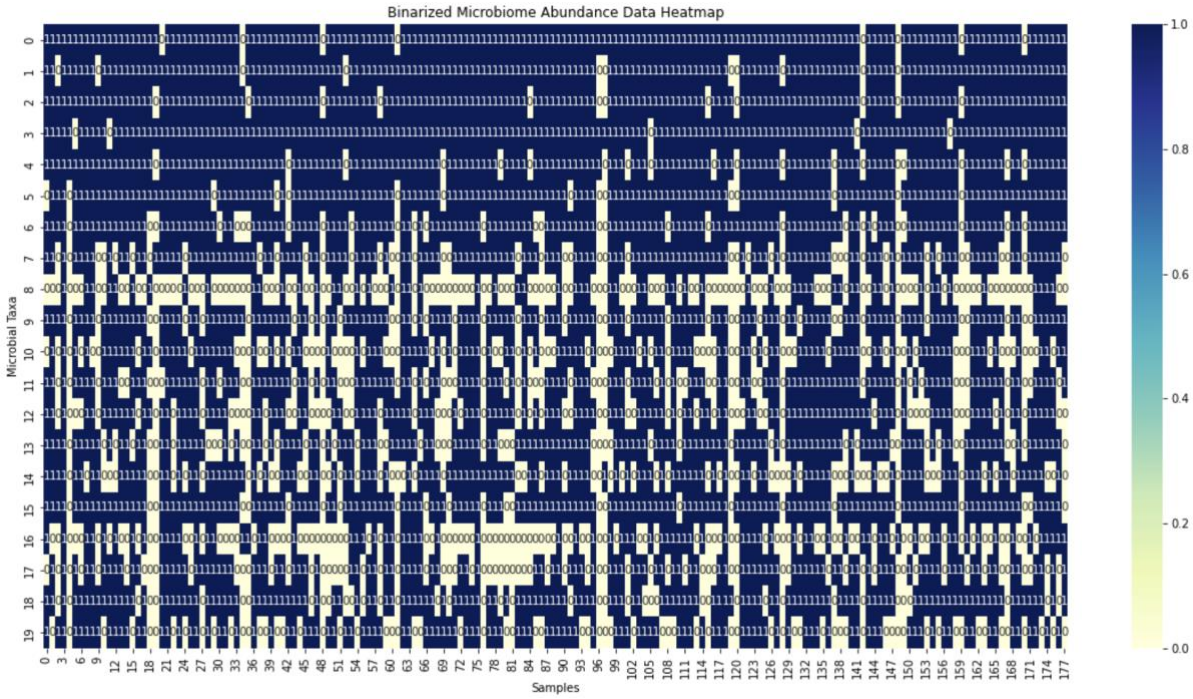

Figure 16: Heat map of 20 highly abundant species with a threshold of 1.

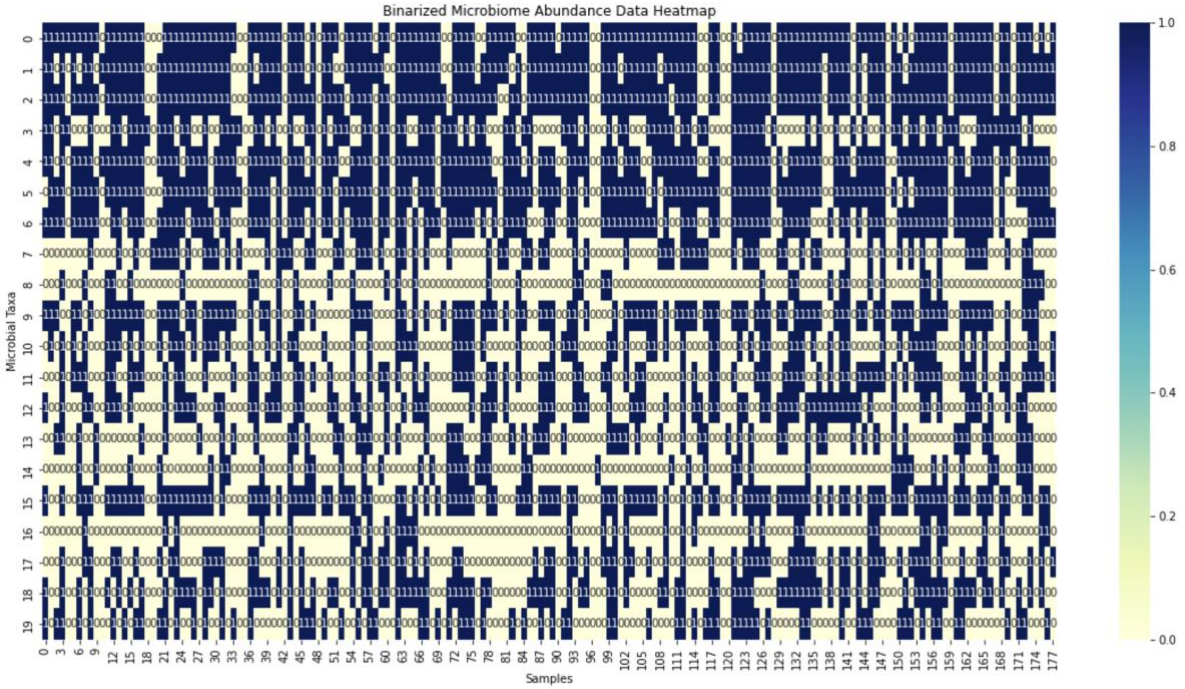

Figure 17: Heat map of 20 high abundant species with optimal threshold

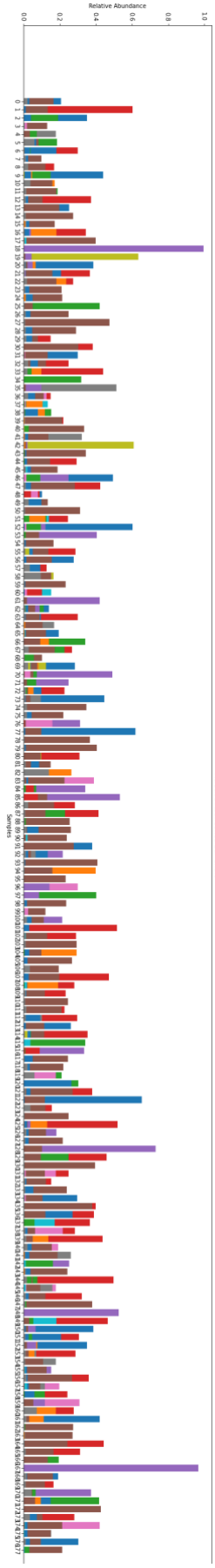

Figure 18: Stacked bar plot showing the relative abundances of species in the microbiome