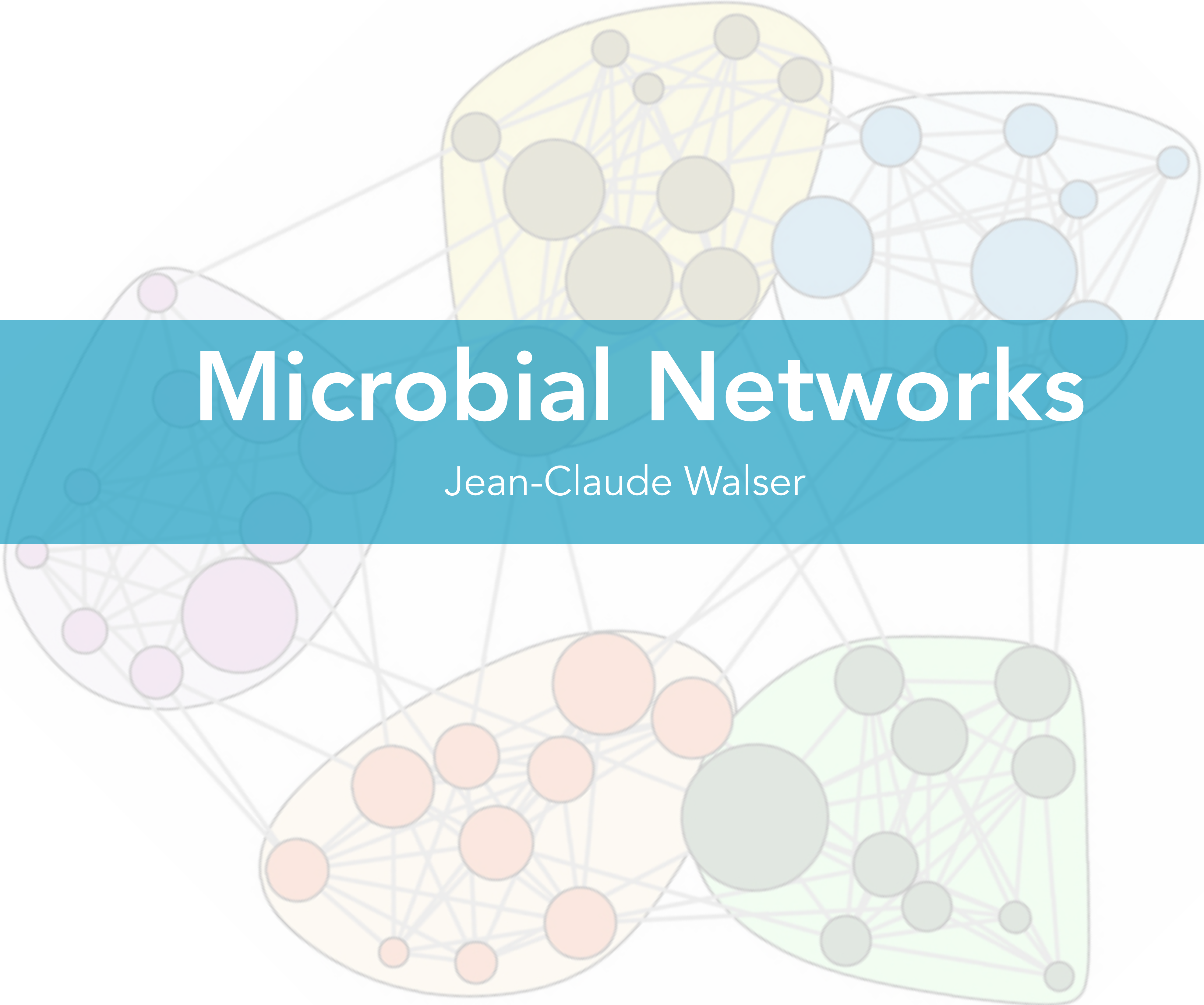
Microbiota Data Analysis 2020 Workshop



Microbial Networks

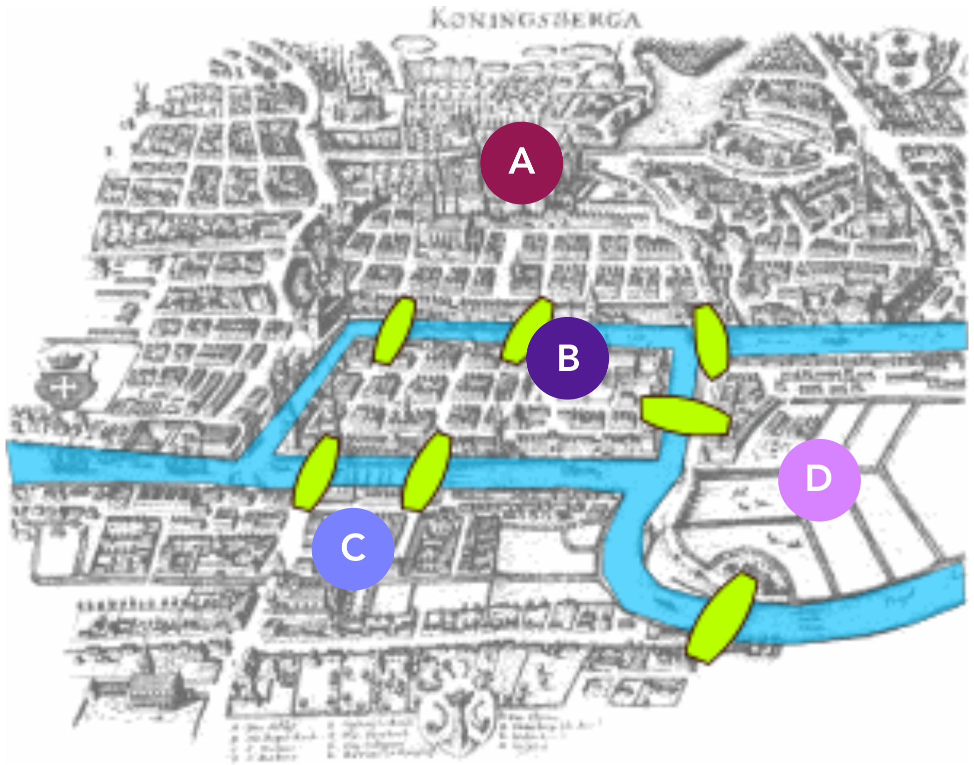
Jean-Claude Walser

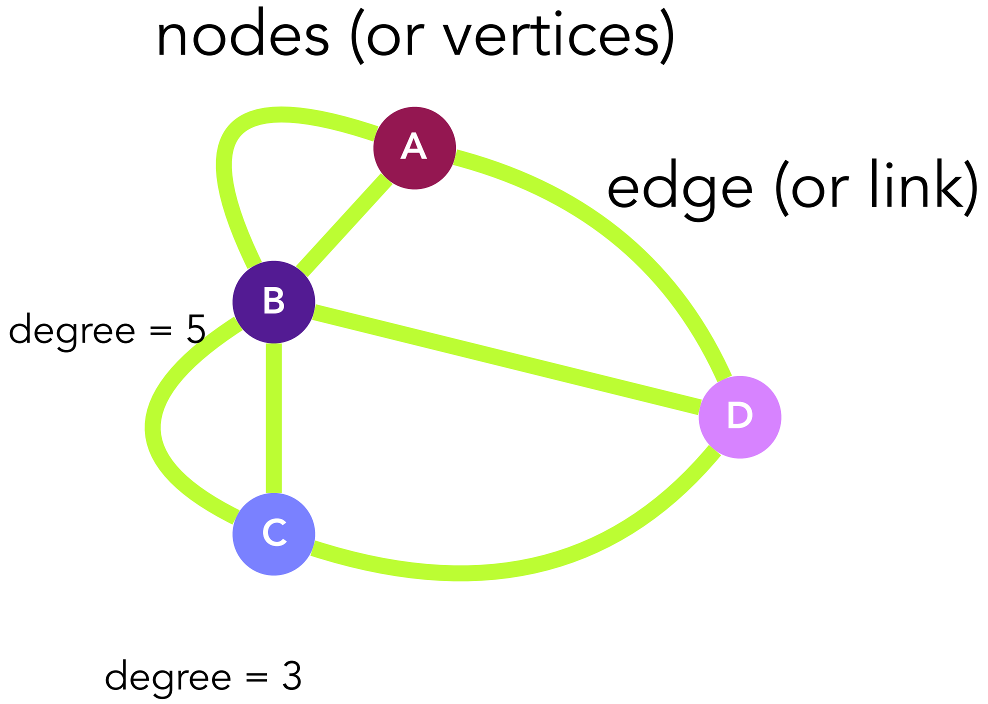
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

GDC
Genetic Diversity Centre Zurich

MDA ▷ Microbiota Network

The Seven Bridges of Königsberg is a historically notable problem in mathematics. Its negative resolution by Leonhard Euler (1736) laid the foundations of **graph theory.**





nodes (or vertices)

edge (or link)

degree = 5

degree = 3

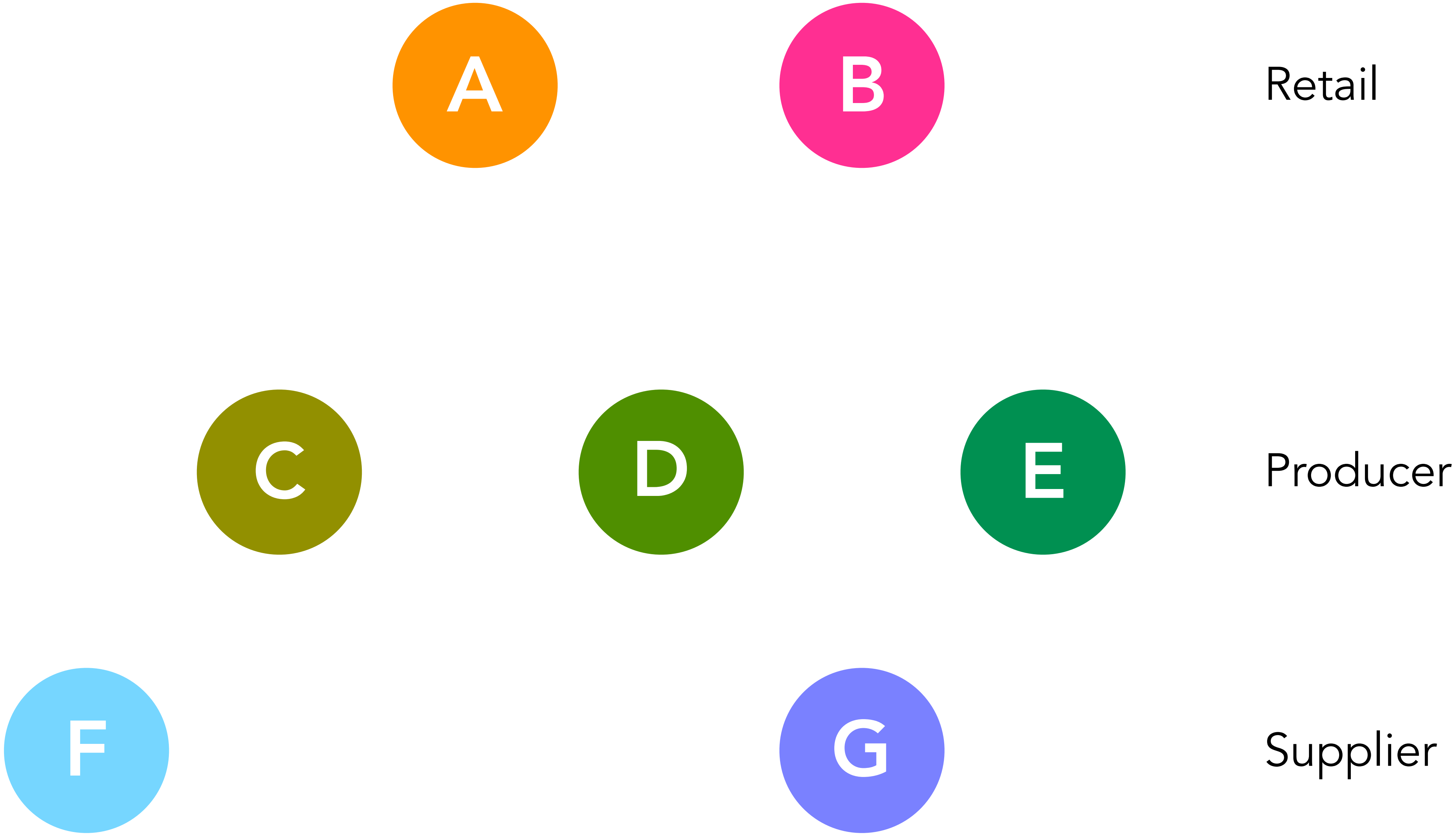
An **edge** is a visual representation of a relation (associations). It is a line that connects two nodes. Edges can be undirected or directed.
A **node** is a visual representation of an entity (e.g. species/OTU or sample).

Euler, Leonhard (1736). "Solutio problematis ad geometriam situs pertinentis". Comment. Acad. Sci. U. Petrop 8, 128–40.

MDA ▷ Microbiota Network

Retail

MDA ▷ Microbiota Network

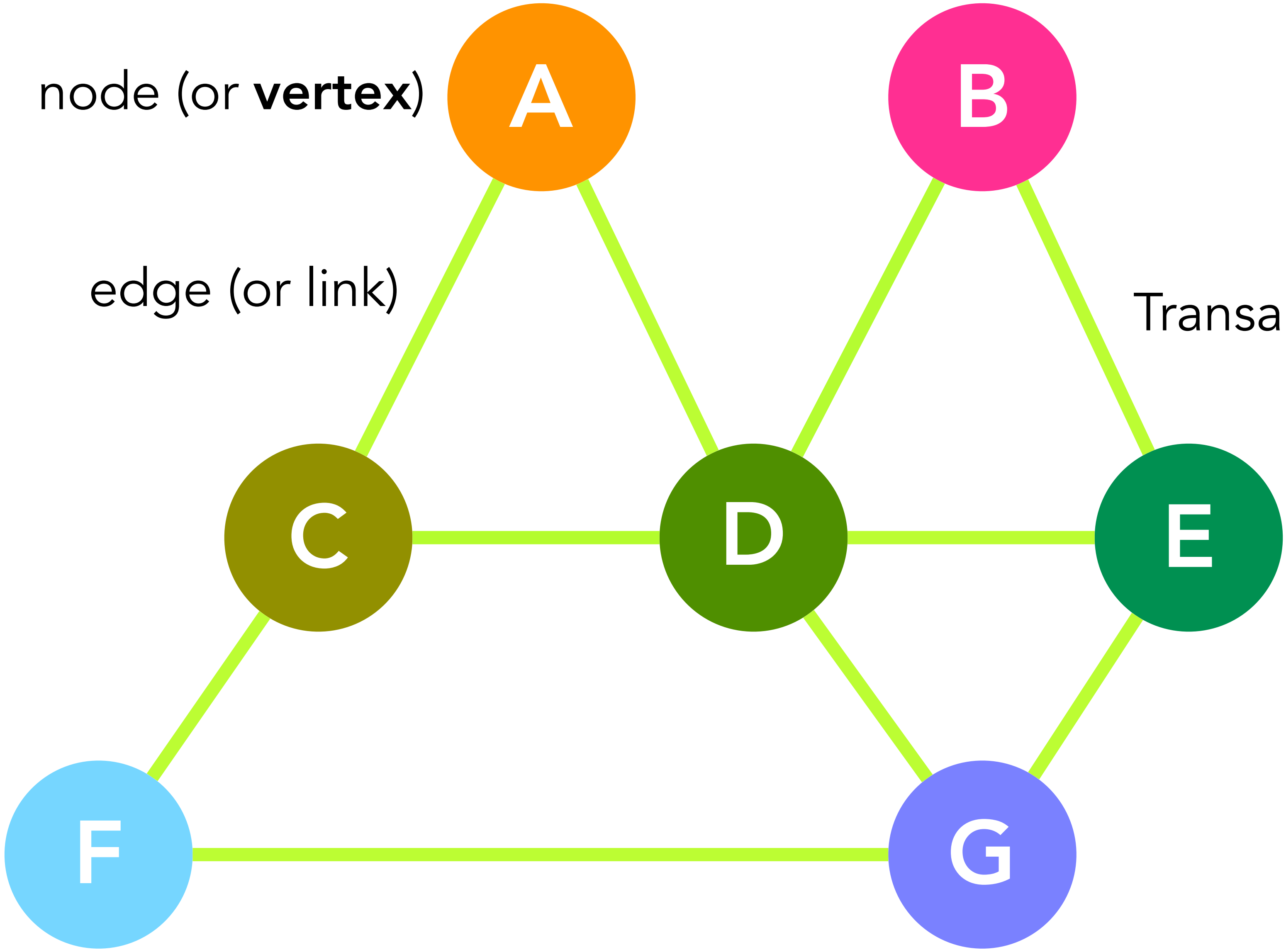


Retail

Producer

Supplier

MDA ▷ Microbiota Network



node (or **vertex**)

edge (or link)

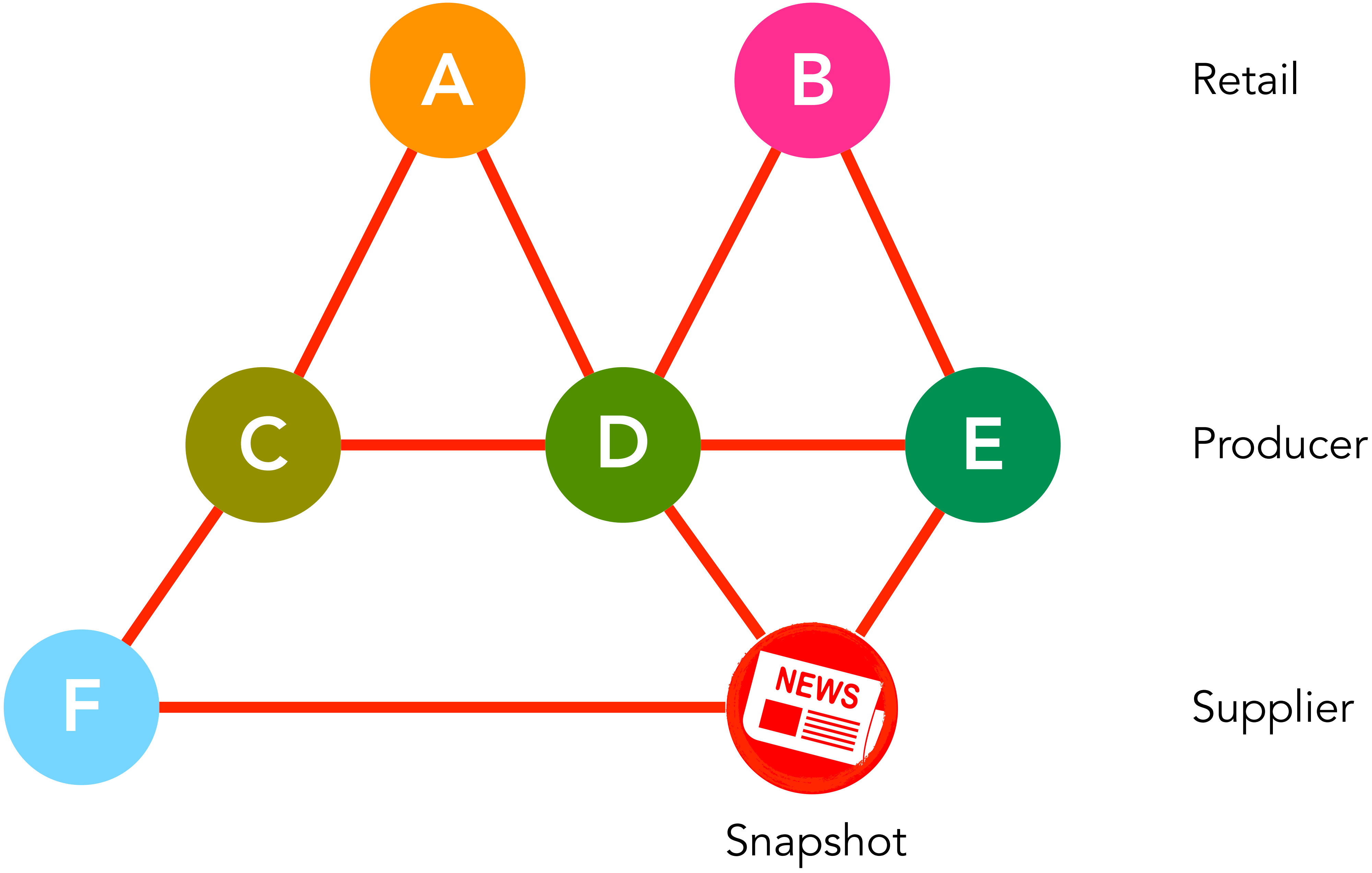
Retail

Transactions (undirected)

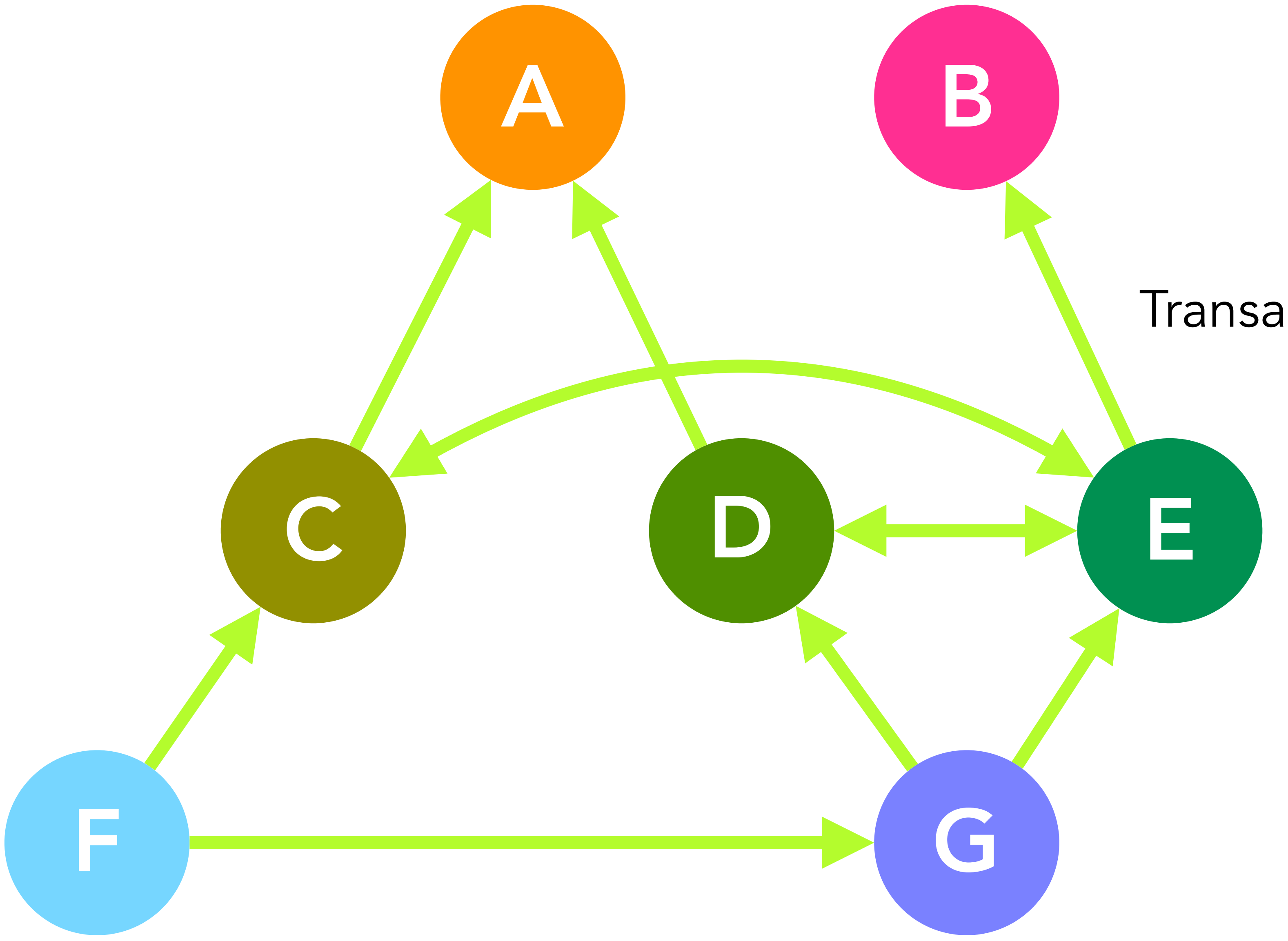
Producer

Supplier

MDA ▷ Microbiota Network



MDA ▷ Microbiota Network



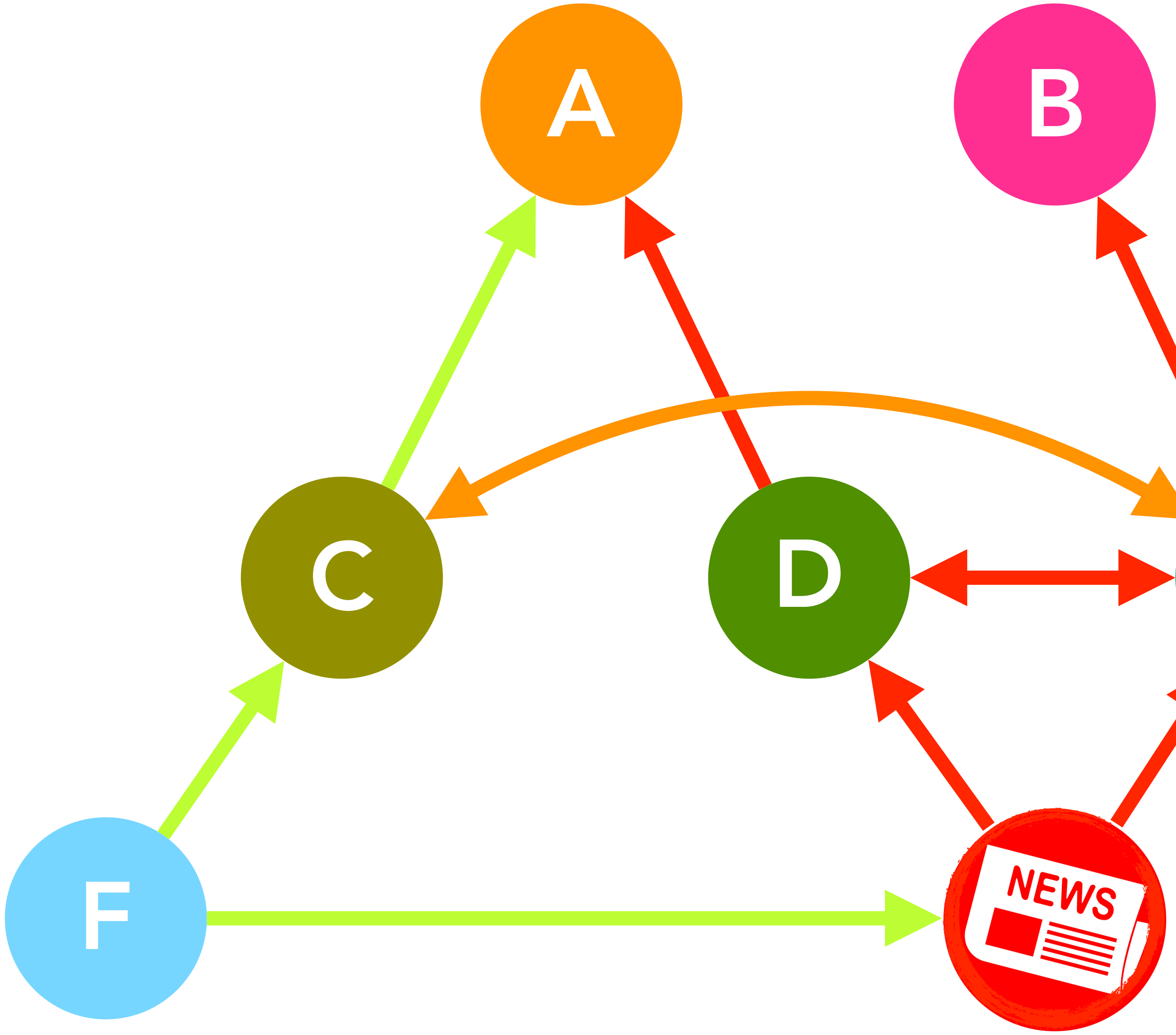
Retail

Transactions (**directed**)

Producer

Supplier

Microbiota Network



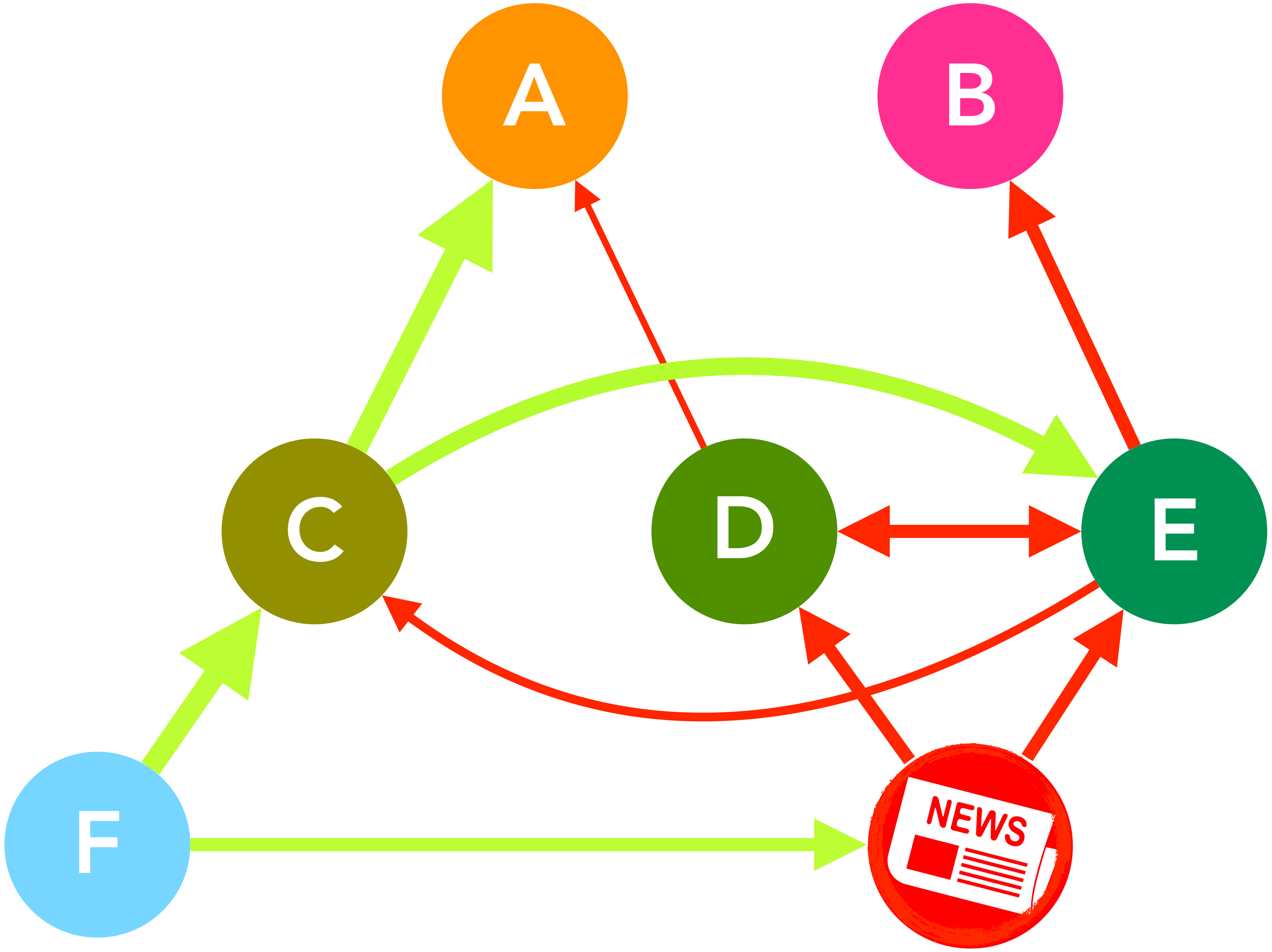
Retail

Transactions (**directed**)

Producer

Supplier

MDA ▷ Microbiota Network

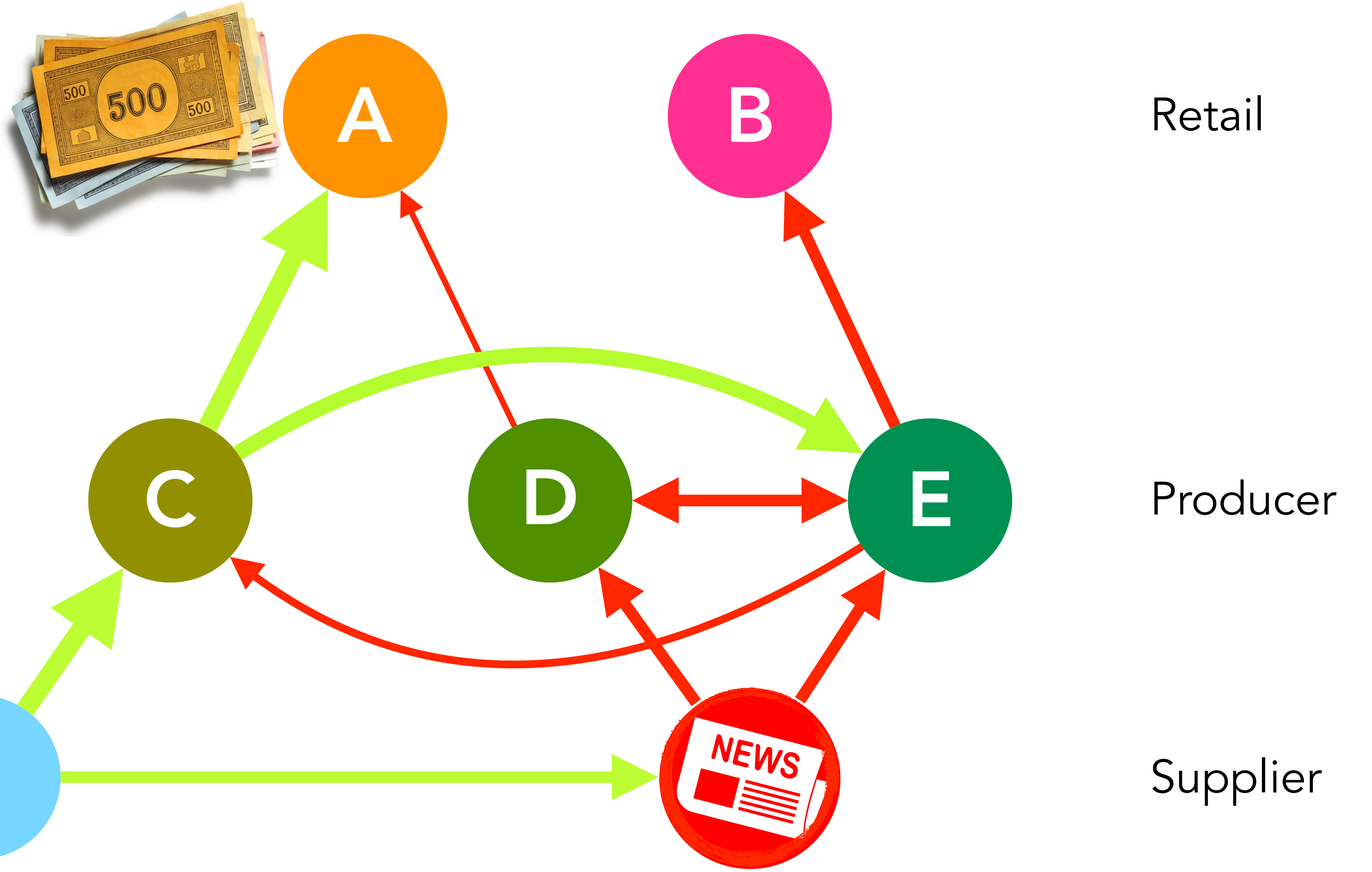


Retail

Producer

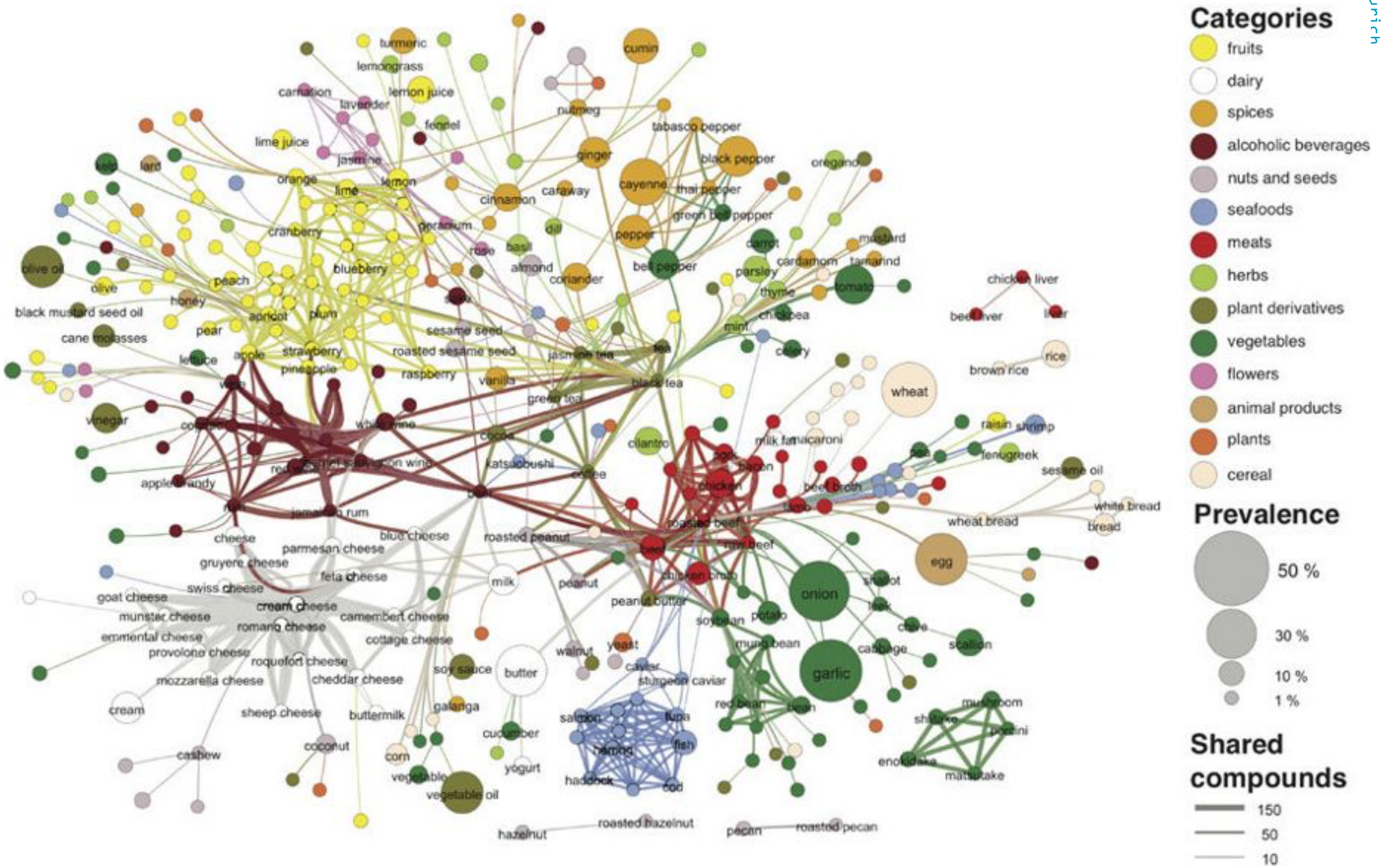
Supplier

Microbiota Network

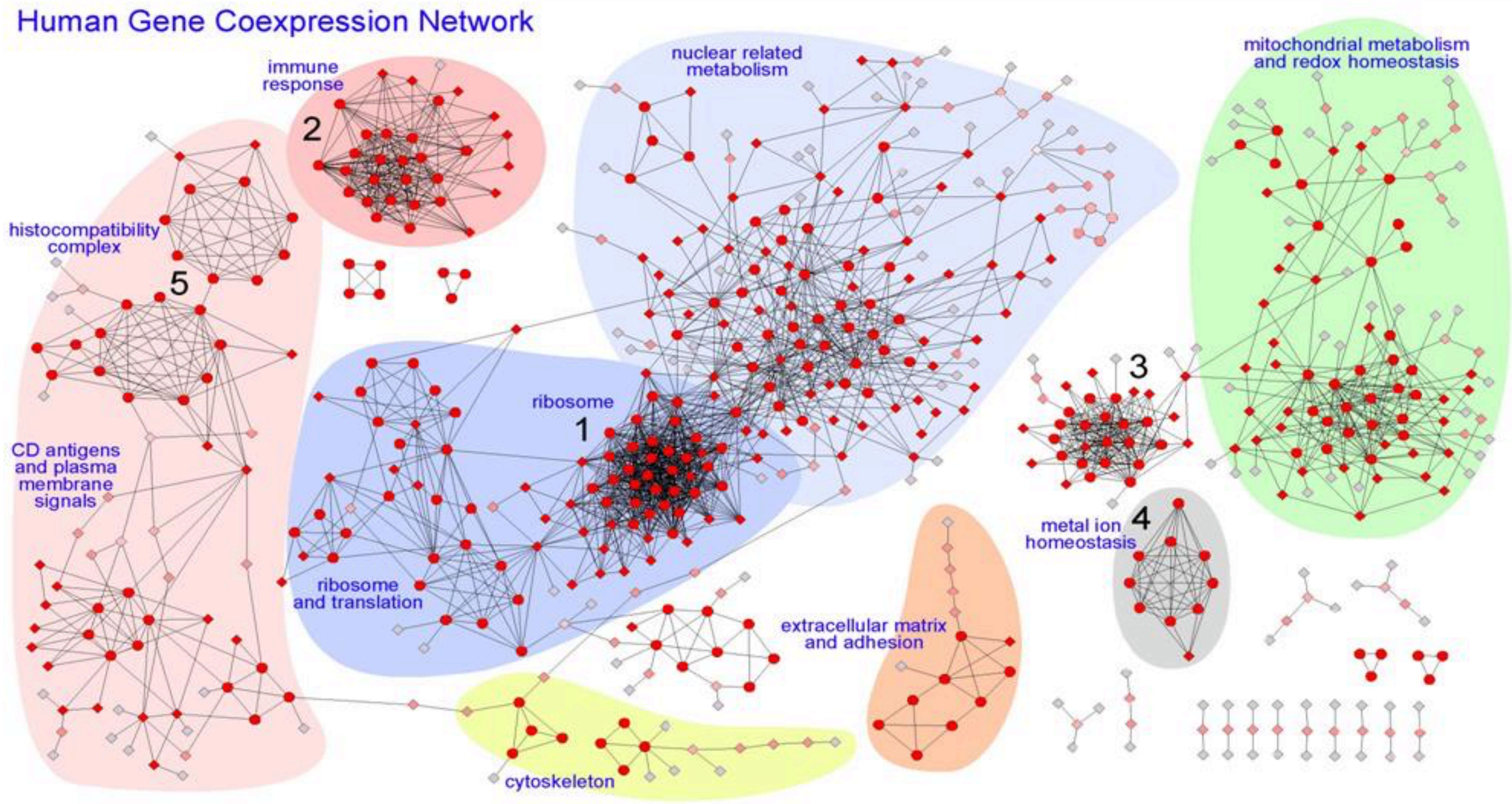


Networks are interesting because of their specific structural patterns, and how those structures affect the members of the network. Stated more simply, networks affect their members based on where those members are located in the networks.

MDA ▷ Microbiota Network



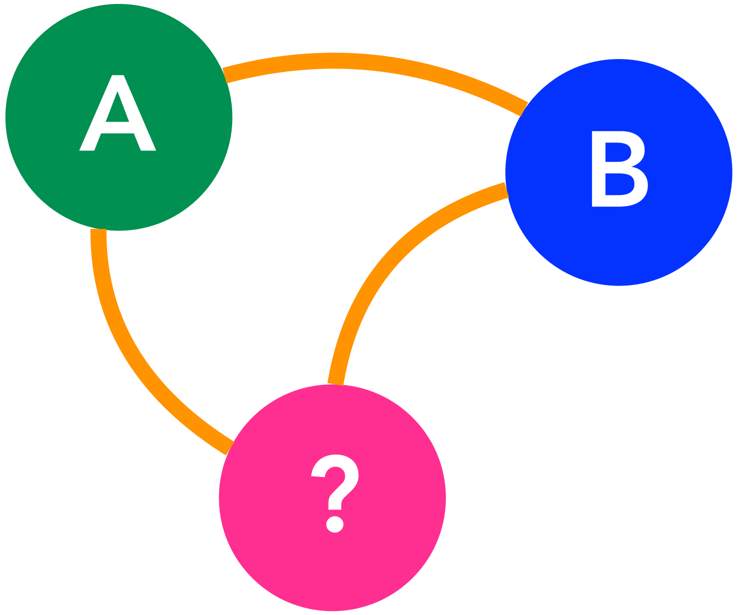
Each node denotes an ingredient, the node color indicates food category, and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients. The map shows only the statistically significant links. Ahn et al. (2011) Flavor network and the principles of food pairing. Scientific Reports volume 1, Article number: 196.


Human Gene Coexpression Network
immune response
2
histocompatibility complex
5
nuclear related metabolism
mitochondrial metabolism and redox homeostasis
ribosome
1
3
CD antigens and plasma membrane signals
ribosome and translation
metal ion homeostasis
4
extracellular matrix and adhesion
cytoskeleton

Graphical view of the coexpression network where the nodes correspond to genes and the edges to coexpression links. The network was produced as the intersection of two datasets (MAS5-Spearman and RMA-Pearson datasets with PPV > 0.60) to provide a confident human coexpression network that includes 615 genes and 2190 pairwise coexpression interactions. The most significant regions have been marked with background colors and labels. The color of the nodes (from red to grey) and the shape (circles or diamonds) were obtained following MCODE algorithm, being: circular nodes, the ones found with high cluster coefficient; diamond nodes, the ones with lower cluster coefficient; intensity of red color, degree of clustering changing till pale grey for the most peripheral nodes that only have one link.

Prieto et al. (2008) Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLOS ONE.

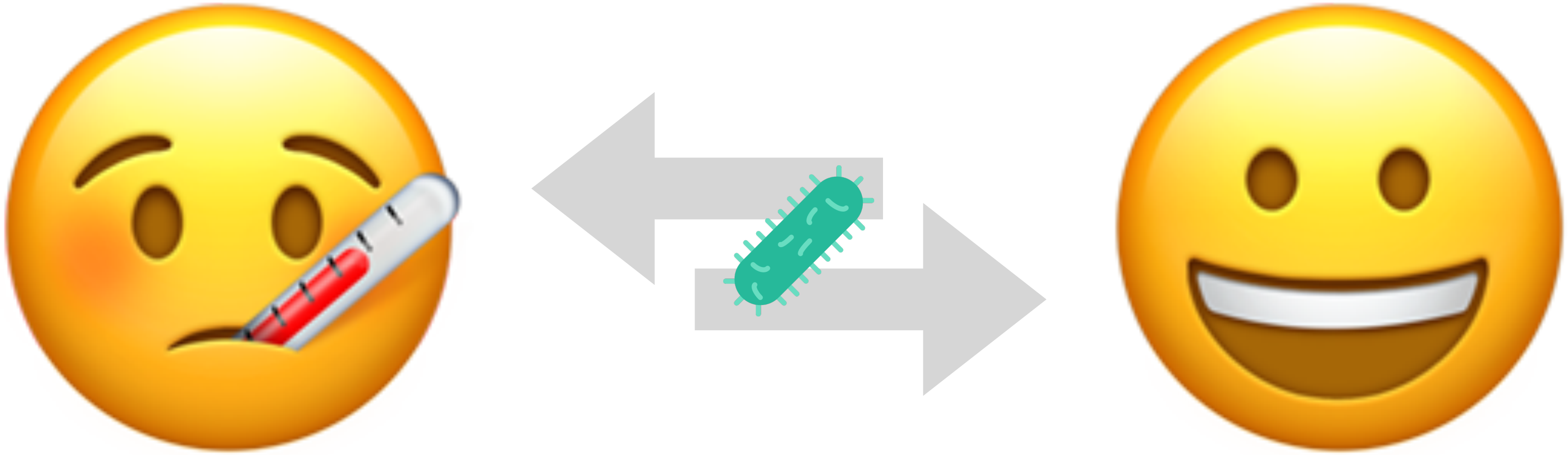
Gene co-expression networks (GCNs) are transcript–transcript association networks, generally reported as **undirected graphs**, where genes are connected when an appreciable co-expression association between them exists.



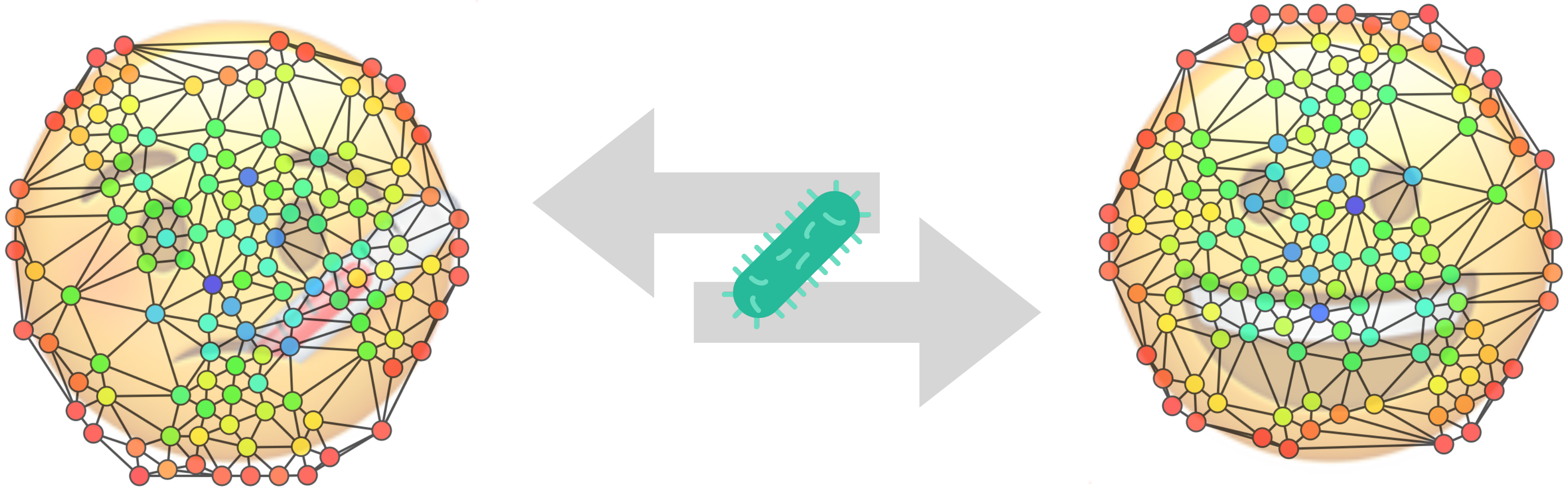
GCNs are built from gene expression data by calculating co-expression values in terms of **pairwise gene similarity score and choosing a significance threshold**. Debates on normalization methods, co-expression correlation (e.g., based on Pearson's or Spearman's correlation measures, among others) and significance and relevance are still alive and ongoing. Graphical Gaussian Models are also popular in the field of GCNs, the key idea behind them being the use of partial correlations, able to discriminate between direct and indirect interactions, as a degree of independence of any pair of genes. Other strategies for GCN inference include edge removal based on gene triplets analysis (e.g., ARACNE), regression methods and Bayesian networks (aimed to discover the best gene set predictor of a target gene expression).

Source: Tieri et al. (2019) Network Inference and Reconstruction in Bioinformatics. Encyclopedia of Bioinformatics and Computational Biology.

MDA ▷ Microbiota Network



Microbiomes are complex microbial communities. **Interactions** influence the structure and the function of such communities.



"Our world is dominated by, and wholly dependent on, complex microbial communities (i.e., microbiota) that are not a mere collection of independent individuals but a **complex of interconnected ecological communities** that communicate, cross-feed, recombine, and coevolve."

— Layeghifard et al. (2016) Trends in Microbiology

"Using **network theory**, one can model and **analyze a microbiome** and all its **complex interactions** in a single network. An interesting aspect of network theory is that the architectural features of networks appear to be **universal to most complex systems**, such as microbiomes, molecular interaction networks, computer networks, microcircuits, and social networks."

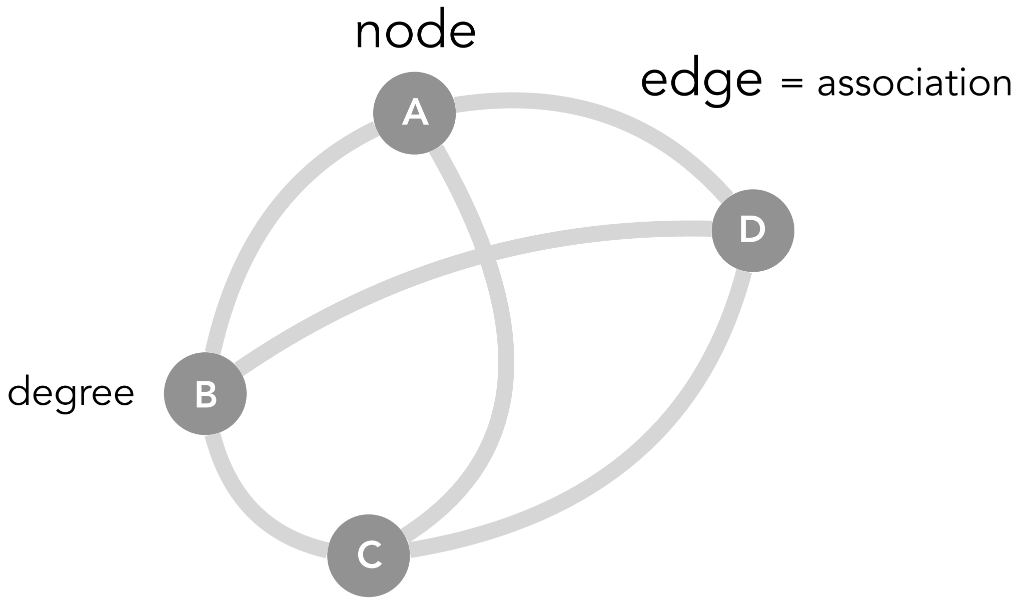
Layeghifard et al. (2018) Constructing and Analyzing Microbiome Networks in Microbiome Analysis: Methods and Protocols, Methods in Molecular Biology, vol. 1849.

"**Biological networks** are useful **visualizations of microbiome data**, as they can handle both their **scale** and **diversity**. In addition to data visualization, a major strength of networks is their ability to **represent emergent properties**. Emergent properties are those that would not be observed if parts of the network are investigated on their own. These properties may help explain the behaviour of complex systems, such as their apparent robustness or modularity."

Röttjers and Faust (2018) From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiology Reviews, 42(761–780).

"The primary tool for network scientists is network analysis, which is a set of methods that are used to (1) **visualize networks**, (2) **describe specific charachteristics** of overall network structure as well as details about the individual nodes, ties, and subgroups within the networks, and (3) build **mathematical and statistical models** of network structures and dynamics."

Luke (2015) A User's Guide to Network Analysis in R. Springer ISBN: 3319238825

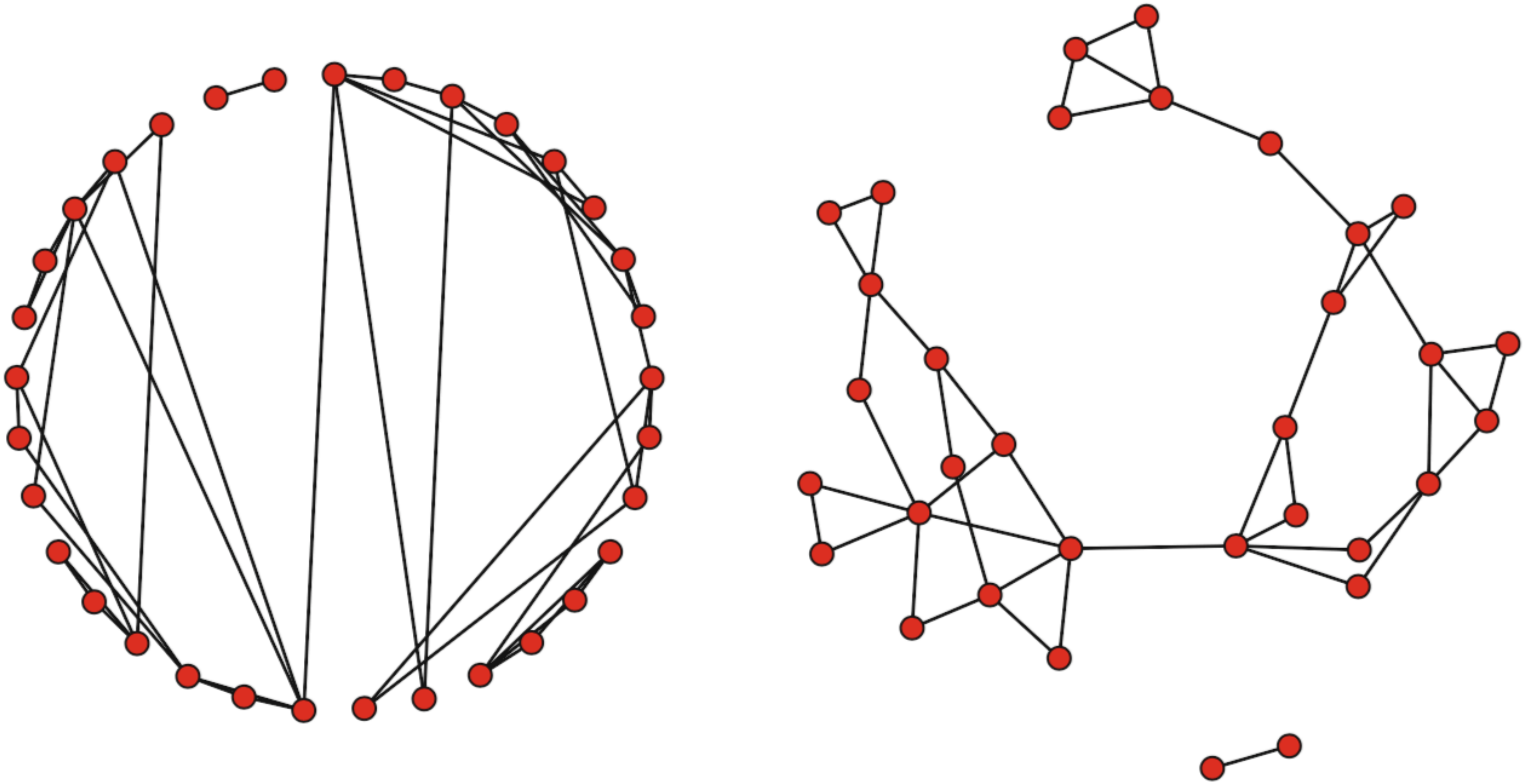


n	nID	nP
A	OTU1	K-P-C-F-G-S
B	OTU2	K-P-C-F-G-S
C	OTU3	K-P-C-F-G-S
D	OTU4	K-P-C-F-G-S

Microbial networks are **temporary or spatial snapshots of ecosystems** that are made up of two components: **nodes** and **edges**. Nodes usually represent **microbes** (OTUs), but they can also represent other **variables of interest**. Edges represent statistically significant **associations** between nodes, with the number of edges connected to a node referred to as the node's **degree**.

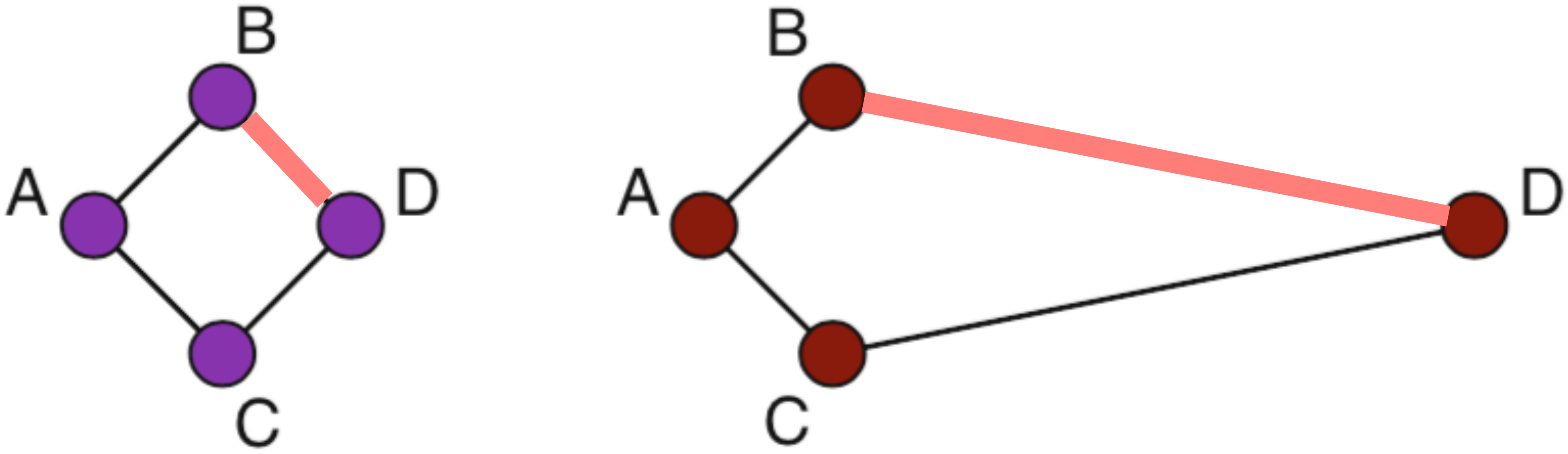
Röttjers and Faust (2018) From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiology Reviews, 42(761–780).

MDA ▷ Microbiota Network



At first glance it may appear that the figures are showing two quite different networks. In fact, they are two different visual representations of the same underlying network. **Same network but different layouts**.

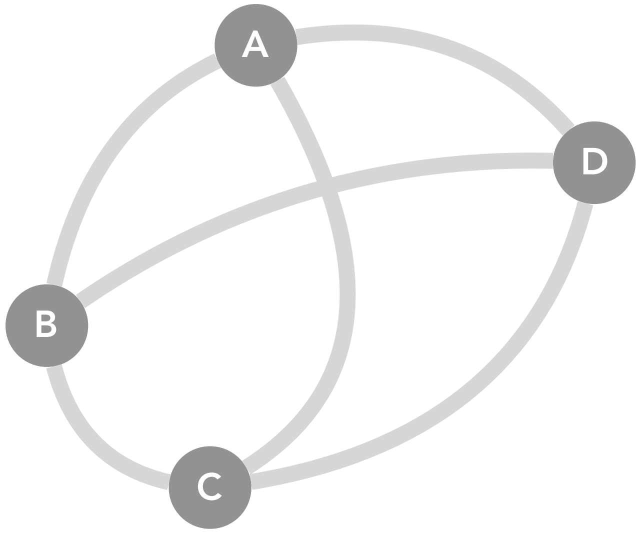
Luke (2015) A User's Guide to Network Analysis in R. Springer ISBN: 3319238825

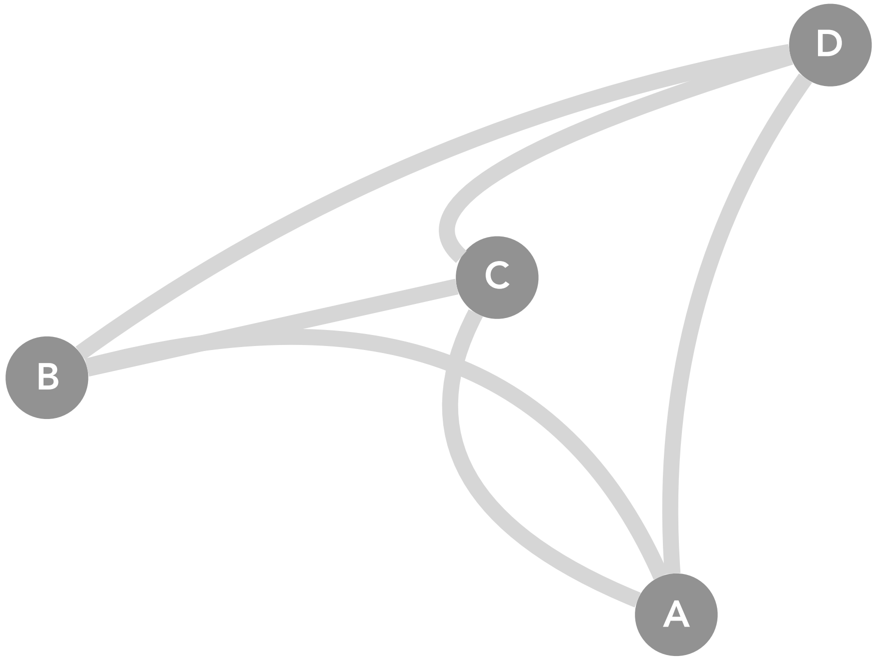


The ties (edge) indicate which nodes are adjacent to one another but the length of each line does not communicate any substantive information.

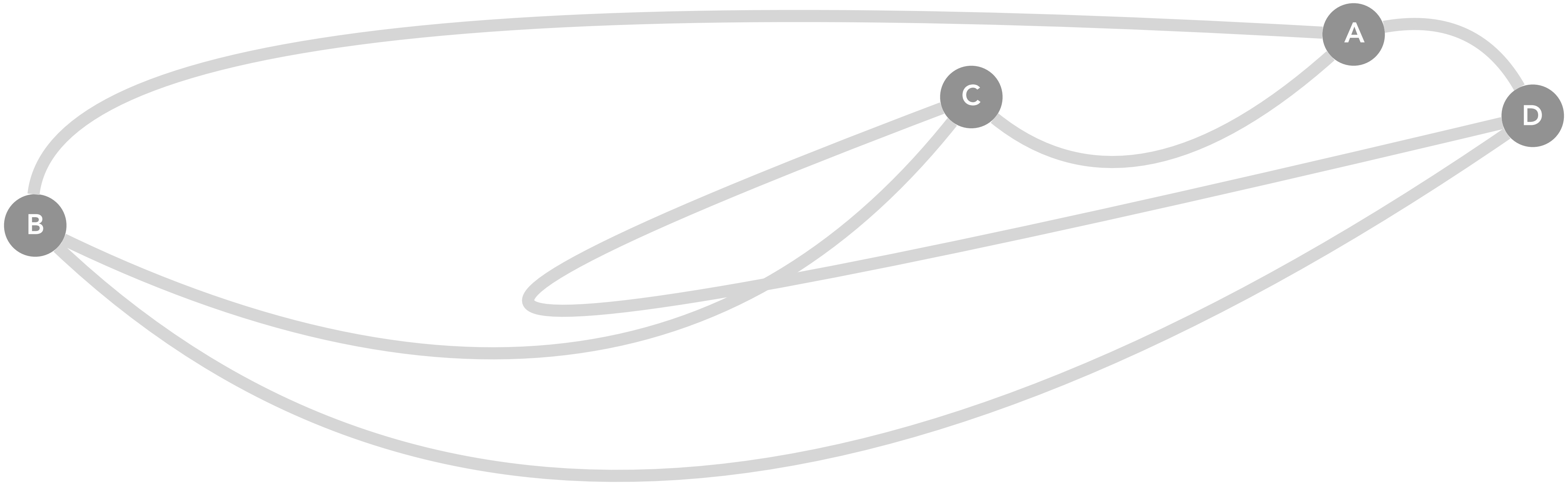
Luke (2015) A User's Guide to Network Analysis in R. Springer ISBN: 3319238825

MDA ▷ Microbiota Network

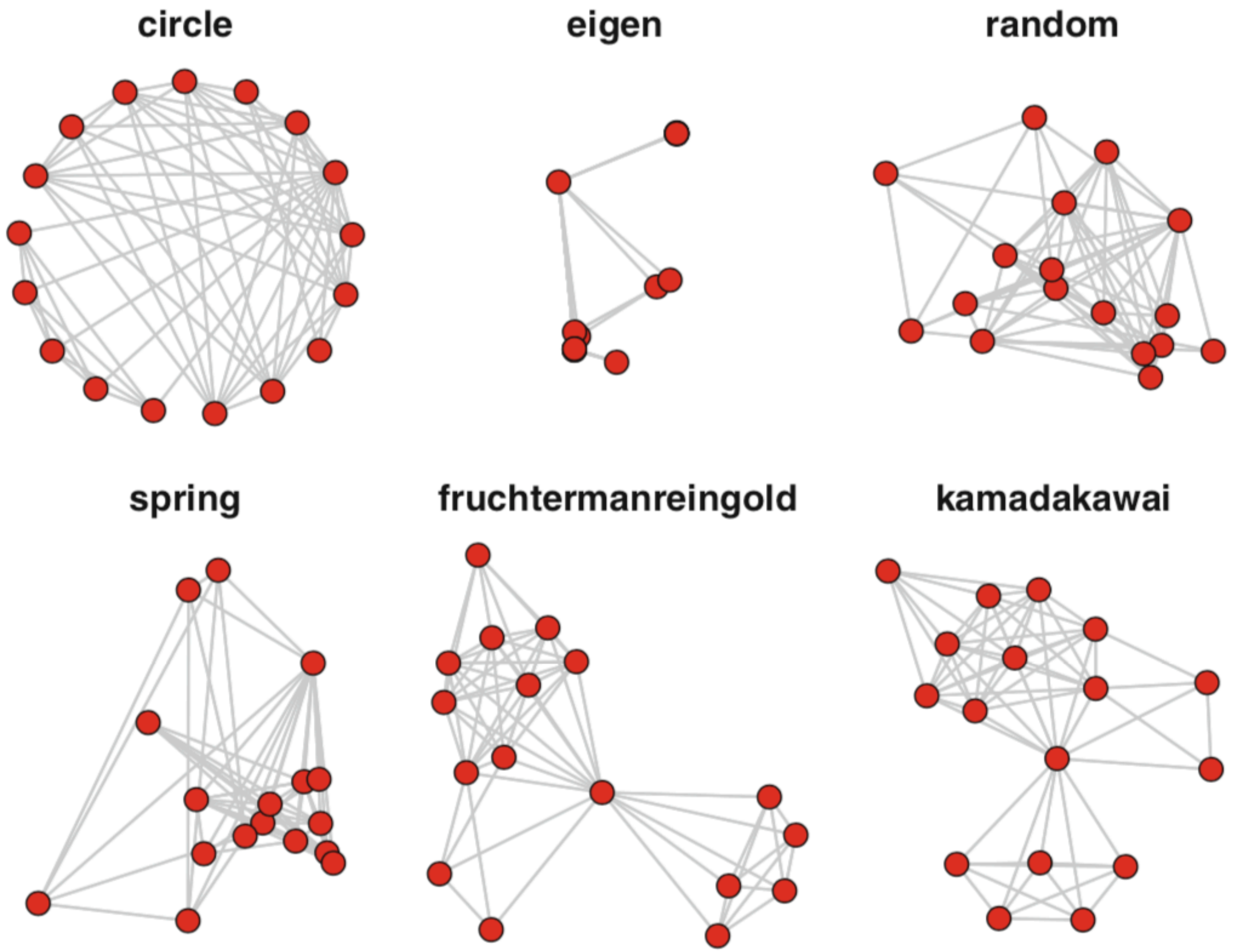




The length of the lines (edges) and the layout have no real meaning.

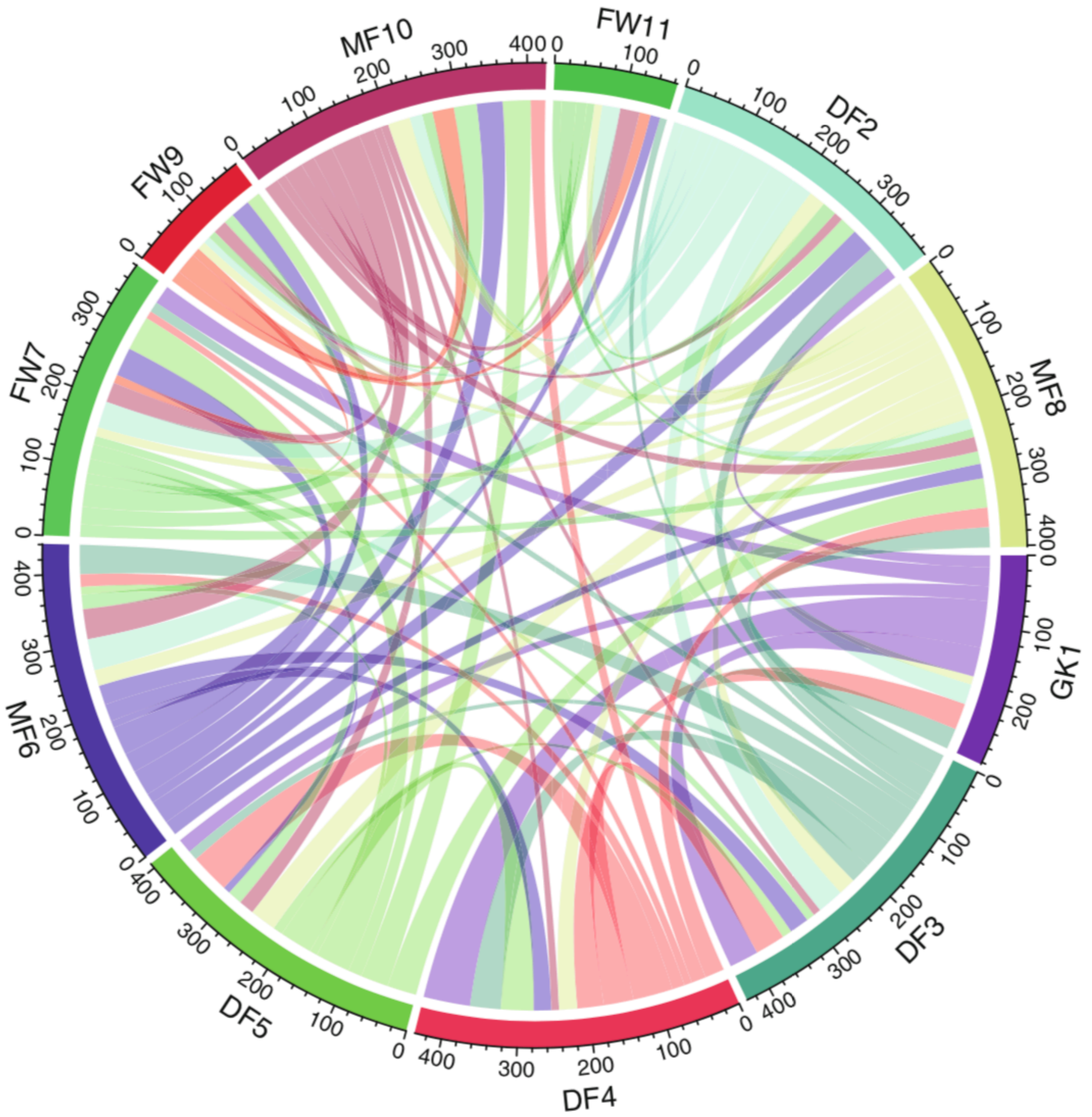


Microbiota Network



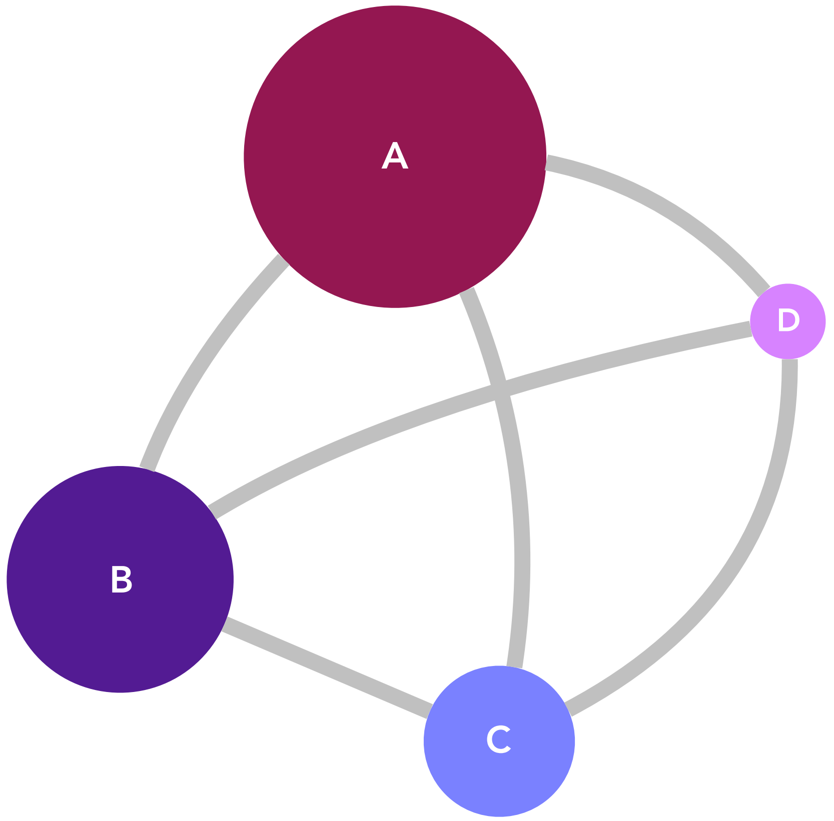
Aesthetics of network layouts - Although there are not in fact an infinite number of ways to display a network on a screen, the number of possibilities might as well be.

Luke (2015) A User's Guide to Network Analysis in R. Springer ISBN: 3319238825



Specialiced network diagrams: chord diagram.

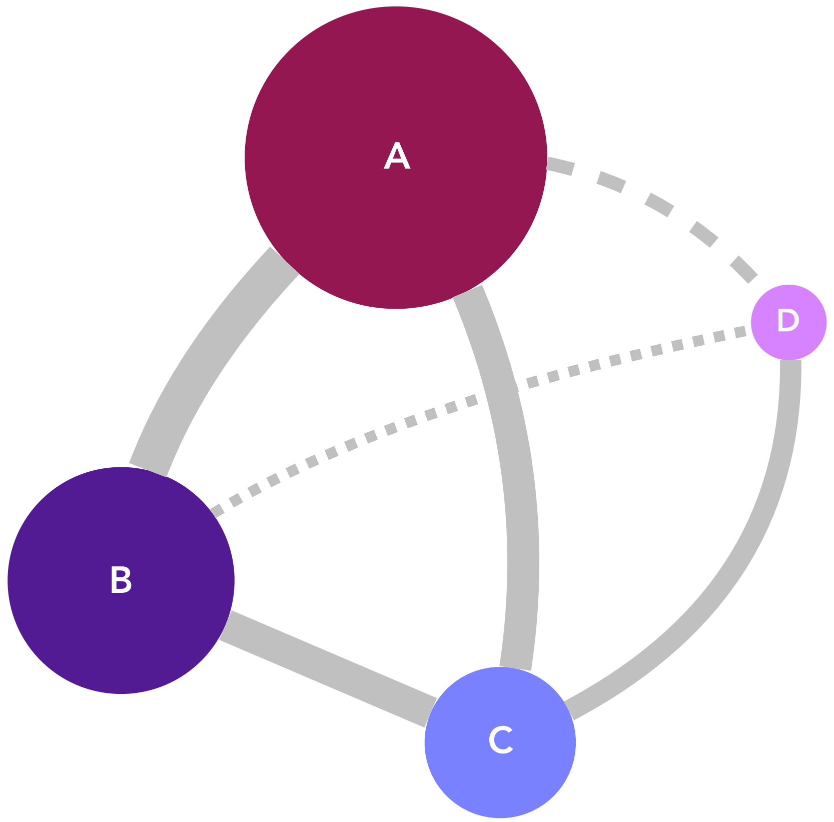
Luke (2015) A User's Guide to Network Analysis in R. Springer ISBN: 3319238825



node: color, shape, size, label

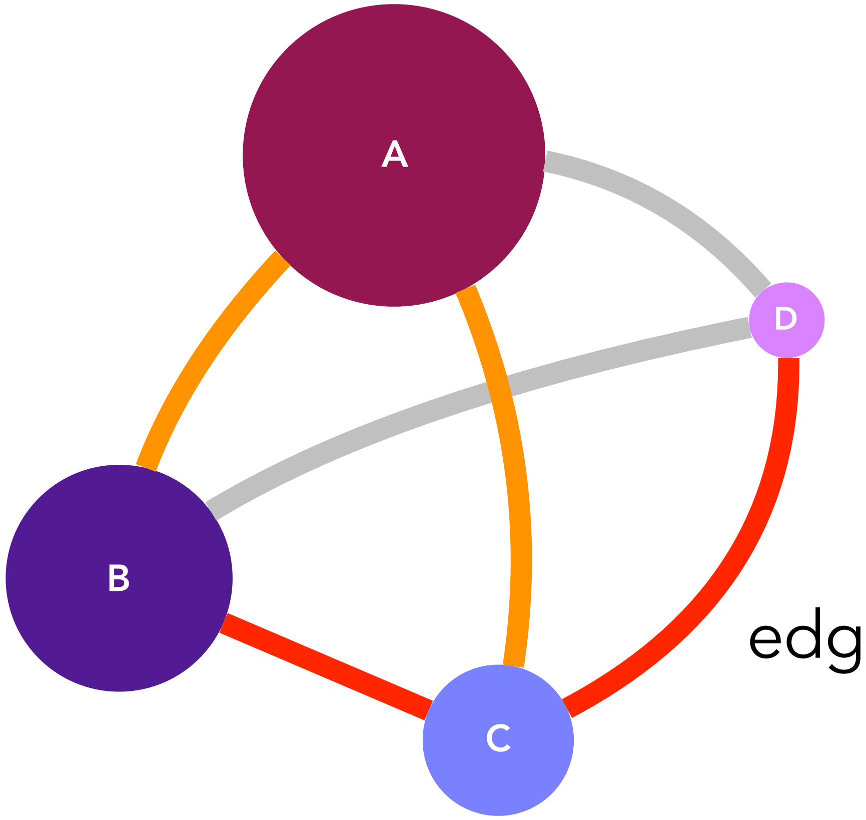
Node attributes

n	nL	d	p	**abundance**
A	OTU1	3	K-P-C-F-G-S	**0.4**
B	OTU2	3	K-P-C-F-G-S	**0.3**
C	OTU3	3	K-P-C-F-G-S	**0.2**
D	OTU4	3	K-P-C-F-G-S	**0.1**

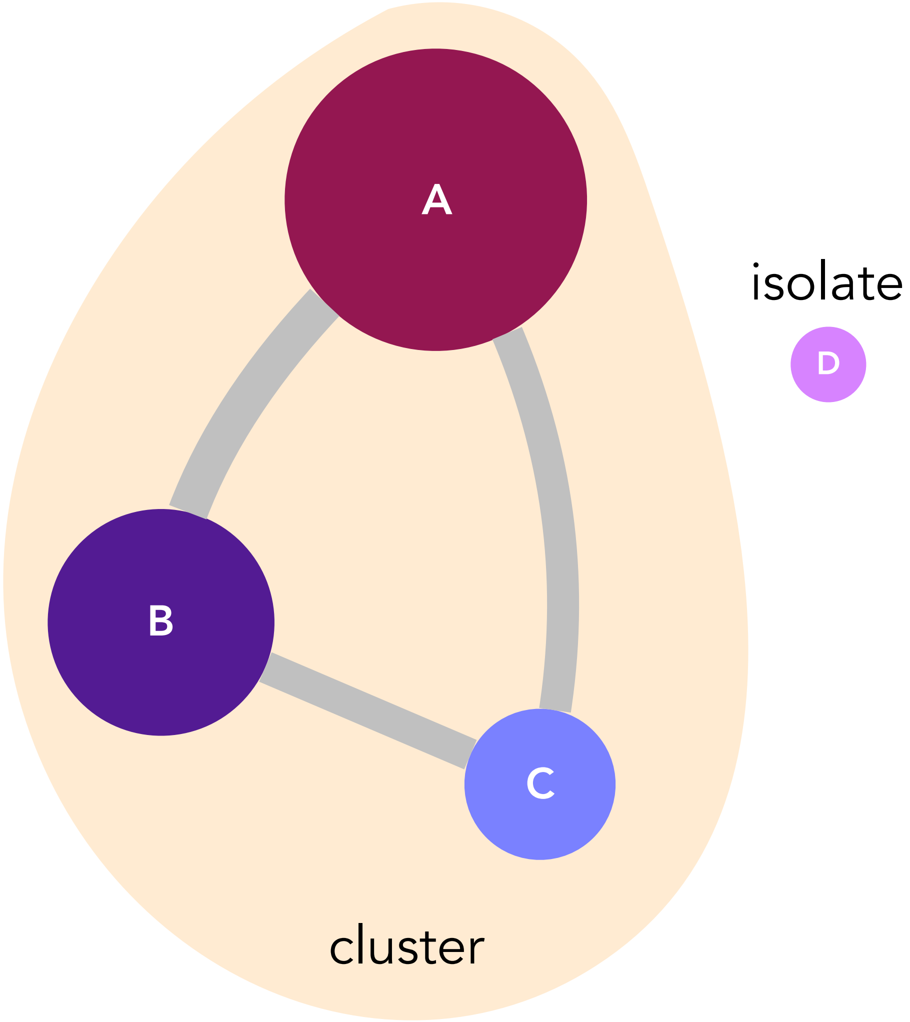


Edge attributes

pairs	**m**
A/B	**f(AB)**
A/C	**f(AC)**
A/D	**f(AD)**
B/C	**f(BC)**
B/D	**f(BD)**
C/D	**f(CD)**

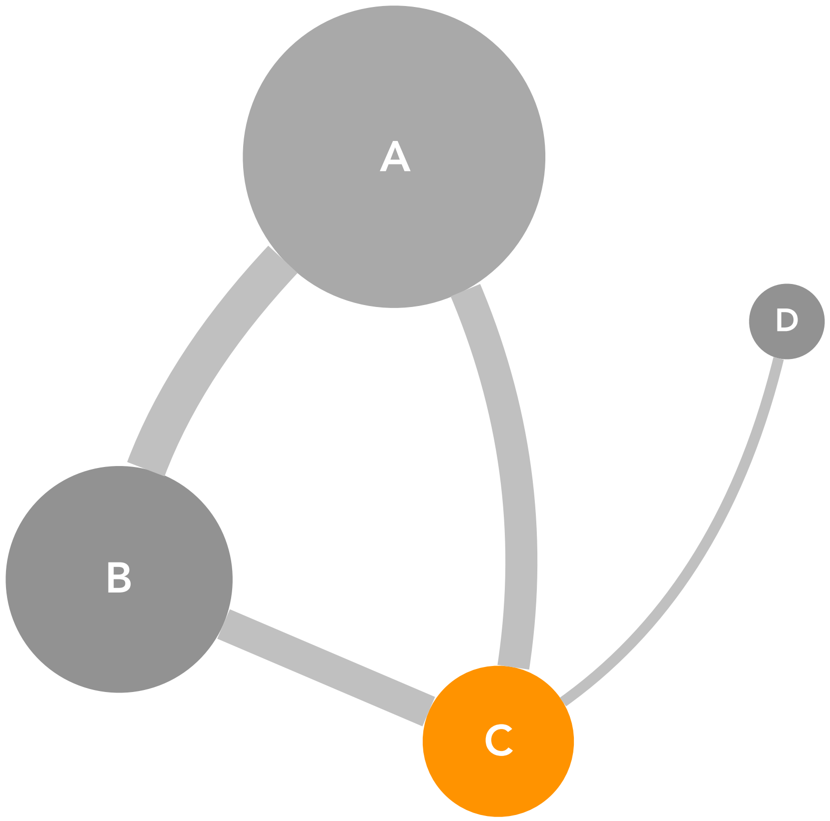


edge: width, color, type



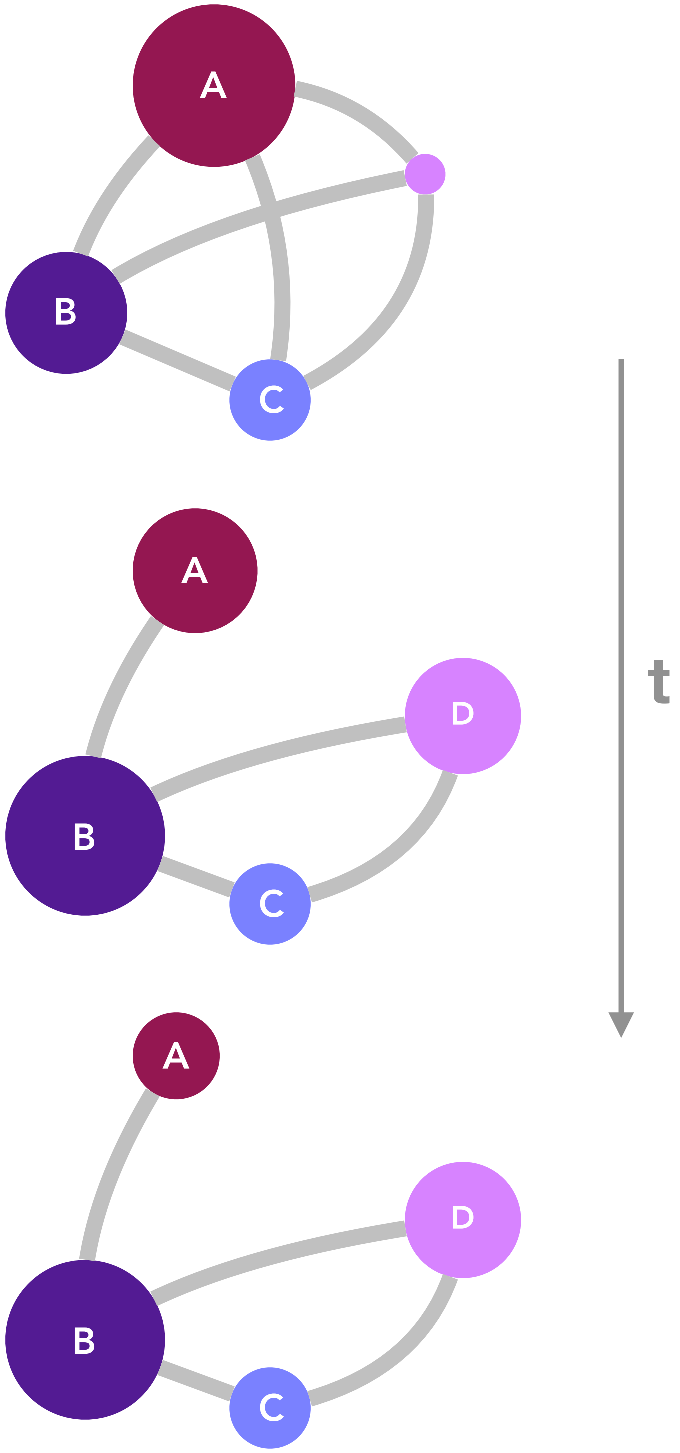
Cluster detection

n	nL	p	**cluster**
A	OTU1	K-P-C-F-G-S	**1**
B	OTU2	K-P-C-F-G-S	**1**
C	OTU3	K-P-C-F-G-S	**1**
D	OTU4	K-P-C-F-G-S	**2**



Hub (keystone) detection

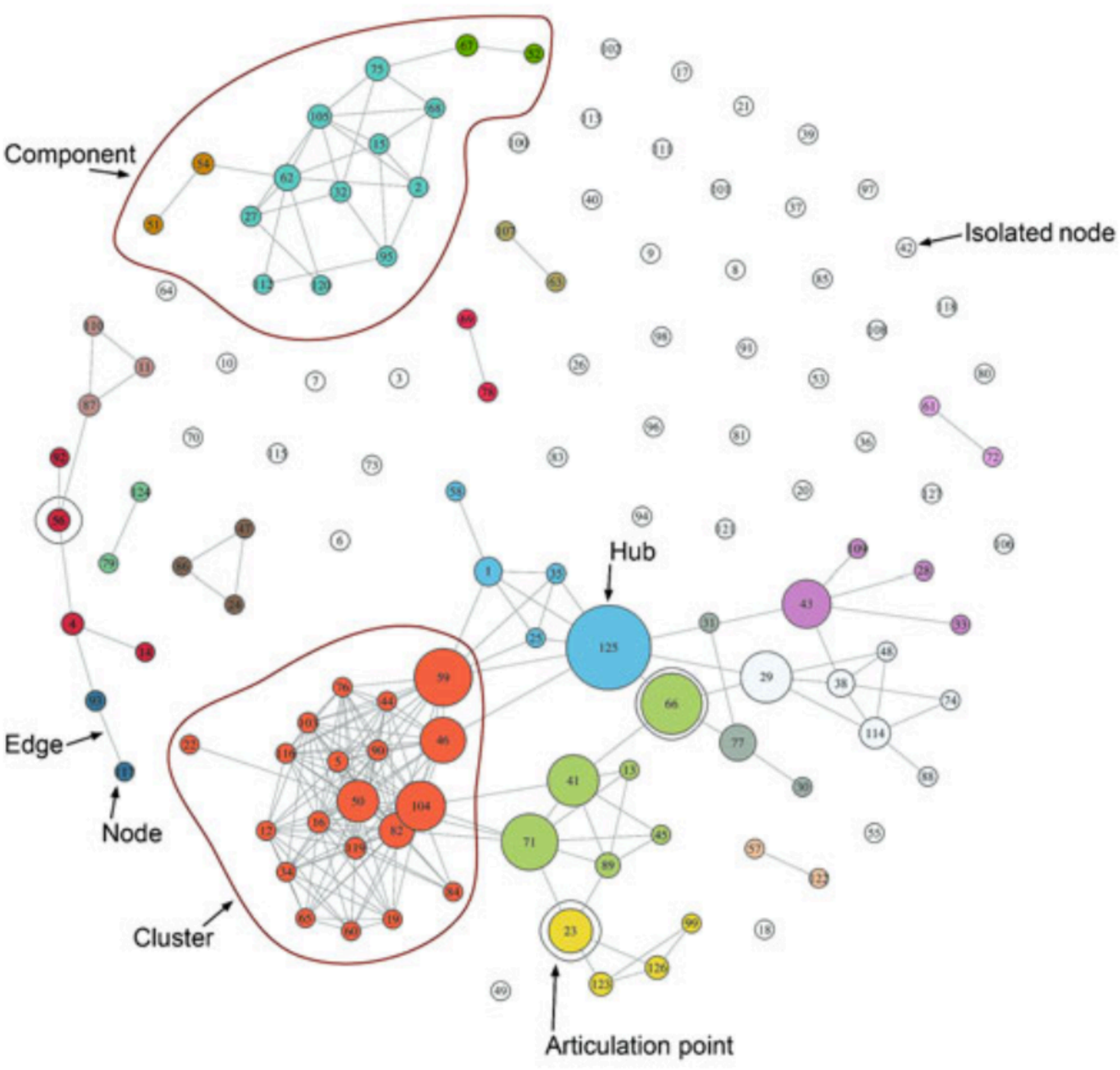
n	nL	p	**degree**
A	OTU1	K-P-C-F-G-S	**2**
B	OTU2	K-P-C-F-G-S	**2**
C	OTU3	K-P-C-F-G-S	**3**
D	OTU4	K-P-C-F-G-S	**1**



Dynamics (time-series)

n	nL	d	p	**t**	**ab**
A	OTU1	3	K-P-C-F-G-S	**1**	**0.4**
B	OTU2	3	K-P-C-F-G-S	**1**	**0.3**
C	OTU3	3	K-P-C-F-G-S	**1**	**0.2**
D	OTU4	3	K-P-C-F-G-S	**1**	**0.1**
A	OTU1	1	K-P-C-F-G-S	**2**	**0.25**
B	OTU2	3	K-P-C-F-G-S	**2**	**0.4**
C	OTU3	2	K-P-C-F-G-S	**2**	**0.1**
D	OTU4	2	K-P-C-F-G-S	**2**	**0.25**
A	OTU1	1	K-P-C-F-G-S	**3**	**0.1**
B	OTU2	3	K-P-C-F-G-S	**3**	**0.4**
C	OTU3	2	K-P-C-F-G-S	**3**	**0.2**
D	OTU4	2	K-P-C-F-G-S	**3**	**0.3**

MDA ▷ Microbiota Network



Edge is a link between a pair of nodes.

Each object in a network is

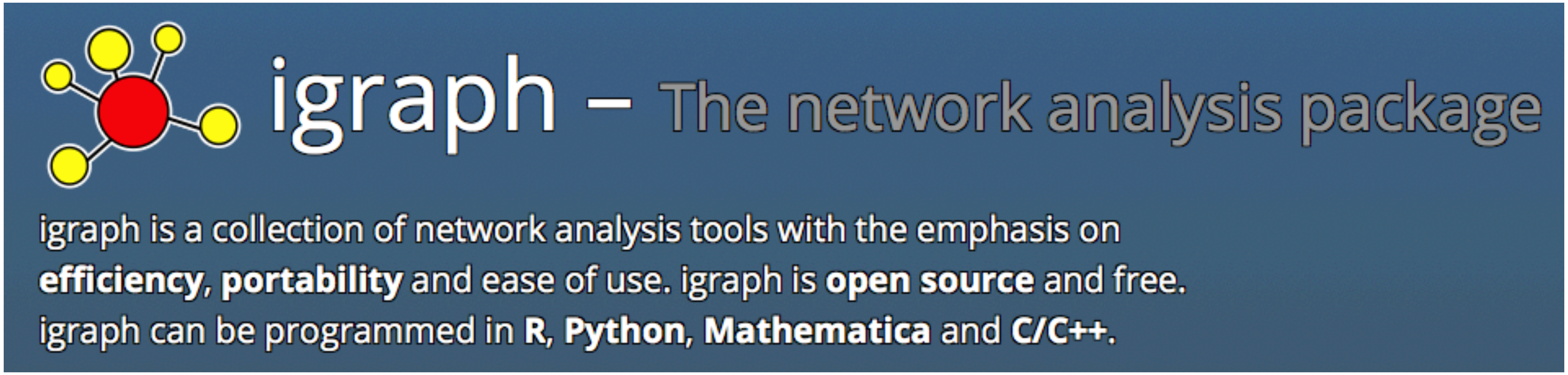
called a **node (vertex)**.

An **hub** is a node (vertex) with large number of links.

A **component** is a subgraph.

An **articulation points** represent vulnerabilities in a connected network.

MDA ▷ Microbiota Network



igraph – The network analysis package

igraph is a collection of network analysis tools with the emphasis on **efficiency**, **portability** and ease of use. igraph is **open source** and free. igraph can be programmed in **R**, **Python**, **Mathematica** and **C/C++**.

https://www.sixhat.net/finding-communities-in-networks-with-r-and-igraph.html

MDA ▷ Microbiota Network

Cytoscape

Network Data Integration, Analysis, and Visualization in a Box

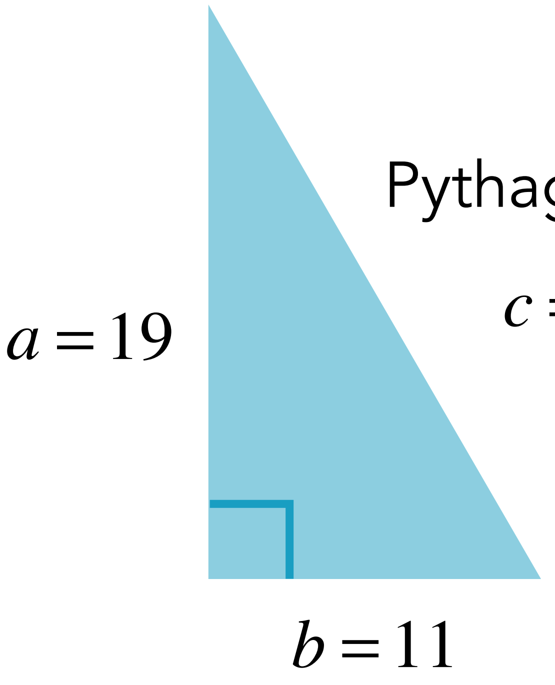
A wide range of methods have been used to construct **ecological networks** based on microbiome data. These approaches vary in their efficiency, accuracy, speed, and computational requirement.

- Dissimilarity-Based Methods
- Correlation-Based Methods
- Regression-Based Methods
- Probabilistic Graphical Models (PGMs)

Dissimilarity-Based Methods

The simplest and fastest way to construct **co-occurrence networks** from OTU microbiome data is to use a **pairwise dissimilarity index** (e.g., Bray–Curtis) in combination with a permutation test.

2x2	Sample1	Sample2
OTU1	20	1
OTU2	12	23



Pythagorean Theorem:

$$c=\sqrt{a^2+b^2}$$

Euclidian distance between sample #1 and sample #2

$$\sqrt{a^2+b^2}=\sqrt{(20-1)^2+(12-23)^2}$$

MDA ▷ Microbiota Network

Jaccard's similarity coefficient (presence-absence)

$$S_{Jaccard} = \frac{a}{a+b+c}$$

a: Number of species in sample A and sample B (joint occurrences)
b: Number of species in sample B but not in sample A
c: Number of species in sample A but not in sample B

```
R> phyloseq::distance(phyloseq, "j")
```

Jaccard's **dis**similarity coefficient (presence-absence)

$$D_{Jaccard} = \frac{b+c}{a+b+c}$$

a: Number of species in sample A and sample B (joint occurrences)
b: Number of species in sample B but not in sample A
c: Number of species in sample A but not in sample B

```
R> phyloseq::distance(phyloseq, "cc")
```

Dissimilarity-Based Methods

The simplest and fastest way to construct **co-occurrence networks** from OTU microbiome data is to use a **pairwise dissimilarity index** (e.g., Bray–Curtis) in combination with a permutation test.

Correlation-Based Methods

A popular alternative to dissimilarity-based network inference is correlation-based techniques. These methods detect significant pairwise interactions between OTUs using a correlationcoefficient such as **Pearson's product-moment correlation coefficient** or **Spearman's nonpara-metric rank correlation coefficient**. Correlation-based methods suffers from limitations such as detecting **spurious correlations** among low-abundance OTUs in **zero-inflated data** or being **sensitive to compositionality**. For more details see Weiss et al. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 10 :1669-1681

Layeghifard et al. (2017) Disentangling Interactions in the Microbiome: A Network Perspectiv. Trends in Microbiology.

MDA ▷ Microbiota Network

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
OUT1	2519	2354	1074	2	452	1233
OUT2	2520	1158	1287	1	34	3184
OUT3	106	4	2	**0**	**0**	**0**
OUT4	490	82	148	**0**	**0**	22
OUT5	120	73	111	**0**	**0**	133
OUT6	13	1	6	**0**	**0**	**0**
OUT7	9	**0**	**0**	**0**	**0**	**0**
OUT8	813	415	142	**0**	**0**	808
OUT9	45	2	7	**0**	**0**	**0**

```
--------------------------------------------
OTU-Table Summary
--------------------------------------------
25414720  Reads (25.4M)
     436  Samples
    8036  OTUs
--------------------------------------------
 3503696  Counts
 3433317  Count   =0  (98.0%)
   12008  Count   =1  (0.3%)
   36286  Count >=10 (1.0%)
--------------------------------------------
       0  OTUs found in all samples (0.0%)
--------------------------------------------
```

MDA ▷ Microbiota Network

```
x2 <- sample(1:100, 10, replace = TRUE)
y2 <- sample(1:100, 10, replace = TRUE)
cor(x2,y2)
# 0.466
```

```
x1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
y1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
cor(x1,y1)
# 0.883
```

```
x3 <- c(x2, rep(0,20))
y3 <- c(y2, rep(0,20))
cor(x3,y3)
# 0.790
```

Regression-Based Methods

Network inference methods based on pairwise association metrics such as Bray–Curtis and Pearson coefficient are not able to capture more complex forms of polymicrobial interactions. One obvious alternative is to use **multiple regression analysis** to infer the abundance of one species from the combined abundances of other taxa. Regression-based methods suffer from **overfitting** that increases with the number of predictor variables. Overfitting can be remedied by using **sparse regressionand cross-validation.**

Probabilistic Graphical Models (PGMs)

PGMs deal with uncertainty and complexitythrough the use of **probability theory** and **graph theory**, respectiv. **Bayesian networks** and **Markov networks** are the most popular graphical models used.

Layeghifard et al. (2017) Disentangling Interactions in the Microbiome: A Network Perspectiv. Trends in Microbiology.

Network Construction Methods Robust to Compositionality

The concerns over correlation-based analyses have led to the development of methods that are robust to compositionality.

SparCC (Sparse Correlations for Compositional data), for example, is a technique that uses linear Pearson's correlations between the log-transformed components to infer associations in compositional data (Friedman 2012). ➜ R package: install.packages("SpiecEasi")

SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) combines data transformations developed for compositional data analysis with a graphical model inference framework with the assumption that the underlying ecological association network is sparse (Kurtz *et al* 2015). ➜ R package: install.packages("SpiecEasi")

REBACCA (Regularized Estimation of the BAsis Covariance based on Compositional dAta) estimates the correlations between pairs of basis abundance using the log ratiotransformation of count or proportional data (Ban *et al.* 2015). ➜ R script

Friedman (2012) Inferring correla- tion networks from genomic survey data. PLoS Comput Biol 8:e1002687.

Kurtz et al. (2015) Sparse and compositionally robust inference of micro- bial ecological networks. PLoS Comput Biol 11:e1004226.

Ban et al. (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics 31: 3322-3329.

Network Construction Methods Robust to Compositionality

ReBoot, a permutation-renormalization bootstrap method to evaluate the significance of e.g. Pearson's correlation coefficients es; plugin tool called CoNet (Faust *et al.* 2012). ➜ Cytoscape plugin http://www.raeslab.org/software/conet.html

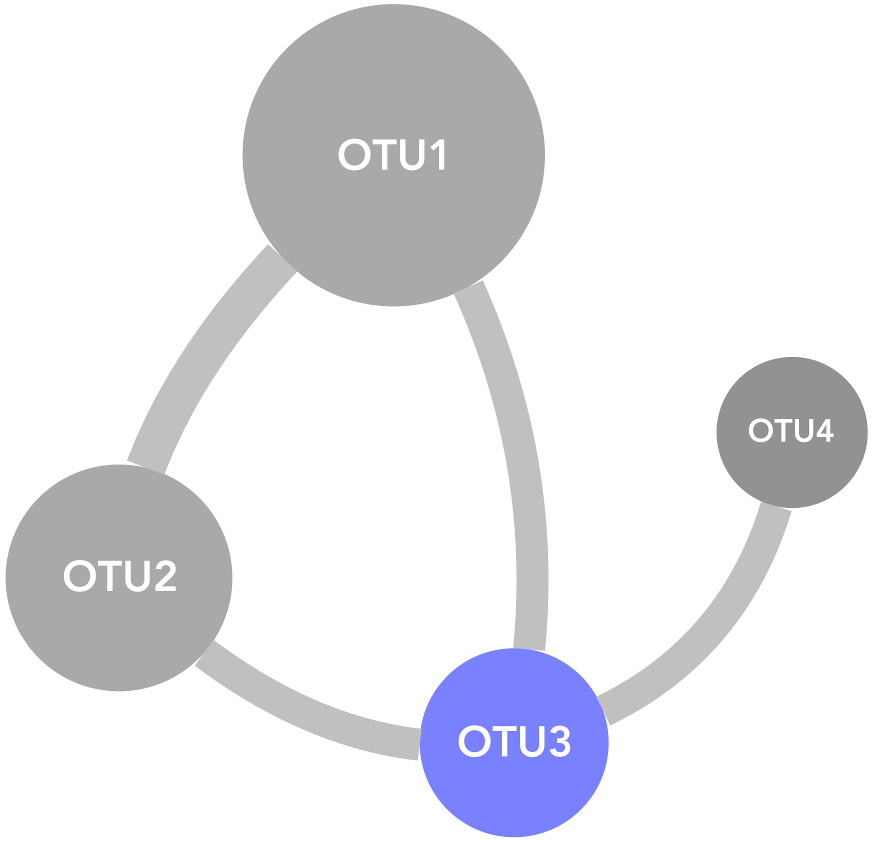
CCLasso (Correlation inference for Compositional data through Lasso) uses least squares with L1 penalty after log ratio transformationfor raw compositional data to infer the correlations among microbes through a latent variable model (Fang *et al.* 2015). ➜ R script

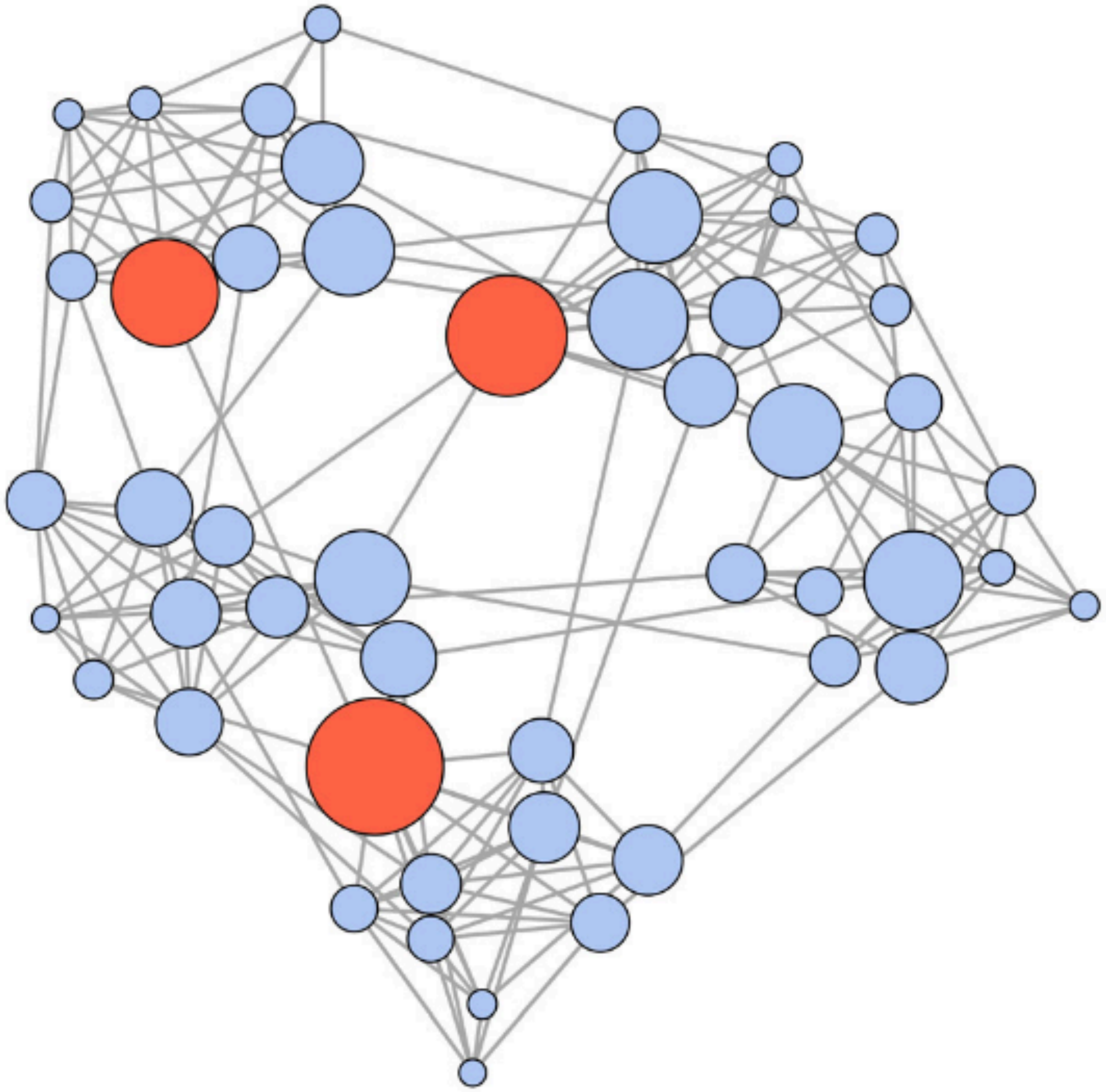
MENAP (Molecular Ecological Network Analysis Pipeline) is a Random Matrix Theory (RMT)-based method that is developed to address the issue of arbitrary choice of threshold used to include or exclude interaction from ecologicalnetworks (Deng *et al.* 2012). ➜ http://129.15.40.240/mena/

MInt (Microbial Interaction) is a Poisson-multivariate normal hierarchical model to find taxon-taxon interactions from metagenomic count data by controlling for confounding predictors at the Poisson layer, and capturing direct microbial interactions at the multivariate normal layer, using an‘1 penalized precision matrix (Biswas *et al.* 2015). Note: MInt was shown to outperform SparCC and graphical lasso methods in both synthetic and experimental experiments. ➜ R package: install.packages("MInt")

Faust et al. (2012) Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol 8:e1002606.
Fang et al. (2015) CCLasso: correlation inference for compositional data through Lasso. Bioinformatics 31: 3172-3180.
Deng et al. (2012) Molecular ecological network analyses. BMC Bioinformatics 13:113.
Biswas et al. (2015) Learning microbial interaction networks from metagenomic count data. in Research in Computational Molecular Biology, Springer, pp. 32-43.

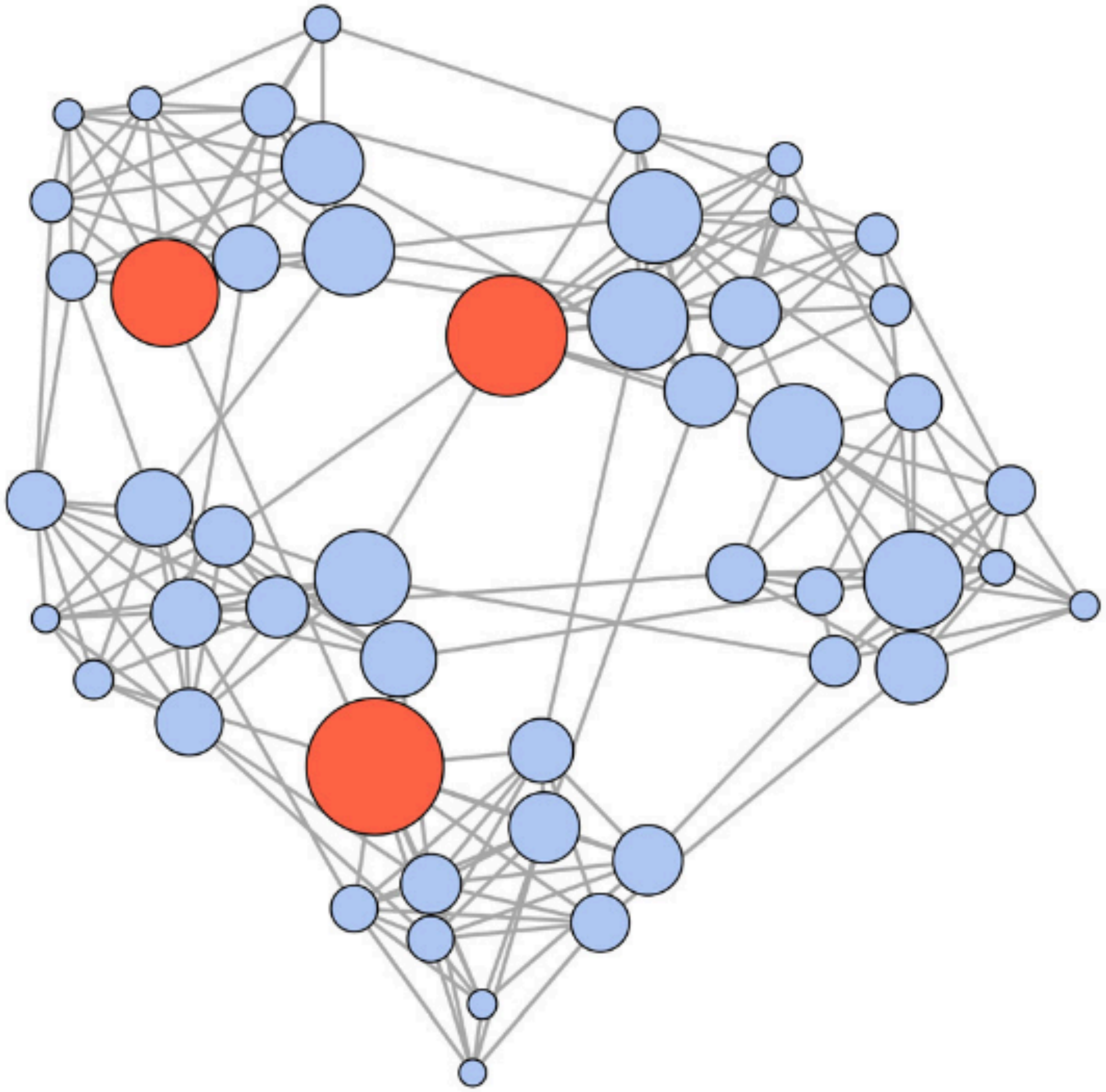
Hub (Keystone) Detection





Keystones

The importance of nodes can take a variety of meanings **depending on the context and application**. For example, the most important node may be the most significant member in the microbial community, the most essential microbe for community stability, the etiological agent of disease, or the organism responsible for disease transmission.

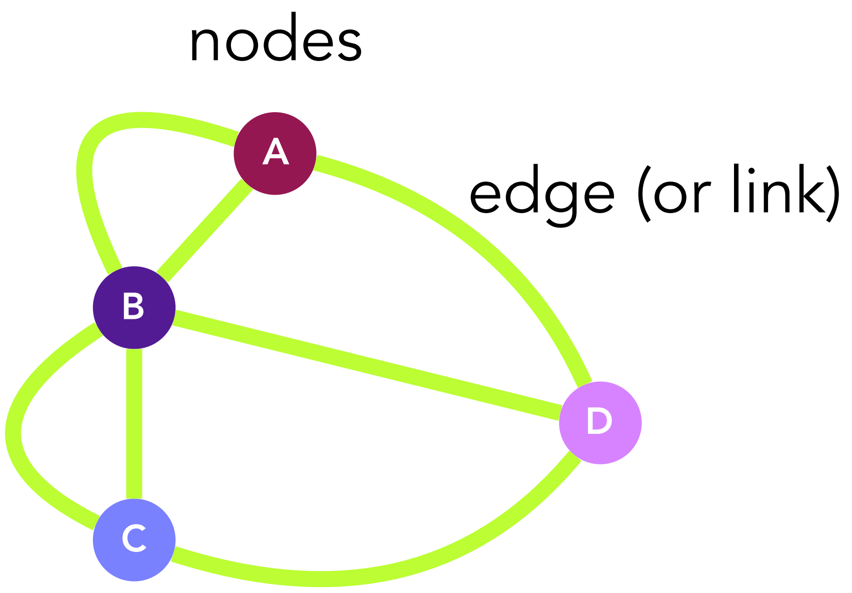


Prominence measures :

- **Degree**
- **Closeness**
- **Betweenness**
- **Eigenvector**
- **Bonacich power**
- Flow betweenness
- Load
- Information
- Stress
- Harary graph
- **Bonacich alpha**
- **Kleinberg authority**
- **Kleinberg hub**
- **PageRank**

***bold: available in igraph**

Microbiota Network



Information about the relative importance of nodes and edges in a graph can be obtained through **centrality** indices.

A node with higher **centrality** would have more control over the network, because **more information will pass through that node**.

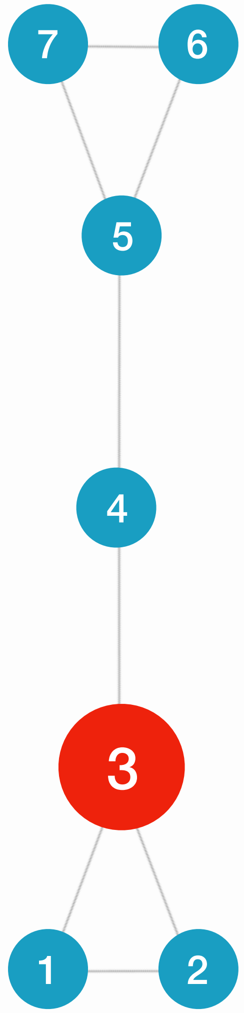
- Characterization by network flows (e.g. paths, walks)
- Characterization by walk structure (e.g. **betweenness centrality**)

The vertex (node) **betweeness centrality** of a vertex v is defined as:

$$BC(v) = \sum_{u,v \in V} \left(\frac{s_{uw}(v)}{s_{uw}}\right)$$

s_{uw}: total number of shortest paths between node u and w

$s_{uw}(v)$: total number of shortest paths between node u and w that pass though v

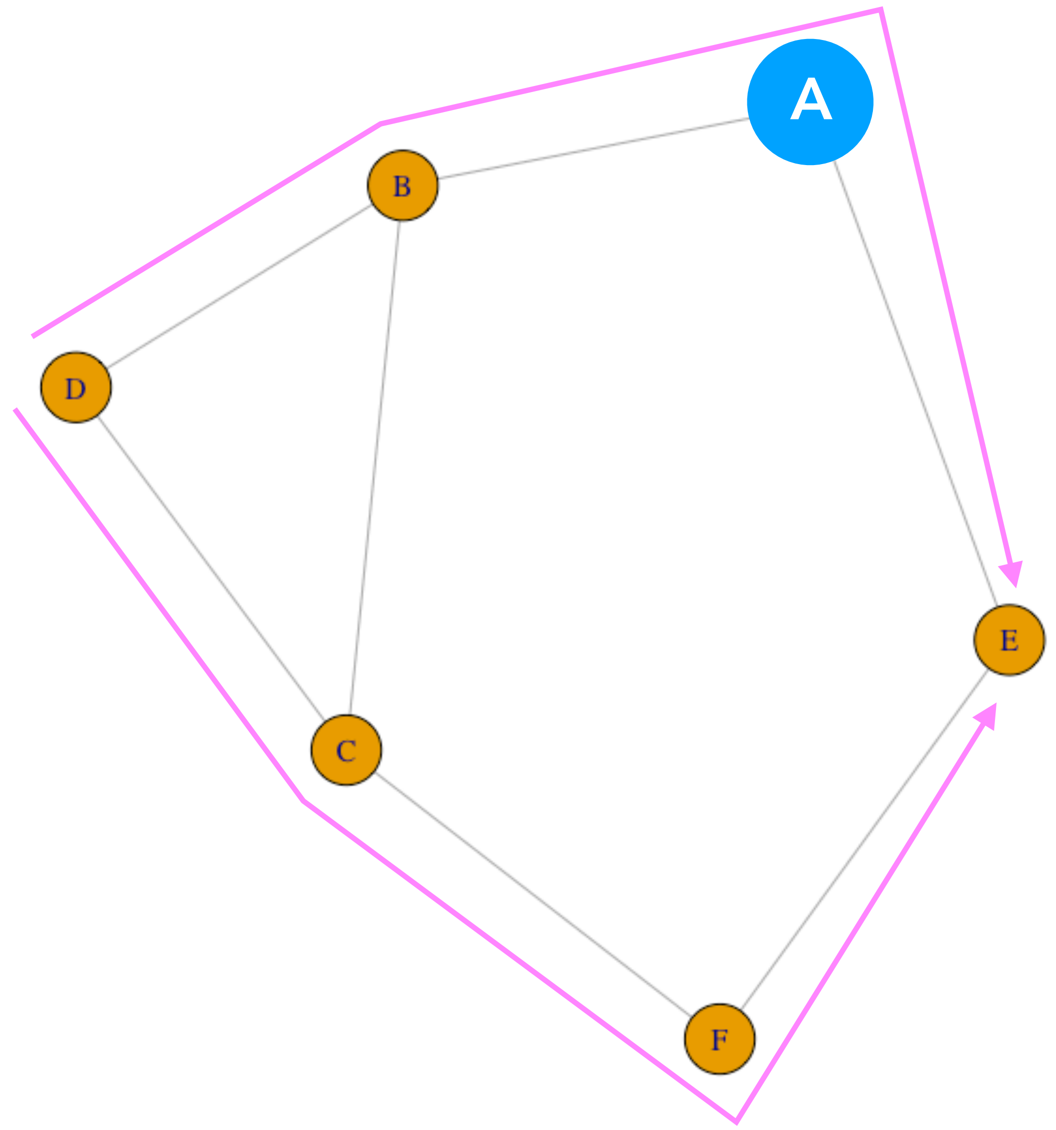


	s_{uw}	$s_{uw}(v)$	$\frac{s_{uw}(v)}{s_{uw}}$
(1,2)	1	0	0
(1,4)	1	1	1
(1,5)	1	1	1
(1,6)	1	1	1
(1,7)	1	1	1
(2,4)	1	1	1
(2,5)	1	1	1
(2,6)	1	1	1
(2,7)	1	1	1
(4,5)	1	0	0
(4,6)	1	0	0
(4,7)	1	0	0
(5,6)	1	0	0
(5,7)	1	0	0
(6,7)	1	0	0
			8

v

	1	2	3	4	5	6	7
degree	2	2	**3**	2	**3**	2	2
BC(v)	0	0	8	**9**	8	0	0

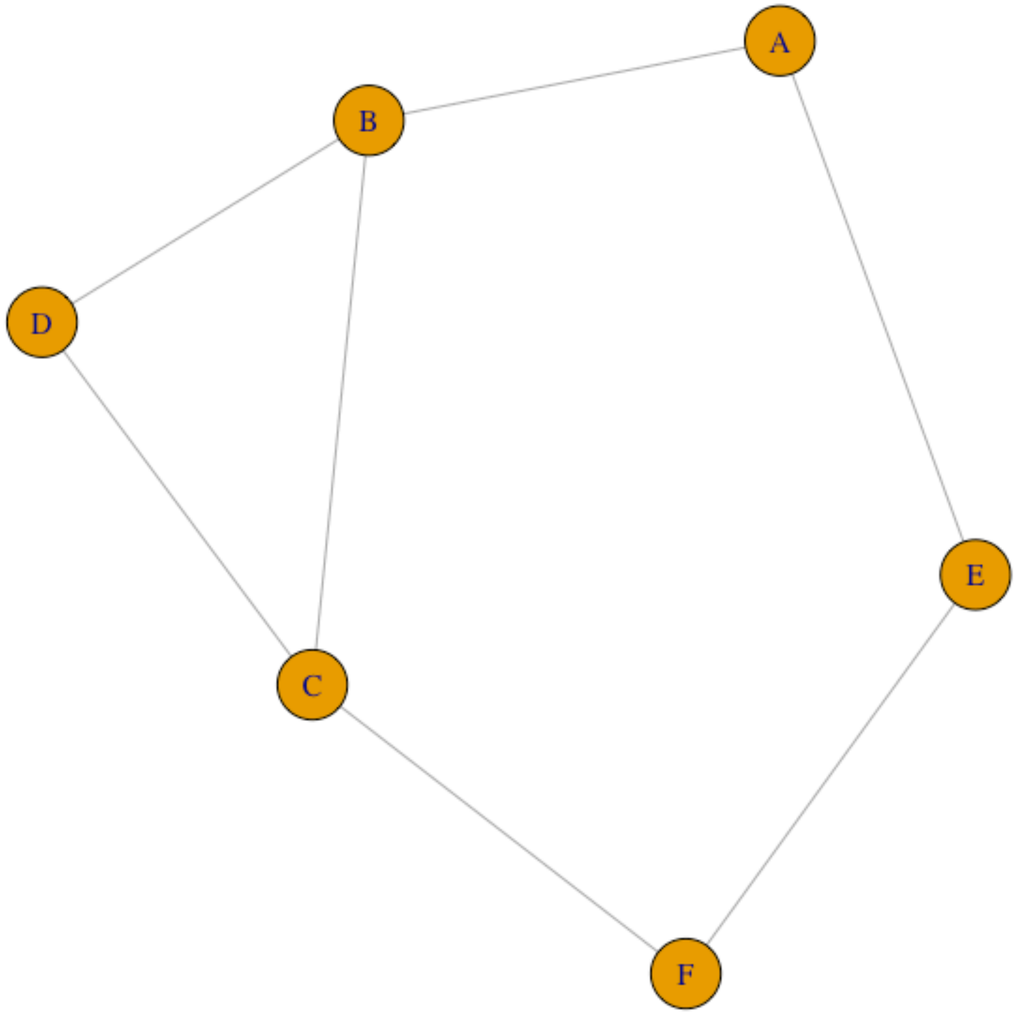
MDA ▷ Microbiota Network



	s_{uw}	$s_{uw}(v)$	$\frac{s_{uw}(v)}{s_{uw}}$
(B,C)	1	0	0
(B,D)	1	0	0
(B,E)	1	1	1
(B,F)	1	0	0
(C,D)	1	0	0
(C,E)	1	0	0
(C,F)	1	0	0
(D,E)	**2**	**1**	**0.5**
(D,F)	1	0	0
(E,F)	1	0	0
			1.5

s_{uw}: total number of shortest paths between node u and w

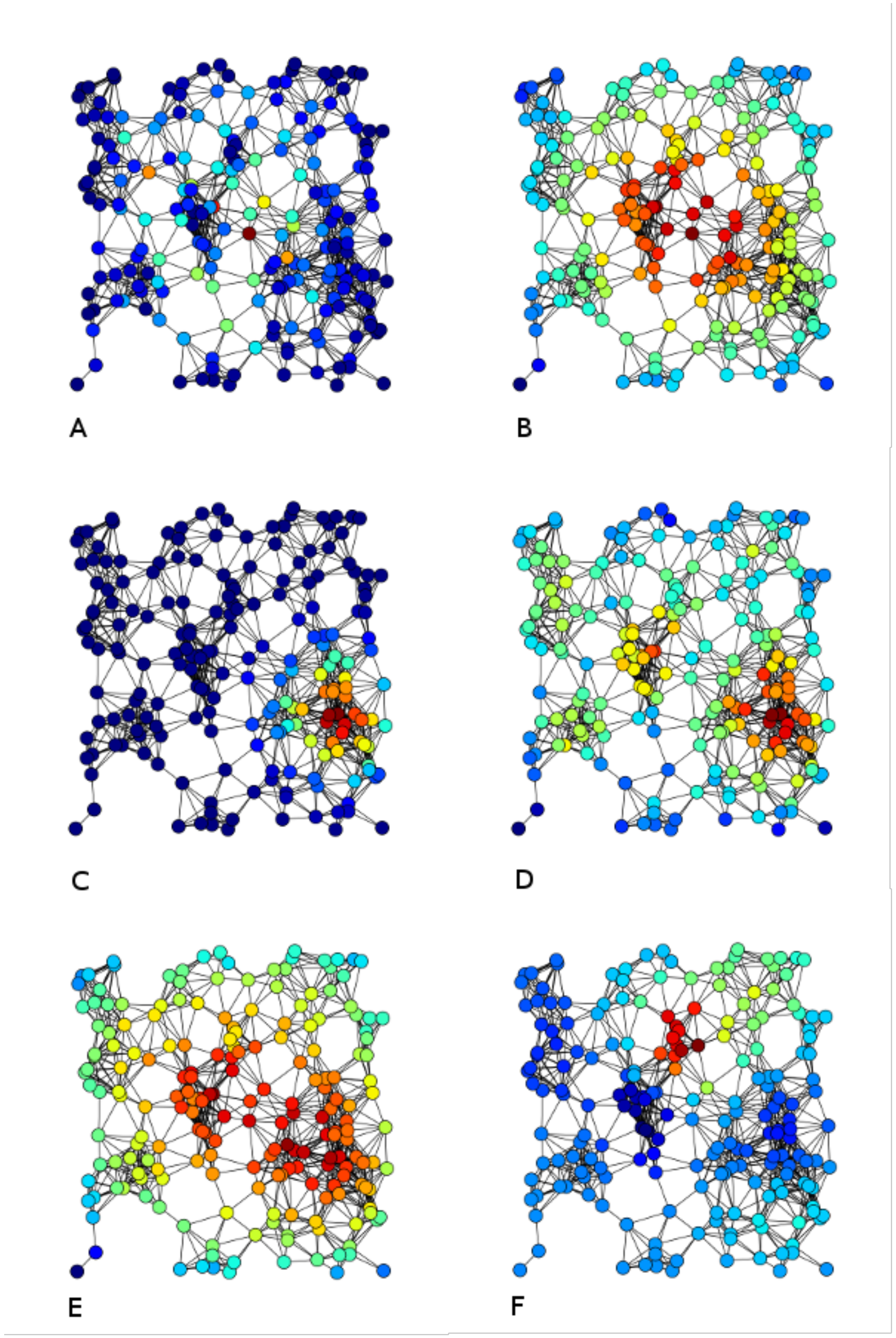
$s_{uw}(v)$: total number of shortest paths between node u and w that pass though v



	A	B	C	D	E	F
degree	2	3	2	3	2	2
BC(v)	1.5	2.5	1.0	2.5	0.0	1.5

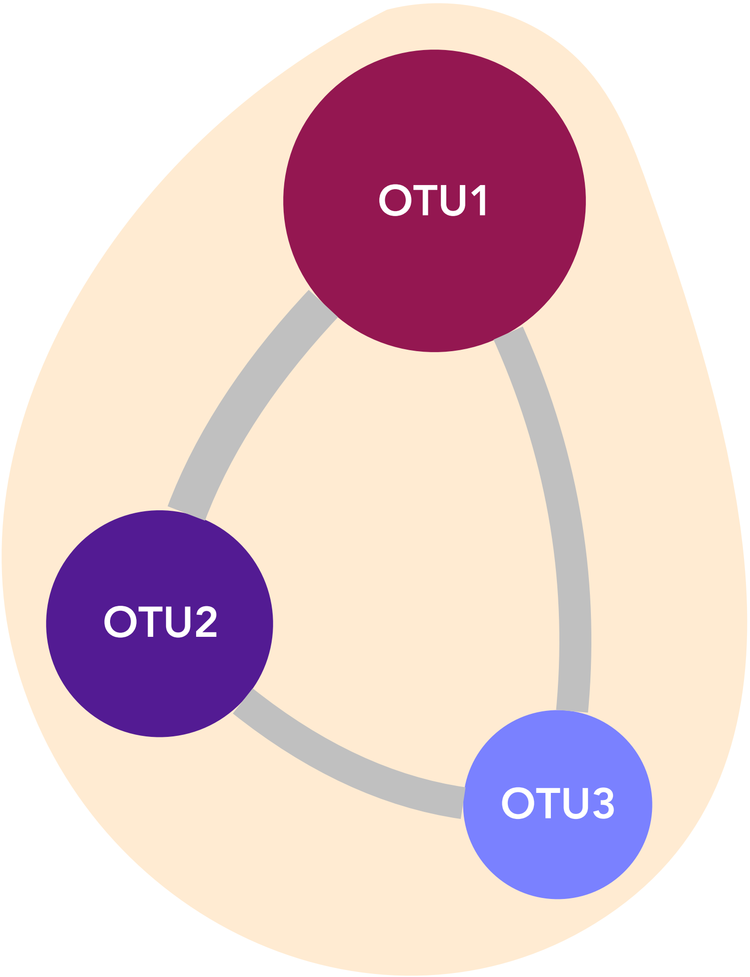
BC: betweeness centrality

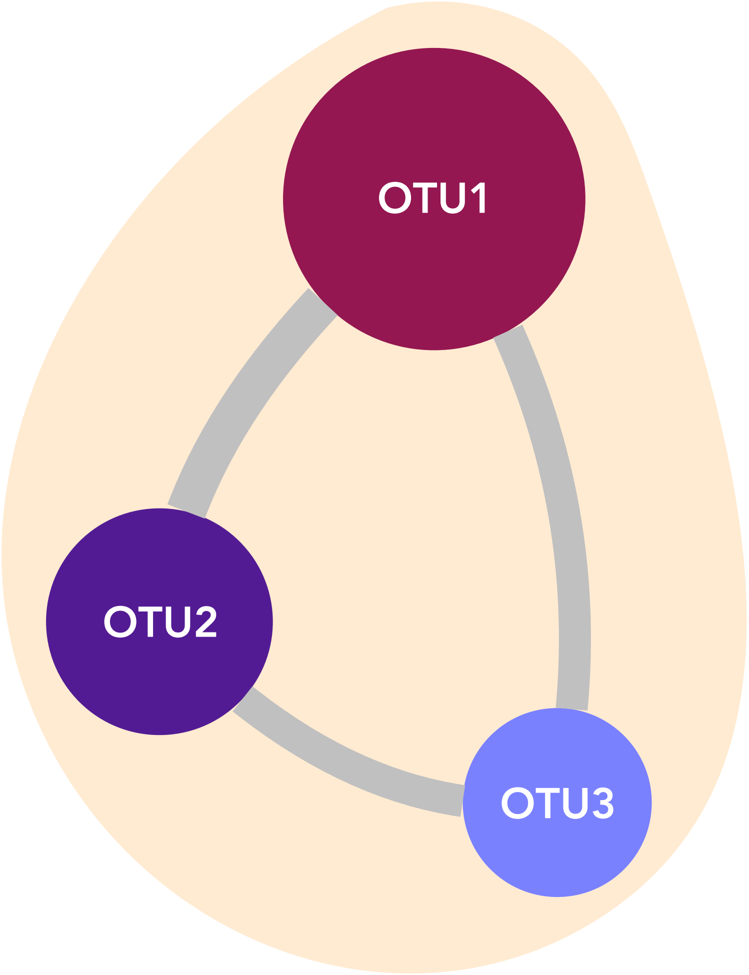
MDA ▷ Microbiota Network



A) **Betweenness centrality**
B) **Closeness centrality**
C) **Eigenvector centrality**
D) Degree centrality
E) Harmonic Centrality
F) Katz centrality

Community (Cluster) Detection

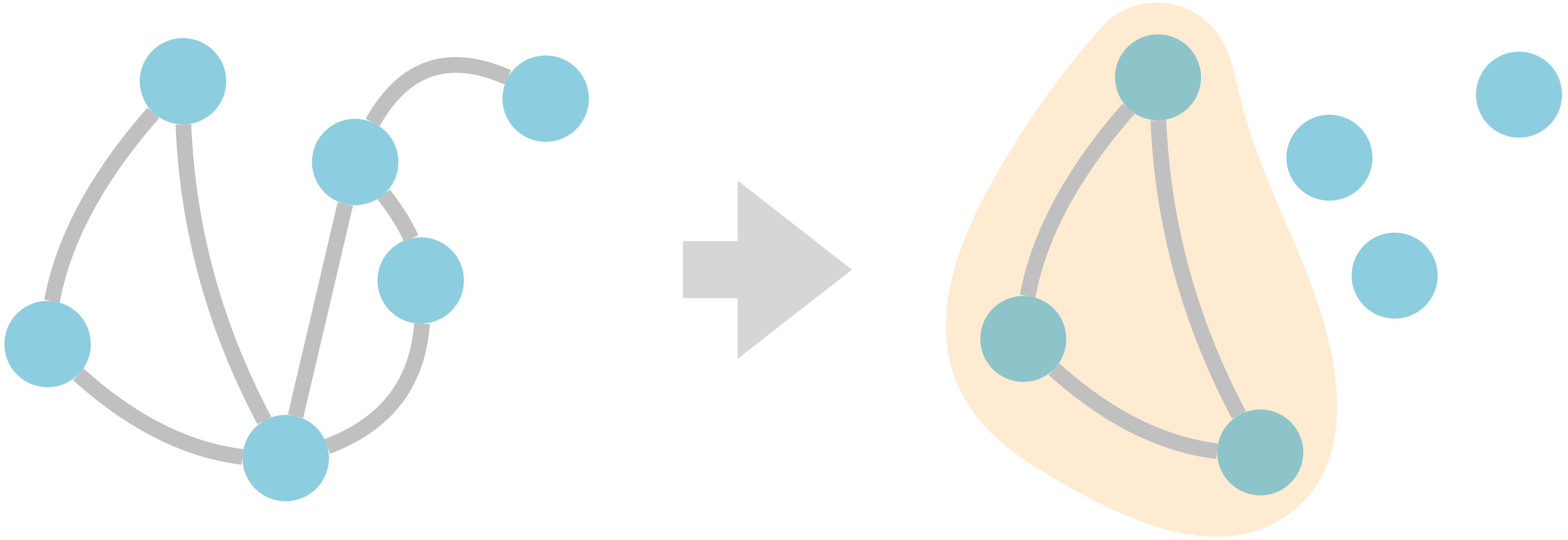




Community detection functions in igraph:

- Edge-betweenness
- Leading eigenvector
- Fast-greedy
- Louvain
- Walktrap
- Label propagation
- InfoMAP
- Spinglass
- Optimal

Community detection: A variety of subgroup identification algorithms and heuristics that define groups not just based on the internal ties have been developed. These approaches vary in their details, but they are all designed to **identify internally cohesive subgroups** that are somewhat separated or isolated from other groups or nodes. These approaches are sometimes called **community detection algorithms**.



Edge betweennessis a popular method in which links are removed in the decreasing order of their betweenness scores (Girvan and Newman 2002).

Girvan and Newman (2002) Community structure in social and biological networks. PNAS 99: 7821-7826.

The concept of assortativity was introduced by Newman in 2002 and is extensively studied since then. **Assortativity is a graph metric**. It represents to what extent nodes in a network associate with other nodes in the network, being of similar sort or being of opposing sort. Generally, the assortativity of a network is determined for the degree (number of direct neighbours) of the nodes in the network.

Newman (2002) Assortative mixing in networks. Phys. Rev. Lett., Vol. 89(20).

network analysis - network structure analysis - clustering

- WGCNA (Weighted Correlation Network Analysis)
- MCL (Markov Cluster Algorithm)
- Walktrap

Clusters can reveal important nodes and correlations to environmental factors:

“We applied WGCNA to the relative abundance tables of eukaryotic, prokaryotic and viral lineages and identified unique subnetworks significantly associated with carbon export within each data set."

Guidi et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. Nature, 532(7600), 465.

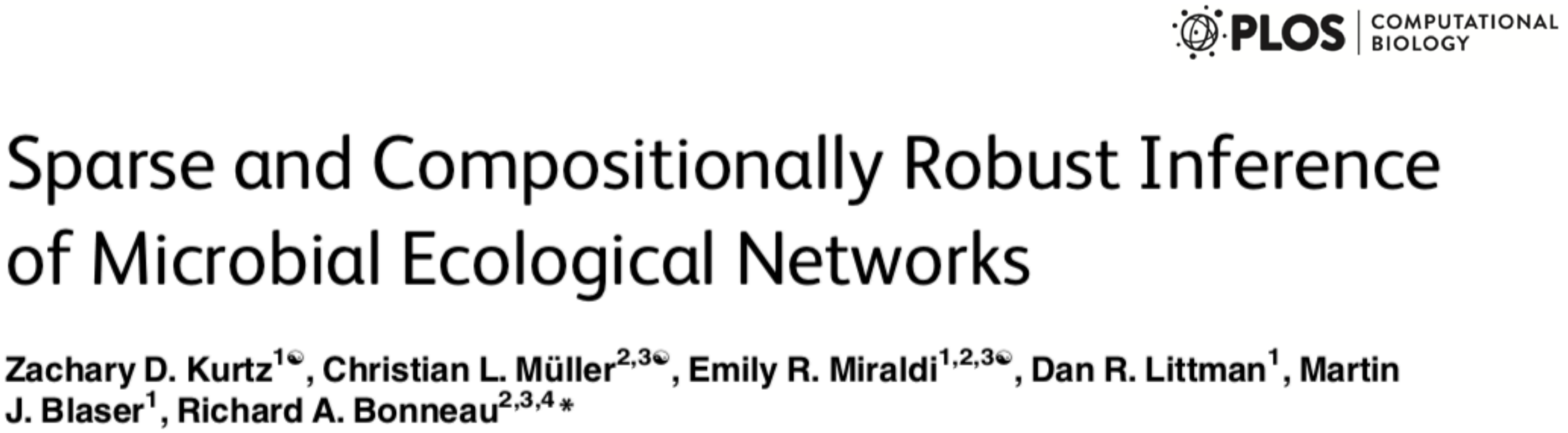
https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/index.html

Robustness of microbial networks - Permutation testing

a) Permute original data
b) Test robustness of edges
c) Remove unstable edges

Guidi et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. Nature, 532(7600), 465.

MDA ▷ Microbiota Network

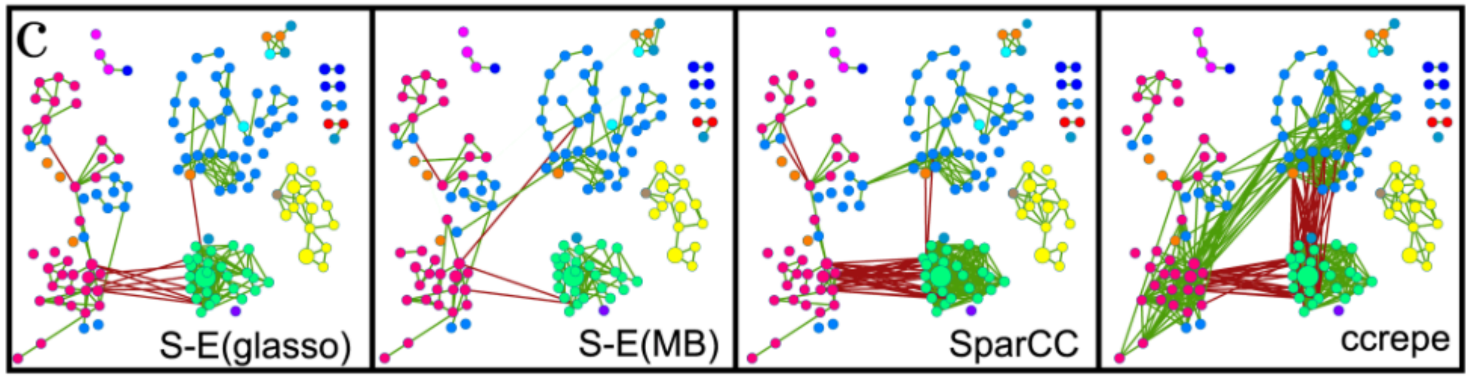


PLOS | COMPUTATIONAL BIOLOGY

Sparse and Compositionally Robust Inference of Microbial Ecological Networks

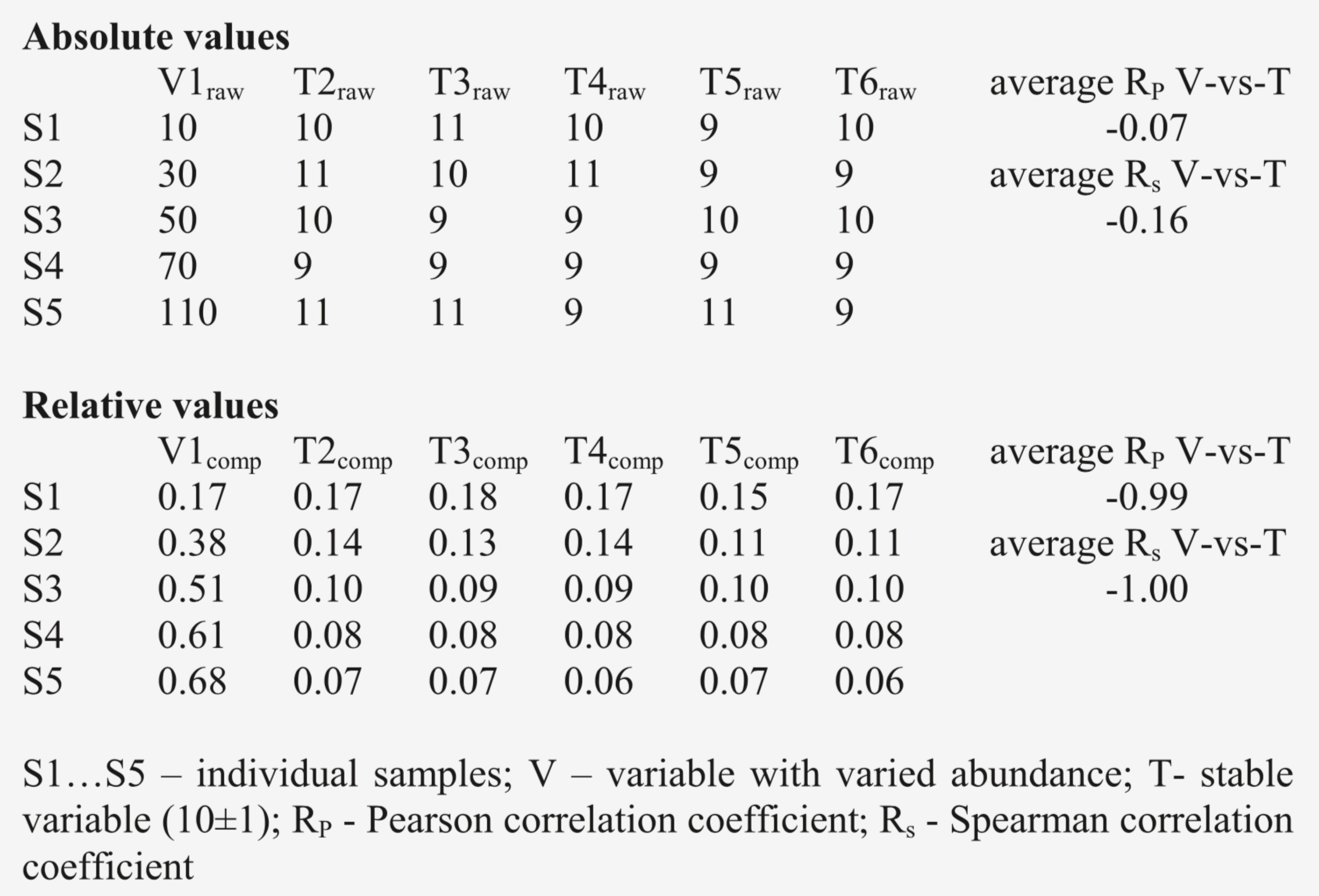
Zachary D. Kurtz[1☯], Christian L. Müller[2,3☯], Emily R. Miraldi[1,2,3☯], Dan R. Littman[1], Martin J. Blaser[1], Richard A. Bonneau[2,3,4*]

SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference)



https://www.rdocumentation.org/packages/SpiecEasi/versions/0.1.4

MDA ▷ Microbiota Network



Absolute values

	$V1_{raw}$	$T2_{raw}$	$T3_{raw}$	$T4_{raw}$	$T5_{raw}$	$T6_{raw}$	average R_P V-vs-T
S1	10	10	11	10	9	10	-0.07
S2	30	11	10	11	9	9	average R_s V-vs-T
S3	50	10	9	9	10	10	-0.16
S4	70	9	9	9	9	9	
S5	110	11	11	9	11	9	

Relative values

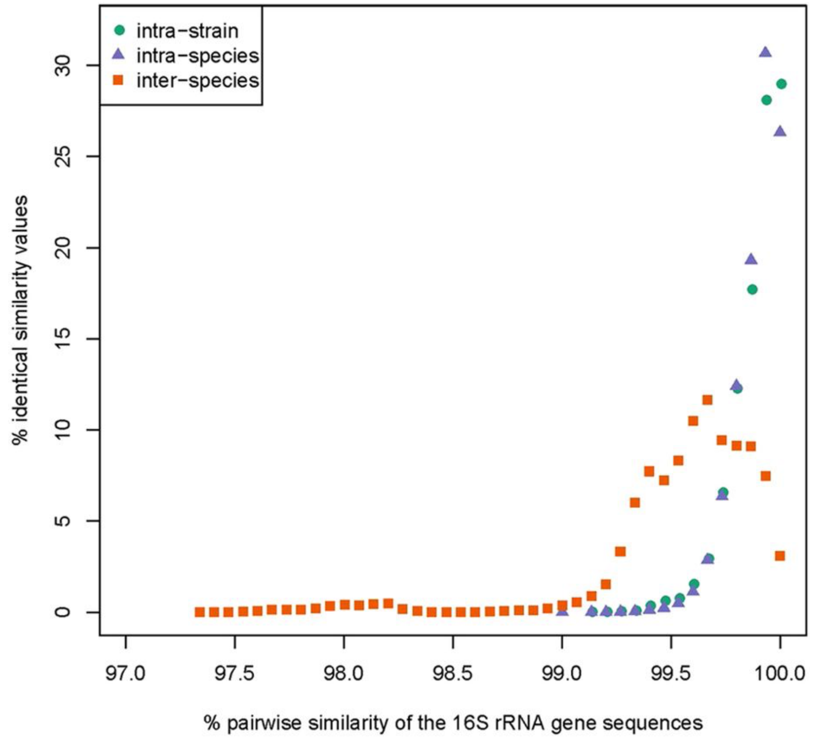
	$V1_{comp}$	$T2_{comp}$	$T3_{comp}$	$T4_{comp}$	$T5_{comp}$	$T6_{comp}$	average R_P V-vs-T
S1	0.17	0.17	0.18	0.17	0.15	0.17	-0.99
S2	0.38	0.14	0.13	0.14	0.11	0.11	average R_s V-vs-T
S3	0.51	0.10	0.09	0.09	0.10	0.10	-1.00
S4	0.61	0.08	0.08	0.08	0.08	0.08	
S5	0.68	0.07	0.07	0.06	0.07	0.06	

S1…S5 – individual samples; V – variable with varied abundance; T- stable variable (10±1); R_P - Pearson correlation coefficient; R_s - Spearman correlation coefficient

Paliy and Shankar (2016) Application of multivariate statistical techniques in microbial ecology. Molecular Ecology 25, 1032–1057

When defining nodes, limitations in marker gene data (e.g. 16S rRNA) can be problematic. Various aspects of these data are also relevant to network construction: **resolution**, varying **sequencing depth**, **compsitionality**, and **sparsity**.

Resolution - Qualitative issue of species assignment caused by the variation- and length-dependend resolution strength of the amplicon, the limitation or errors of the reference database, the shortcommings of the annoation or "clustering" method.

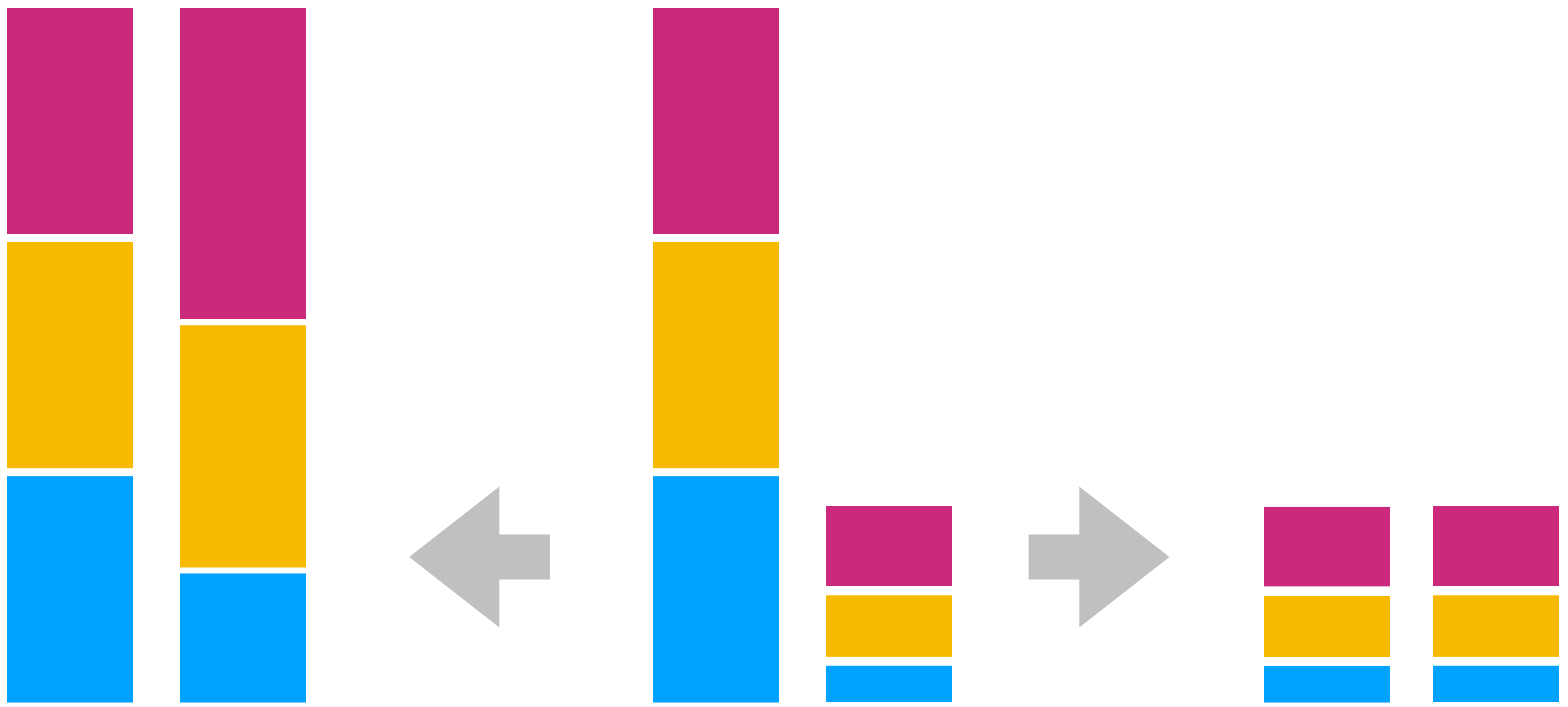


The 16S rRNA gene sequence was not sufficient to differentiate the bacteria within Bacillus cereus group due to its high conservation.

A total of 1007 16S rRNA gene sequences were analyzed, sharing all sites from positions 352 to 1051 in the complete 16S rRNA gene; more sequences could not be considered due to the condition of some of the draft genome sequences. The x-axis indicated the pairwise similarity (in %) of the 16S rRNA gene sequences, whereas the y-axis represents the proportion of each respective similarity value.

Liu *et al.* (2015) Genomic insights into the taxonomic status of the Bacillus cereus group. Sci Rep 5:14082.

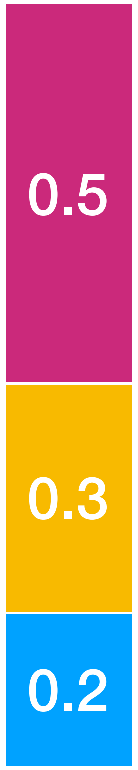
Sequencing Depth & Compsitionality - Technical variation during sequencing results in varying sequencing depths. To reduce/remove sequencing depth variation, counts should be converted either into relative abundances, rarefied or normalized. As a result, we are dealing with compositional rather than absolute data. Certain statistical methods, such as correlations, can lead to erroneous results when applied to compositional data.



MDA ▷ Microbiota Network

Relative abundances could be converted into absolute abundances when the total number of microorganisms in a sample is known by spiking with DNA or qPCR.

qPCR: 5E+06 gene copy numbers



⟹ 2.5E+06

⟹ 1.5E+06

⟹ 1.0E+06

MDA ▷ Microbiota Network

Sparsity - Microbiome data are zero-rich. While log-ratios (network inference in general) can be used to tackle compositionality it is sensitive to zeros (i.e. negative infinities). Pseudo-counts could resolve the issue but might impact the results as they alter the covariance structure of data. Alternative treatments of zeros have been proposed but are problematic since zeros could indicate absence or undersampling.

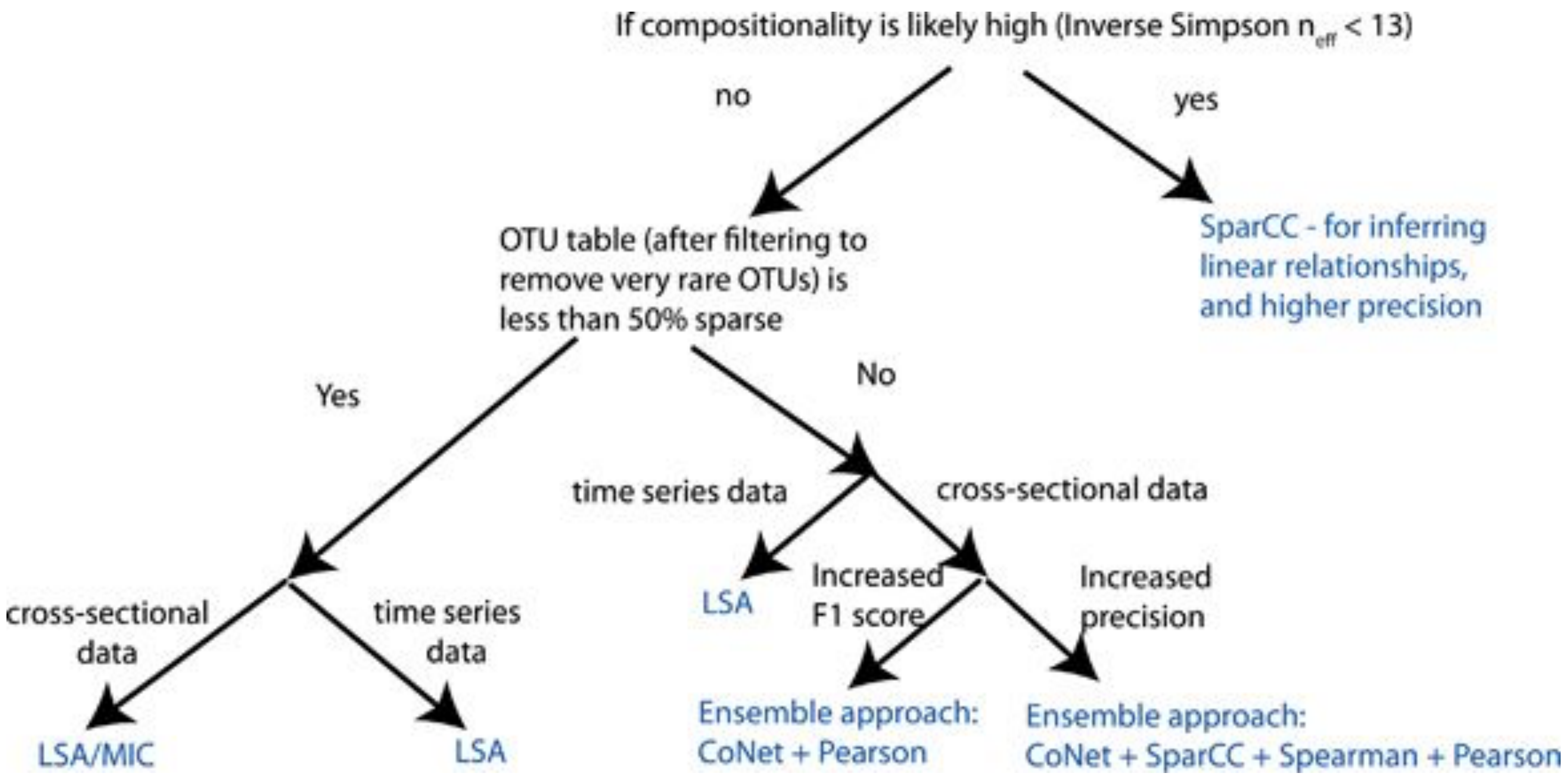
```
set.seed(190801)
x2 <- sample(1:100, 10, replace = TRUE)
y2 <- sample(1:100, 10, replace = TRUE)
cor(x2,y2)
# 0.466

x1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
y1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
cor(x1,y1)
# 0.883

x3 <- c(x2, rep(0,20))
y3 <- c(y2, rep(0,20))
cor(x3,y3)
# 0.790
```

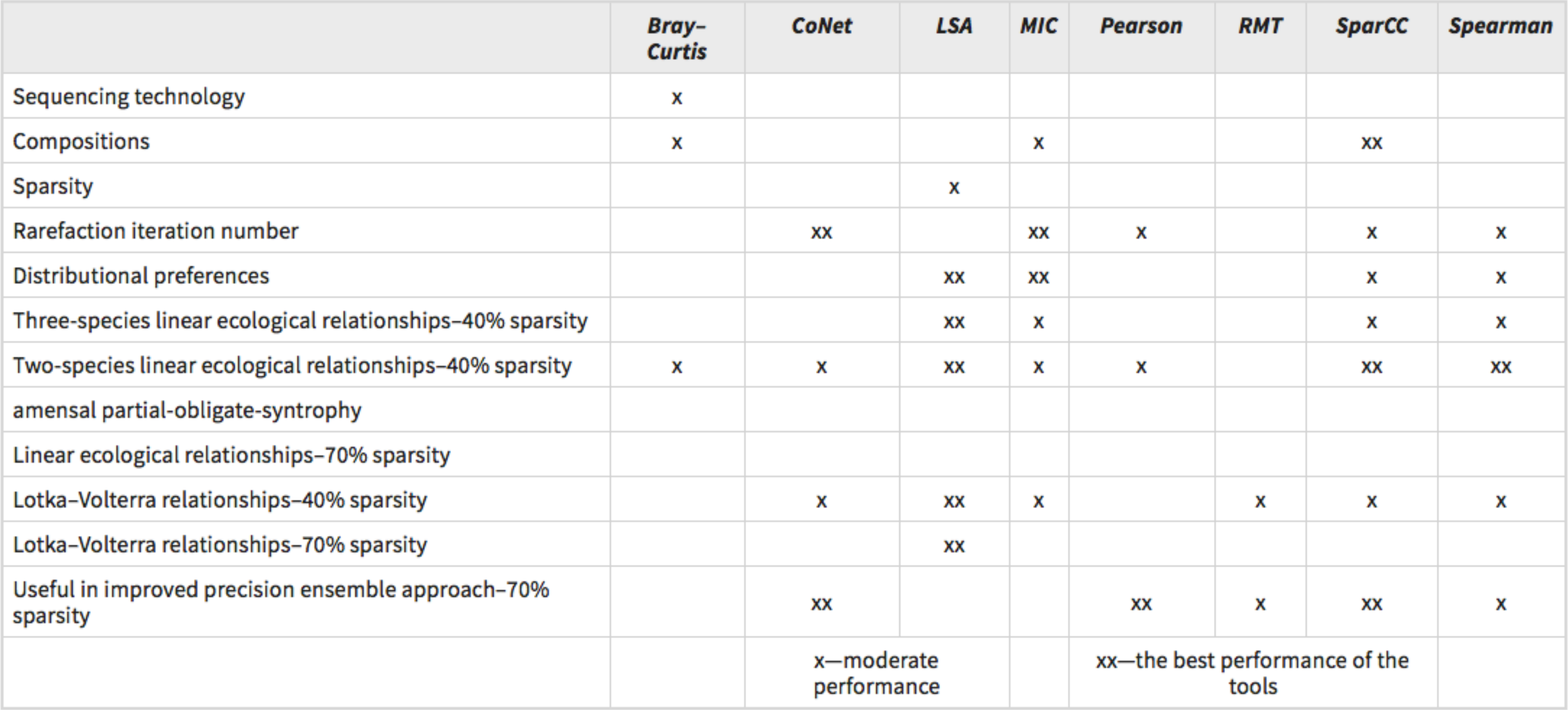
MDA ▷ Microbiota Network

The best correlation technique depending upon data set characteristics and desired ecological relationship discovery.


If compositionality is likely high (Inverse Simpson n_{eff} < 13)
no
yes
SparCC - for inferring linear relationships, and higher precision
OTU table (after filtering to remove very rare OTUs) is less than 50% sparse
Yes
No
time series data
cross-sectional data
LSA
Increased F1 score
Increased precision
Ensemble approach: CoNet + Pearson
Ensemble approach: CoNet + SparCC + Spearman + Pearson
cross-sectional data
time series data
LSA/MIC
LSA

Weiss *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 10:1669–81.

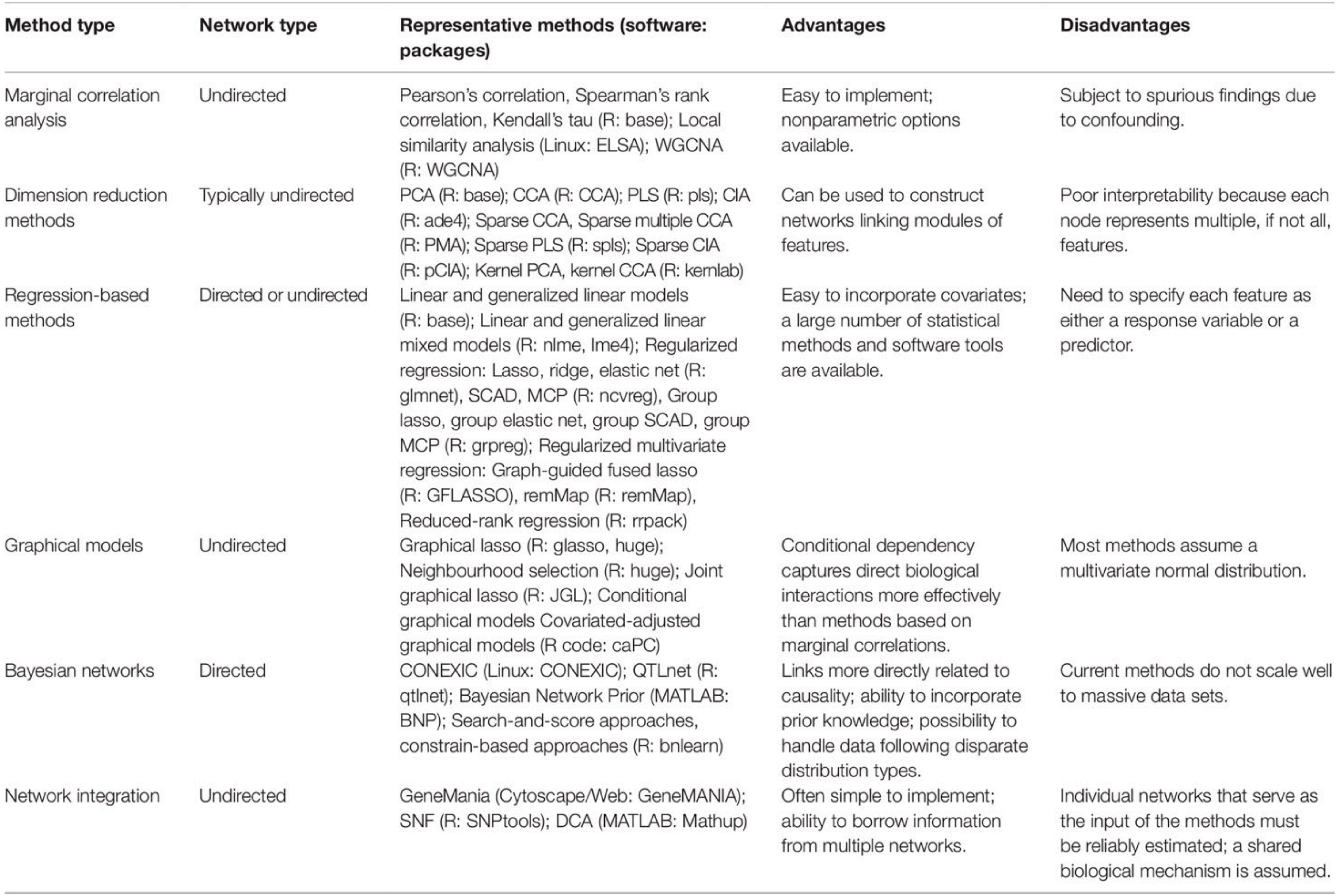
Summary of strengths and weaknesses for each correlation technique



	Bray–Curtis	*CoNet*	*LSA*	*MIC*	*Pearson*	*RMT*	*SparCC*	*Spearman*
Sequencing technology	x							
Compositions	x			x			xx	
Sparsity			x					
Rarefaction iteration number		xx		xx	x		x	x
Distributional preferences			xx	xx			x	x
Three-species linear ecological relationships–40% sparsity			xx	x			x	x
Two-species linear ecological relationships–40% sparsity	x	x	xx	x	x		xx	xx
amensal partial-obligate-syntrophy								
Linear ecological relationships–70% sparsity								
Lotka–Volterra relationships–40% sparsity		x	xx	x		x	x	x
Lotka–Volterra relationships–70% sparsity			xx					
Useful in improved precision ensemble approach–70% sparsity		xx			xx	x	xx	x
		x—moderate performance			xx—the best performance of the tools			

Weiss *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 10:1669–81.

MDA ▷ Microbiota Network



Method type	Network type	Representative methods (software: packages)	Advantages	Disadvantages
Marginal correlation analysis	Undirected	Pearson's correlation, Spearman's rank correlation, Kendall's tau (R: base); Local similarity analysis (Linux: ELSA); WGCNA (R: WGCNA)	Easy to implement; nonparametric options available.	Subject to spurious findings due to confounding.
Dimension reduction methods	Typically undirected	PCA (R: base); CCA (R: CCA); PLS (R: pls); CIA (R: ade4); Sparse CCA, Sparse multiple CCA (R: PMA); Sparse PLS (R: spls); Sparse CIA (R: pCIA); Kernel PCA, kernel CCA (R: kernlab)	Can be used to construct networks linking modules of features.	Poor interpretability because each node represents multiple, if not all, features.
Regression-based methods	Directed or undirected	Linear and generalized linear models (R: base); Linear and generalized linear mixed models (R: nlme, lme4); Regularized regression: Lasso, ridge, elastic net (R: glmnet), SCAD, MCP (R: ncvreg), Group lasso, group elastic net, group SCAD, group MCP (R: grpreg); Regularized multivariate regression: Graph-guided fused lasso (R: GFLASSO), remMap (R: remMap), Reduced-rank regression (R: rrpack)	Easy to incorporate covariates; a large number of statistical methods and software tools are available.	Need to specify each feature as either a response variable or a predictor.
Graphical models	Undirected	Graphical lasso (R: glasso, huge); Neighbourhood selection (R: huge); Joint graphical lasso (R: JGL); Conditional graphical models Covariated-adjusted graphical models (R code: caPC)	Conditional dependency captures direct biological interactions more effectively than methods based on marginal correlations.	Most methods assume a multivariate normal distribution.
Bayesian networks	Directed	CONEXIC (Linux: CONEXIC); QTLnet (R: qtlnet); Bayesian Network Prior (MATLAB: BNP); Search-and-score approaches, constrain-based approaches (R: bnlearn)	Links more directly related to causality; ability to incorporate prior knowledge; possibility to handle data following disparate distribution types.	Current methods do not scale well to massive data sets.
Network integration	Undirected	GeneMania (Cytoscape/Web: GeneMANIA); SNF (R: SNPtools); DCA (MATLAB: Mathup)	Often simple to implement; ability to borrow information from multiple networks.	Individual networks that serve as the input of the methods must be reliably estimated; a shared biological mechanism is assumed.

Jiang et al. (2019) Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. Front. Genet.

MDA ▷ Microbiota Network

ALL
THINGS
CONSIDERED

The first **mathematically described network model** is the random network introduced in 1960 by Paul Erdös and Alfred Rényi. This model assumes a network of randomly interconnected nodes. The nodes' degrees follow a Poisson distribution and most nodes have a number of connections comparable to the network's average degree. Most natural networks show a power-law degree distribution. A few nodes have a very large number of connections, while other nodes have no or few connections.

Erdös and Rényi (1960) On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Sciences.

The concerns over correlation-based analyses have led to the development of methods that are robust to compositionality.

SparCC (Sparse Correlations for Compositional data), for example, is a new technique that uses linear Pearson's correlations between the log-transformed components to infer associations in composi- tional data (Friedman 2012).

SPIEC-EASI (SParse InversE Covariance Estima- tion for Ecological Association Inference) is another statistical method for the inference of microbial ecological networks that combines data transformations developed for compositional data analysis with a graphical model inference framework with the assumption that the underlying ecological association network is sparse .

Friedman J, Alm EJ (2012) Inferring correla- tion networks from genomic survey data. PLoS Comput Biol 8:e1002687.

Kurtz ZD, Muller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA (2015) Sparse and compositionally robust inference of micro- bial ecological networks. PLoS Comput Biol 11:e1004226.

An alternative approach uses probabilistic graphical models (**PGMs**), which provides a probability theory framework based on discrete data structures in computer science, to measure uncer- tainty in high dimensional data.

One method, EBIC- glasso, estimates sparse undirected graphical models for continuous data with multivariate Gaussian distribution through the use of L1 (lasso) regularization before using an extended Bayesian informa- tion criteria (EBIC) to select the most fitting model.

Tibshirani R (2018) Regression shrinkage and selection via the lasso: a retrospective. J Roy Stat Soc Ser B (Stat Method) 73:273–282.

Hubs are nodes in the network that have a significantly larger number of links compared to the other nodes in the network. A hub in a microbiome network can be considered as an equiva- lent to a keystone species in the microbial community.

Two popular approaches have been used to detect keystone taxa from microbiome networks: centrality indices and link-analysis methods.

Tibshirani R (2018) Regression shrinkage and selection via the lasso: a retrospective. J Roy Stat Soc Ser B (Stat Method) 73:273–282.

igraph and **qgraph** are network analysis tools that can be used in R to construct, simulate, analyze, and visualize networks.

The **vegan** package will be used to calculate pairwise (dis)similarity/distance between OTUs as well as to permute the OTU table.

The **MCL** package is needed for cluster detection and **SpiecEasi** will be used to construct microbiome networks.

The effect of P treatment on abundance of each OTU was investigated with edgeR (Robinson et al., 2010a) on TMM-normalized data (Robinson & Oshlack, 2010) and visualized with ternary plots. TMM-normalized data was used to calculate Spearman rank correlations between OTUs for co-occurrence networks. Positive (<0.7) and significant relationships were visualized with igraph (Csardi & Nepusz, 2006).

Scripts, functions and support files are available as Notes S3. The Figure S1 visualizes the workflow of analysis steps.

MDA ▷ Microbiota Network

A wide range of methods, with varying levels of efficiency and accuracy, have been used to construct networks based on microbiome data. **The simplest methods are (dis)similarity- or distance-based techniques. The most popular methods, however, are correlation-based techniques**, where significant pairwise associations between operational taxonomic units (OTUs, a grouping of organisms circumscribed by a specified level of DNA sequence similarity at a marker gene) are detected using a correlation coefficient such as **Pearson's correlation coefficient** or **Spearman's non-parametric rank correlation coefficient**.

The use of correlation coefficients to detect dependencies between members of a microbiome suffers from limitations such as detecting **spurious correlations** due to compositionality, and being severely underpowered owing to the relatively low number of samples.

Layeghifard et al. (2018) Constructing and Analyzing Microbiome Networks in Microbiome Analysis: Methods and Protocols, Methods in Molecular Biology, vol. 1849.