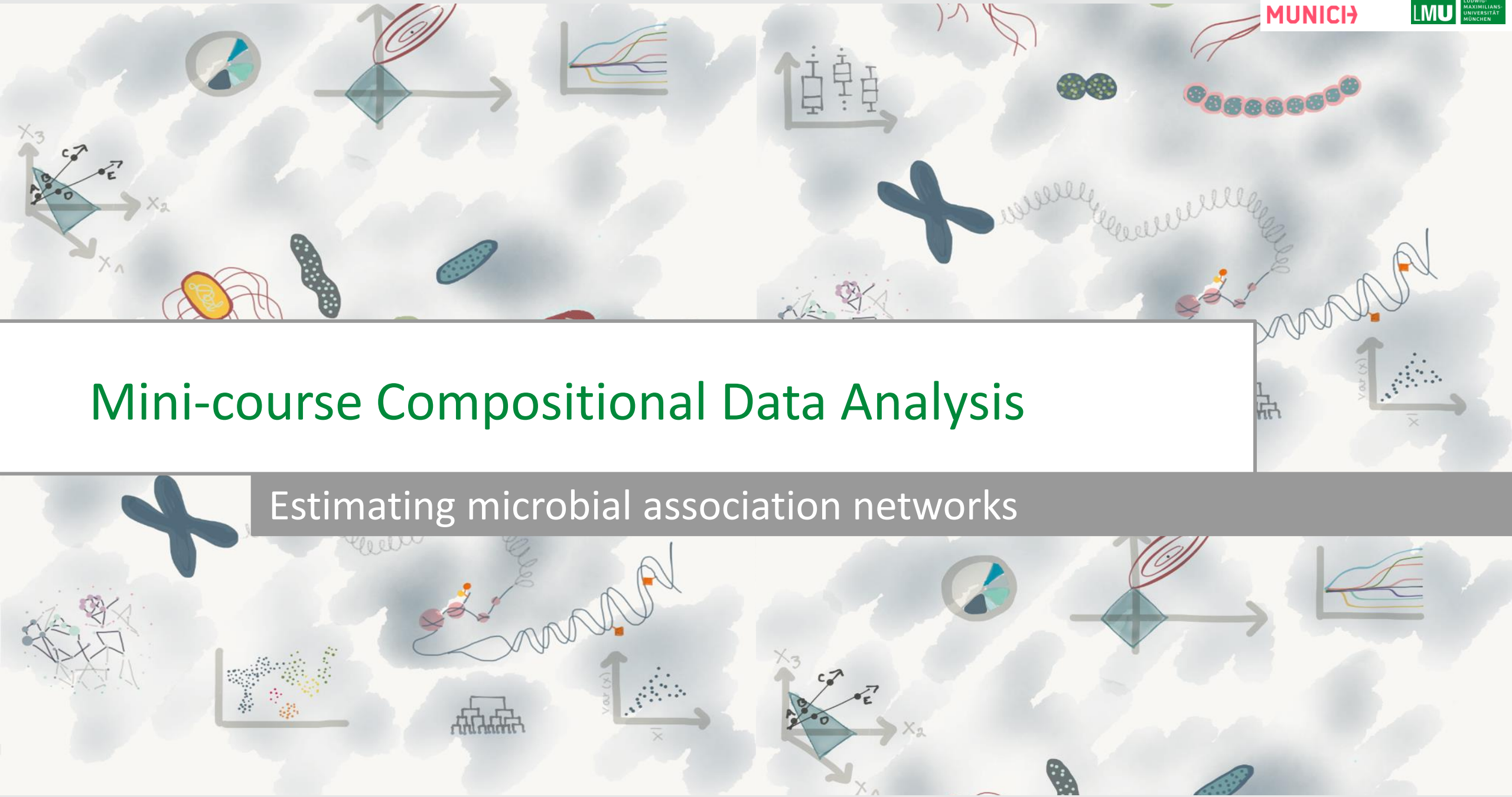
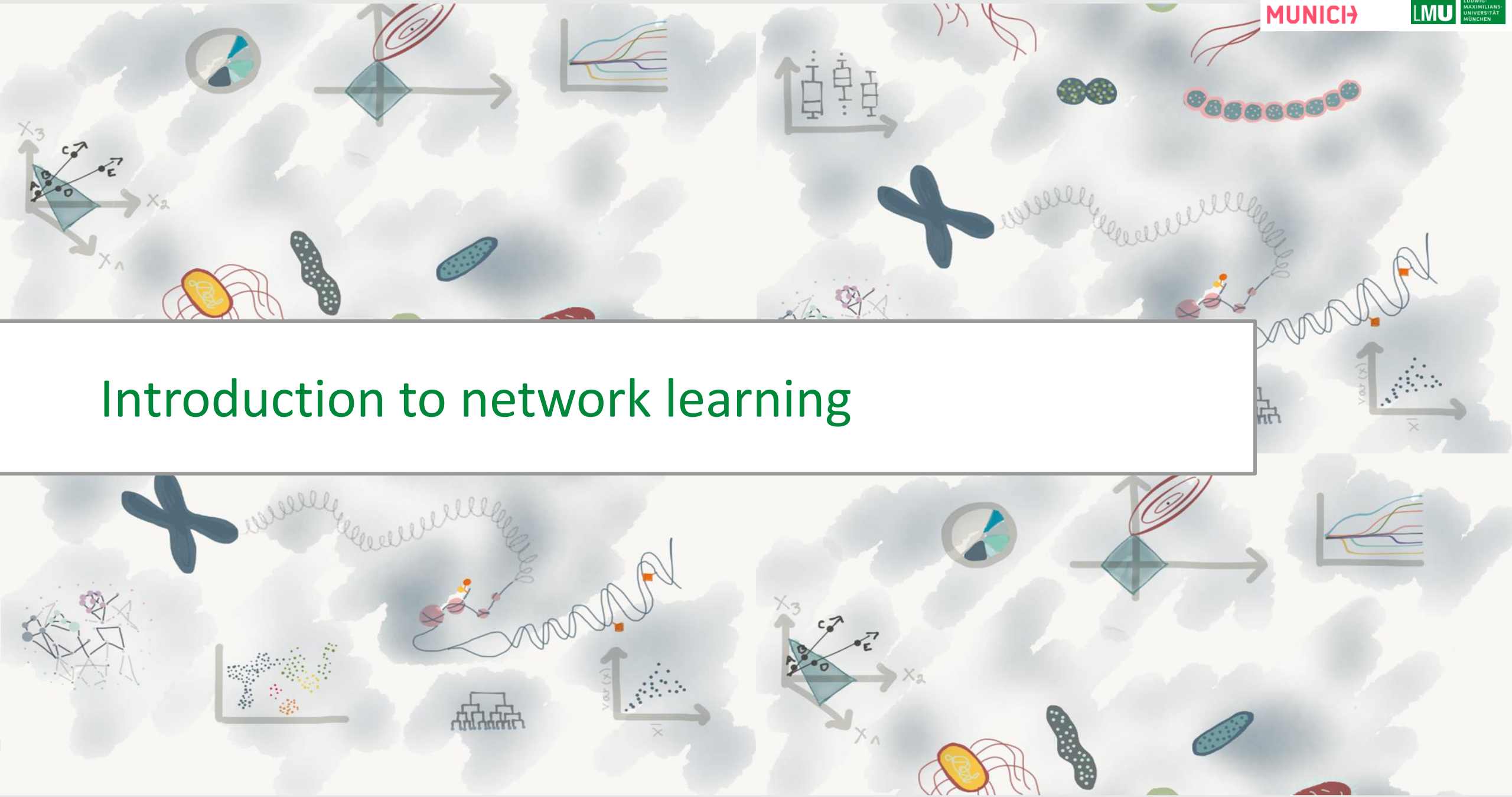


Mini-course Compositional Data Analysis

Estimating microbial association networks

Agenda

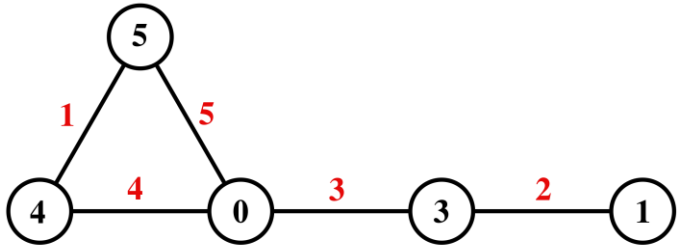
1. Introduction to network learning
2. Association estimation
3. The SPIEC-EASI approach and R package
4. From associations to adjacencies
5. Network analysis



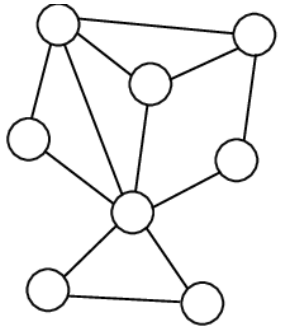
Introduction to network learning

Networks and graphs

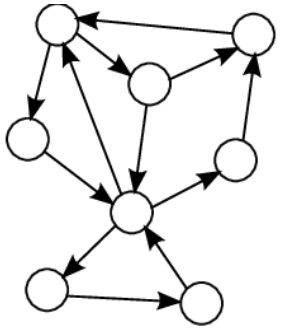
- The term **graph** is used in computer science and math



Graph with labeled **vertices** and **edges**

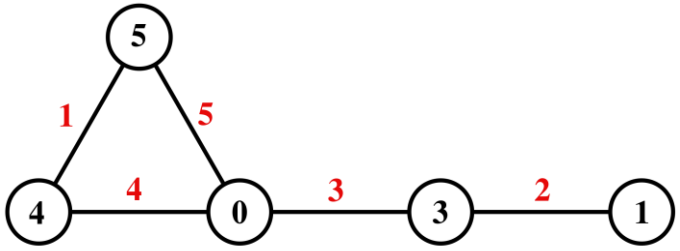


Directed graph



Undirected graph

- The term **network** is used in physics, biology, and social sciences



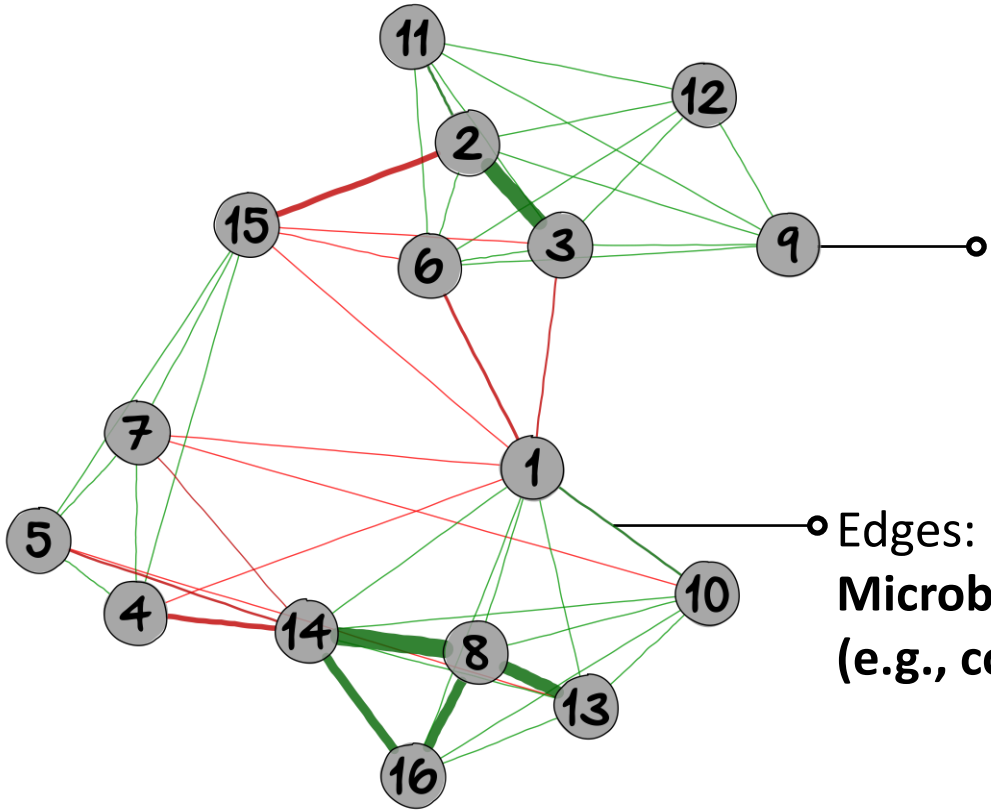
Network with labeled **nodes** and **links**

Networks in cell / molecular biology

- Protein-protein interaction (PPI) networks
- Gene regulatory (co-expression) networks
- Metabolic networks
- Signaling networks
- Neuronal networks
- Microbial (ecological) interaction networks

Microbiome networks

→ Insights into the organizational structure of a microbial community



Nodes / vertices:
ASVs, OTUs, Taxa, …

Edges:
Microbial associations (e.g., correlations)

Association strength and direction:

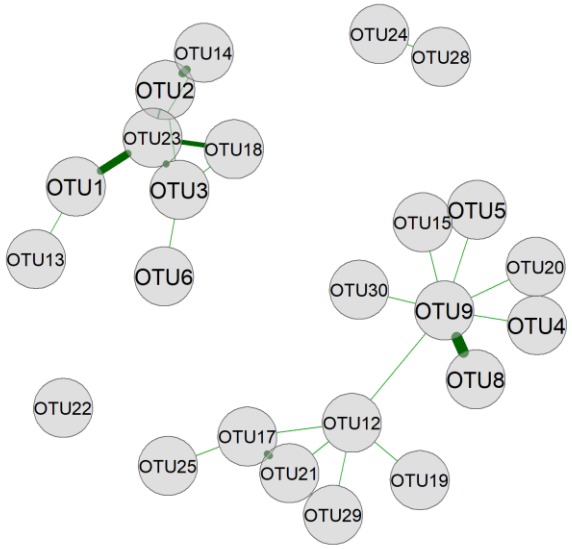
- Strongly positive
- Weakly positive
- Strongly negative
- Weakly negative

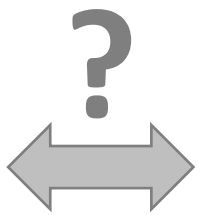
Layout:
Force-directed layout algorithm

Network comparison

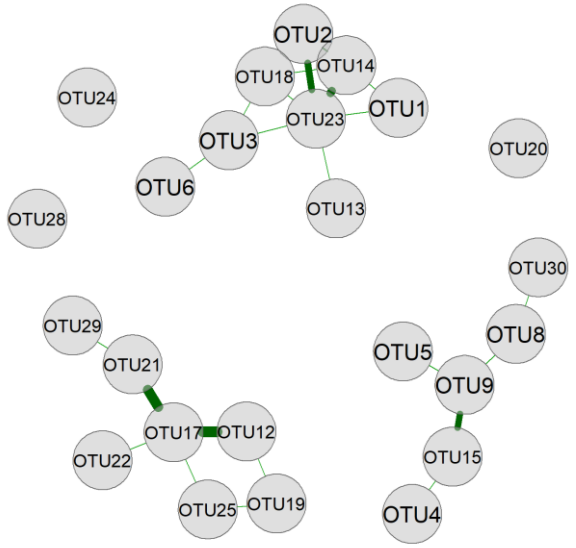
→ Does the microbial composition change across different conditions?

group 1





group 2

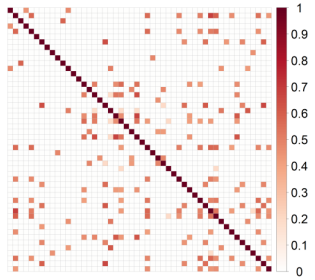


Possible groups:

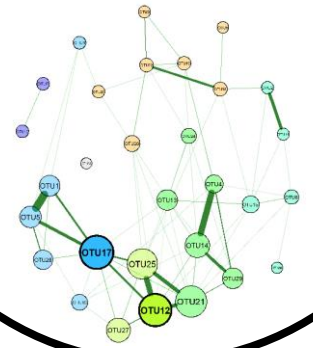
- Cases and controls
- Two environmental states
- Two time points
- …

Typical network analysis workflow

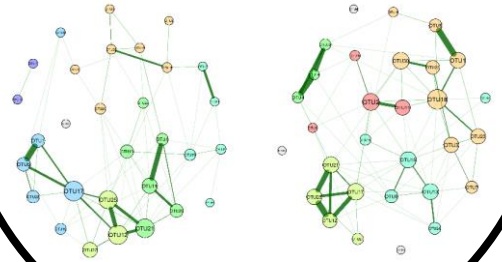
Network construction



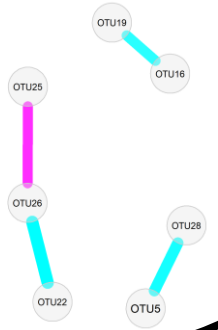
Network analysis



Network comparison



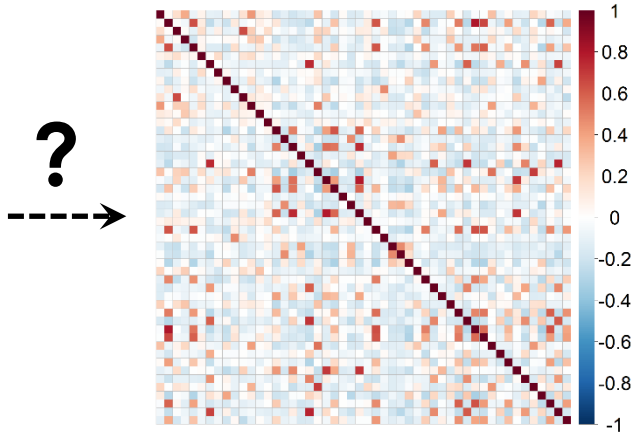
Differential network



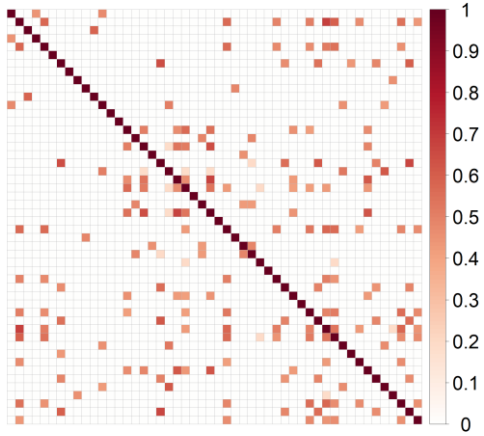
From sequencing data to networks

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}$$

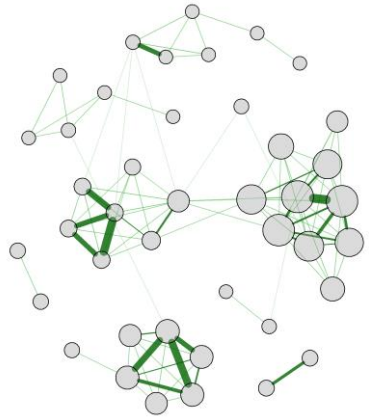
with $p \gg n$



Association matrix



Adjacency matrix



Association network

Data characteristics:

- We observe only a **sample** of the true microbial composition
- **Compositional** (only relative information)
- **Sequencing depth** (total number of reads) **varies** across samples
- **Zero-inflated**
- **High-dimensional** (number of taxa p >> sample size n)

Association estimation

Association estimation

Zero replacement → Normalization / Transformation → Association estimation

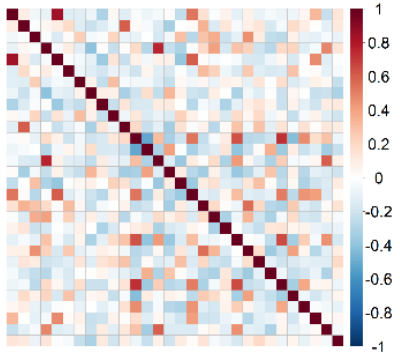
	OTU 1	OTU 2	OTU 3	OTU 4	…
sample 1	1452	0	0	0	
sample 2	5414	0	0	2513	
sample 3	1942	0	0	381	
sample 4	1089	0	0	468	
sample 5	3896	0	0	16	
sample 6	1907	0	0	3089	
sample 7	3915	0	30	263	
sample 8	1626	0	0	0	
sample 9	735	0	0	43	
sample 10	7449	60	0	471	
…					

pseudo →

	OTU 1	OTU 2	OTU 3	OTU 4	…
sample 1	1453	1	1	1	
sample 2	5415	1	1	2514	
sample 3	1943	1	1	382	
sample 4	1090	1	1	469	
sample 5	3897	1	1	17	
sample 6	1908	1	1	3090	
sample 7	3916	1	31	264	
sample 8	1627	1	1	1	
sample 9	736	1	1	44	
sample 10	7450	61	1	472	
…					

clr →

	OTU 1	OTU 2	OTU 3	OTU 4	…
sample 1	3,12	-4,16	-4,16	-4,16	
sample 2	3,29	-5,31	-5,31	2,52	
sample 3	2,85	-4,73	-4,73	1,22	
sample 4	1,74	-5,25	-5,25	0,90	
sample 5	3,43	-4,83	-4,83	-2,00	
sample 6	2,32	-5,23	-5,23	2,80	
sample 7	2,06	-6,21	-2,77	-0,63	
sample 8	2,66	-4,73	-4,73	-4,73	
sample 9	1,90	-4,70	-4,70	-0,92	
sample 10	4,31	-0,49	-4,61	1,55	
…					

→ 

- Pseudo counts
- Methods from **zCompositions** R package (Martín-Fernández et al., 2011)
- …

- Centered log-ratio (**clr**) transformation
- Variance stabilizing transformation (**VST**)
- …

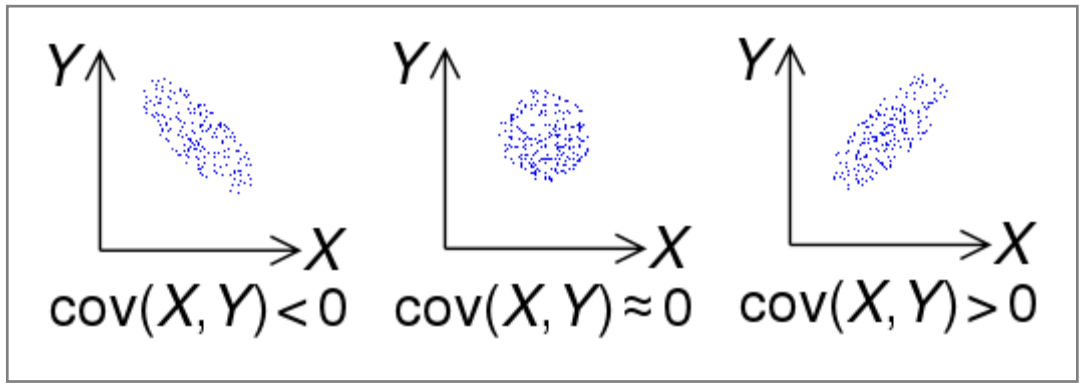
⟶ (Badri et al., 2020)

- Correlation
- Conditional dependence
- Proportionality

Packages often include the preprocessing steps, e.g.,
- SparCC
- SPRING
- SpiecEasi

References:
- Badri M, Kurtz ZD, Bonneau R, et al. Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genom Bioinform* 2020. doi: 10.1093/nargab/lqaa100.
- Martín-Fernández JA, Palarea-Albaladejo J, Olea RA. Dealing with zeros. *Compositional data analysis*, 2011, 43–58.

Correlation as association measure



Source: Wikipedia

Correlation:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

How to estimate the covariance matrix?

Covariance estimation problem

Let $x_1, \dots, x_n$ be independent observations of a p-dimensional random vector $X \in \mathbb{R}^{px1}$, and $\boldsymbol{p \gg n}$.

We want to find an estimate $\hat{\Sigma}$ of the **covariance matrix** $\boldsymbol{\Sigma} = \mathbb{E}\left[\left(x - \mathbb{E}(x)\right)\left(x - \mathbb{E}(x)\right)^T\right] \in S^{pxp}$

The **sample covariance matrix** $\boldsymbol{S}$ has entries

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ and x_{ki} is the k-th observation of the variable X_i.

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}$$

Properties:

- $\boldsymbol{S}$ is an **unbiased** estimator of Σ
- In the Gaussian case, $\boldsymbol{S}$ is the **ML estimator** of the covariance matrix

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & & s_{2p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

$\boldsymbol{p \gg n}$ case:

- $\boldsymbol{S}$ has no full rank
- $\boldsymbol{S}$ singular (non-invertible)
- $\boldsymbol{S}$ not positive definite

$\boldsymbol{S}$ is no good approximation if $p \gg n$!

Solutions for high-dimensional data (a few examples)

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & & s_{2p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

Preprocessed data set (p>>n)

Sample covariance matrix (singular)

Shrinkage [1] (correlation)

$$\begin{pmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & & \hat{\sigma}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}$$

Sparse covariance matrix (invertible)

$$\frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

SparCC [2] (correlation)

$$\begin{pmatrix} \hat{\rho}_{11} & \cdots & \hat{\rho}_{1p} \\ \hat{\rho}_{21} & & \hat{\rho}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\rho}_{p1} & \cdots & \hat{\rho}_{pp} \end{pmatrix}$$

Sparse correlation matrix

$$\hat{\Omega} = \hat{\Sigma}^{-1}$$

SpiecEasi [3] (conditional dependence)

$$\begin{pmatrix} \hat{\omega}_{11} & \cdots & \hat{\omega}_{1p} \\ \hat{\omega}_{21} & & \hat{\omega}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\omega}_{p1} & \cdots & \hat{\omega}_{pp} \end{pmatrix}$$

Sparse inverse covariance matrix

$$\hat{\hat{\rho}}_{ij} = \frac{-\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}}$$

$$\begin{pmatrix} \hat{\hat{\rho}}_{11} & \cdots & \hat{\hat{\rho}}_{1p} \\ \hat{\hat{\rho}}_{21} & & \hat{\hat{\rho}}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\hat{\rho}}_{p1} & \cdots & \hat{\hat{\rho}}_{pp} \end{pmatrix}$$

Partial correlation matrix

[1] Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 1–30.

Shrinkage estimator of the covariance matrix

Given:

- Unstructured estimator:
  Sample covariance matrix S
- Structured estimator:
  "shrinkage target" T
- "Shrinkage constant" $\delta \in [0,1]$

Find a compromise between the two matrices via:

$$\delta T + (1-\delta)S$$

$\Rightarrow S$ is shrunk towards the structured estimator.

Shrinkage estimator:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Shrink}} = \widehat{\boldsymbol{\delta}}^{*}\boldsymbol{T} + \left(\mathbf{1} - \widehat{\boldsymbol{\delta}}\right)\mathbf{S}$$

Commonly used shrinkage targets

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

Target A: "diagonal, unit variance"
0 estimated parameters

$$t_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(s_{ij}) + \sum_i \widehat{\mathrm{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii}-1)^2}$$

$$T = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

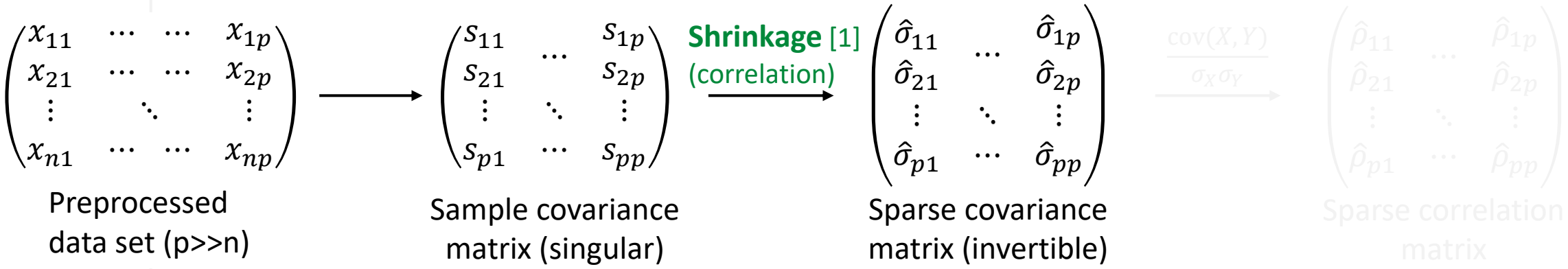
Target B: "diagonal, common variance"
1 estimated parameter: v

$$t_{ij} = \begin{cases} v = \mathrm{avg}(s_{ii}) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(s_{ij}) + \sum_i \widehat{\mathrm{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii}-v)^2}$$

$$T = \begin{pmatrix} v & 0 & \cdots & 0 \\ 0 & v & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v \end{pmatrix}$$

Solutions for high-dimensional data (a few examples)



SparCC [2] (correlation)

$\frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$

Sparse correlation matrix

$\hat{\Omega} = \hat{\Sigma}^{-1}$

SpiecEasi [3] (conditional dependence)

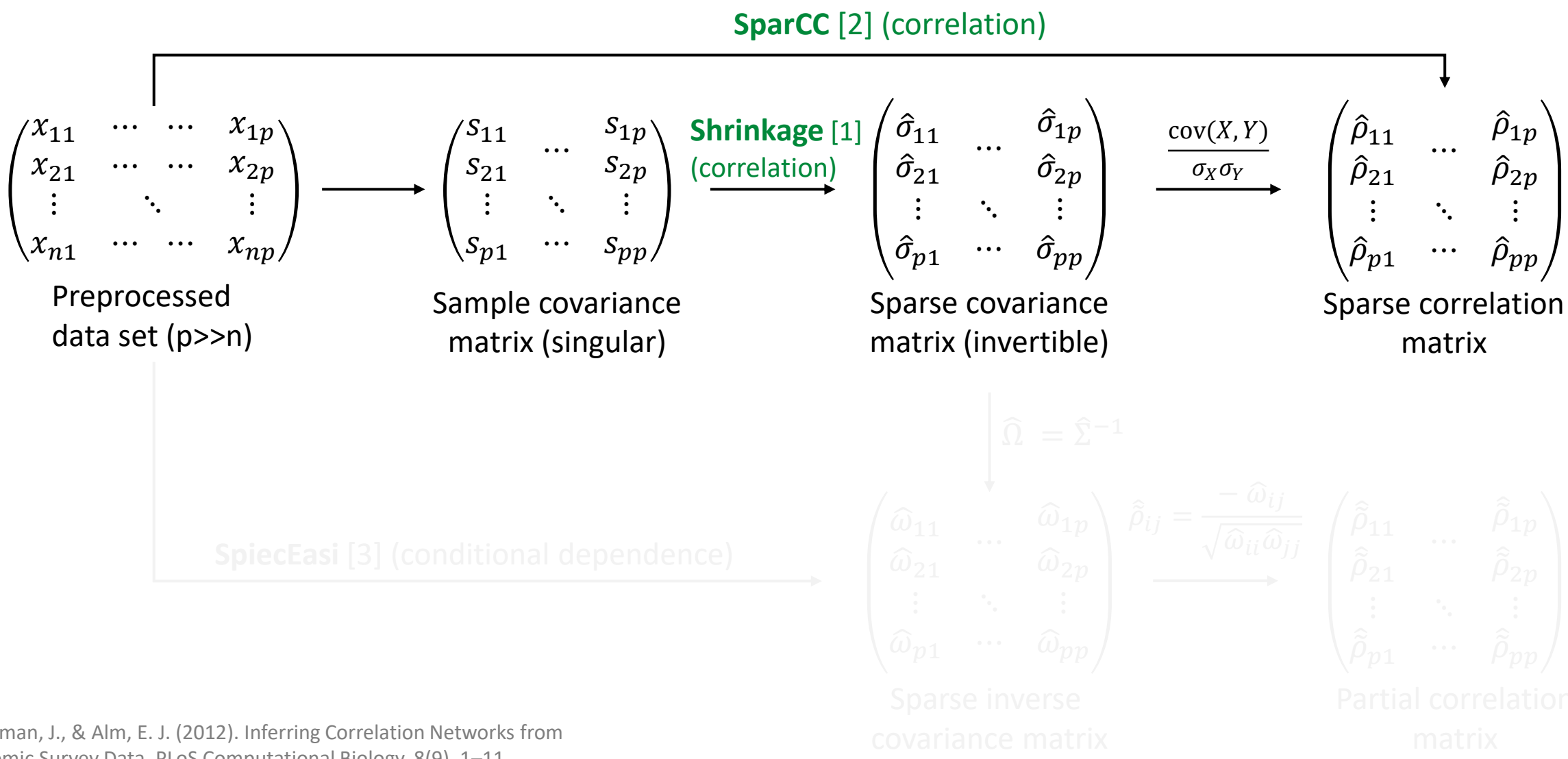
Sparse inverse covariance matrix

$\hat{\hat{\rho}}_{ij} = \frac{-\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}}$

Partial correlation matrix

[1] Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 1–30.

Solutions for high-dimensional data (a few examples)



[2] Friedman, J., & Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology, 8(9), 1–11.

Solutions for high-dimensional data (a few examples)

SparCC [2] (correlation)

$$\begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}$$

Preprocessed data set (p>>n)

$$\begin{pmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & & s_{2p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

Sample covariance matrix (singular)

Shrinkage [1] (correlation)

$$\begin{pmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & & \hat{\sigma}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}$$

Sparse covariance matrix (invertible)

$$\frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\begin{pmatrix} \hat{\rho}_{11} & \cdots & \hat{\rho}_{1p} \\ \hat{\rho}_{21} & & \hat{\rho}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\rho}_{p1} & \cdots & \hat{\rho}_{pp} \end{pmatrix}$$

Sparse correlation matrix

$$\widehat{\Omega} = \hat{\Sigma}^{-1}$$

SpiecEasi [3] (conditional dependence)

$$\begin{pmatrix} \hat{\omega}_{11} & \cdots & \hat{\omega}_{1p} \\ \hat{\omega}_{21} & & \hat{\omega}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\omega}_{p1} & \cdots & \hat{\omega}_{pp} \end{pmatrix}$$

Sparse inverse covariance matrix

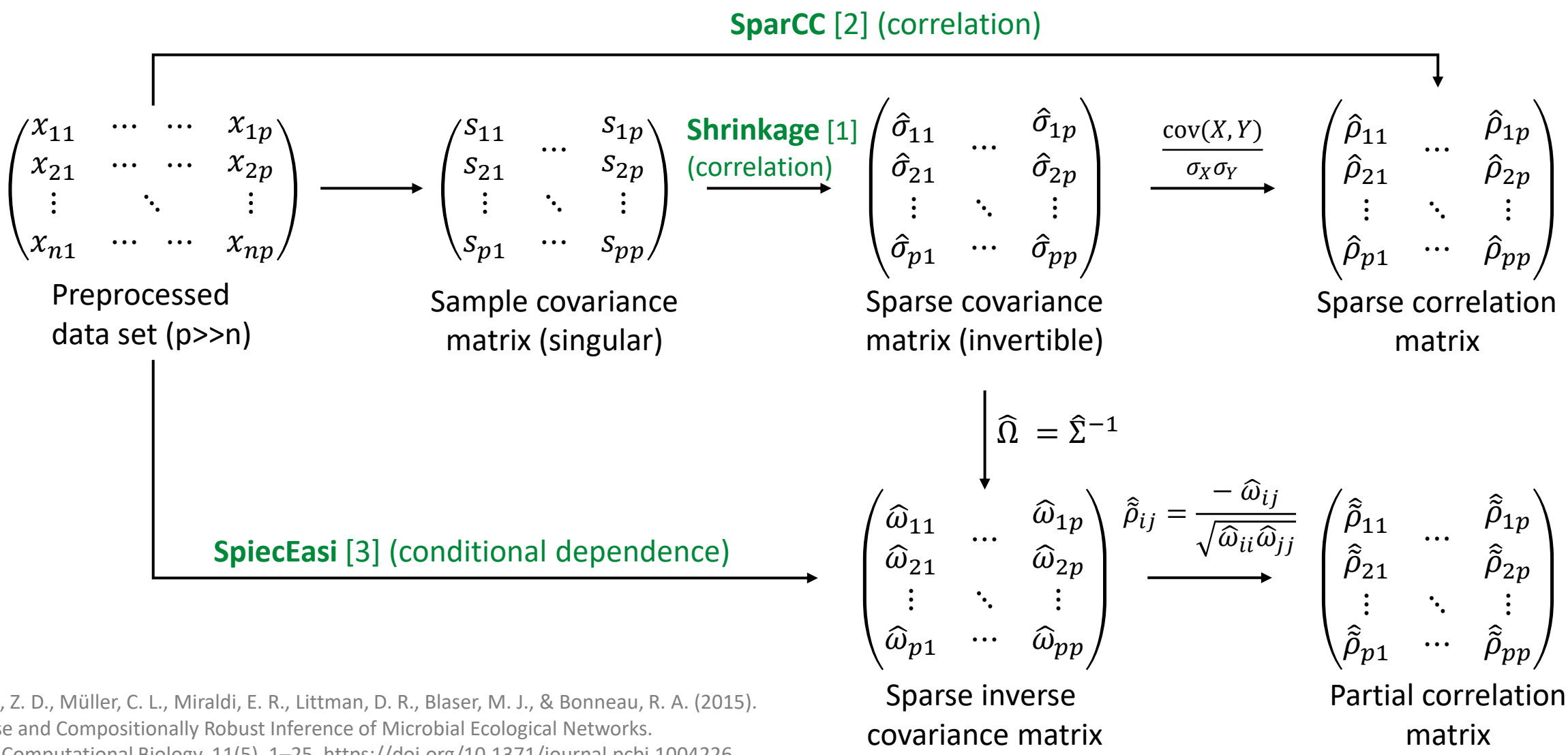
$$\hat{\hat{\rho}}_{ij} = \frac{-\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}}$$

$$\begin{pmatrix} \hat{\hat{\rho}}_{11} & \cdots & \hat{\hat{\rho}}_{1p} \\ \hat{\hat{\rho}}_{21} & & \hat{\hat{\rho}}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\hat{\rho}}_{p1} & \cdots & \hat{\hat{\rho}}_{pp} \end{pmatrix}$$

Partial correlation matrix

[3] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25. https://doi.org/10.1371/journal.pcbi.1004226

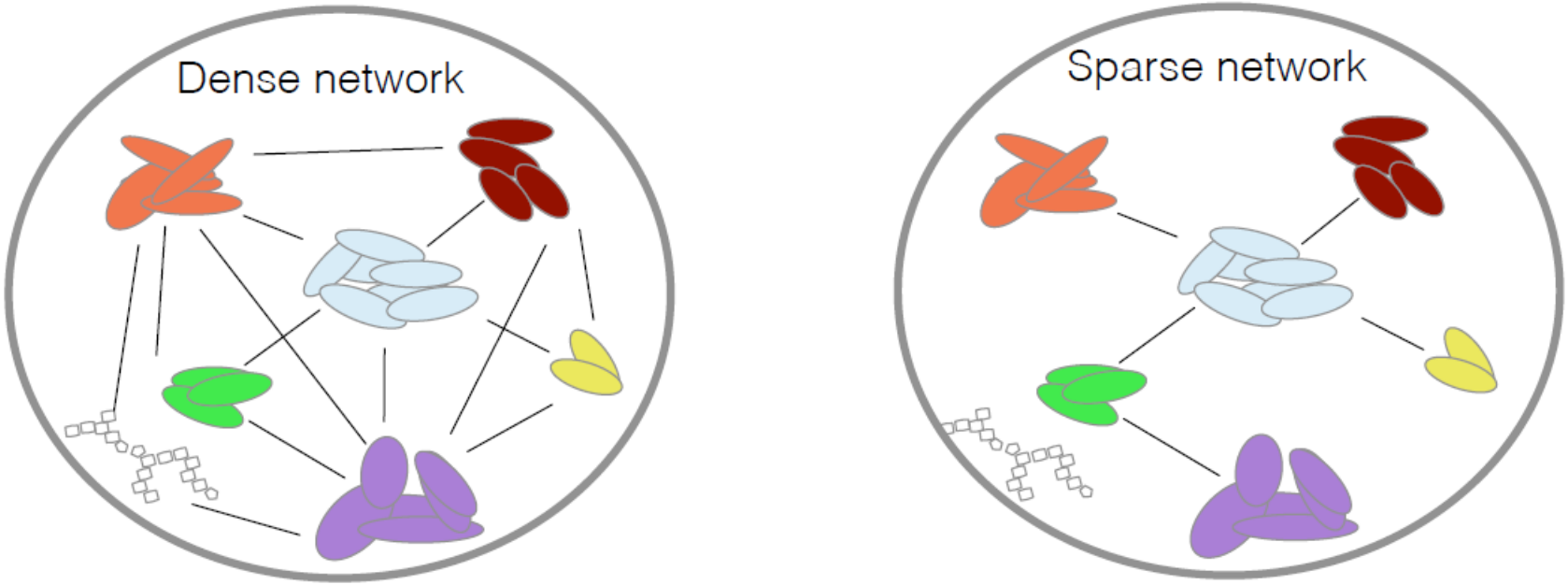
Solutions for high-dimensional data (a few examples)



[3] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25. https://doi.org/10.1371/journal.pcbi.1004226

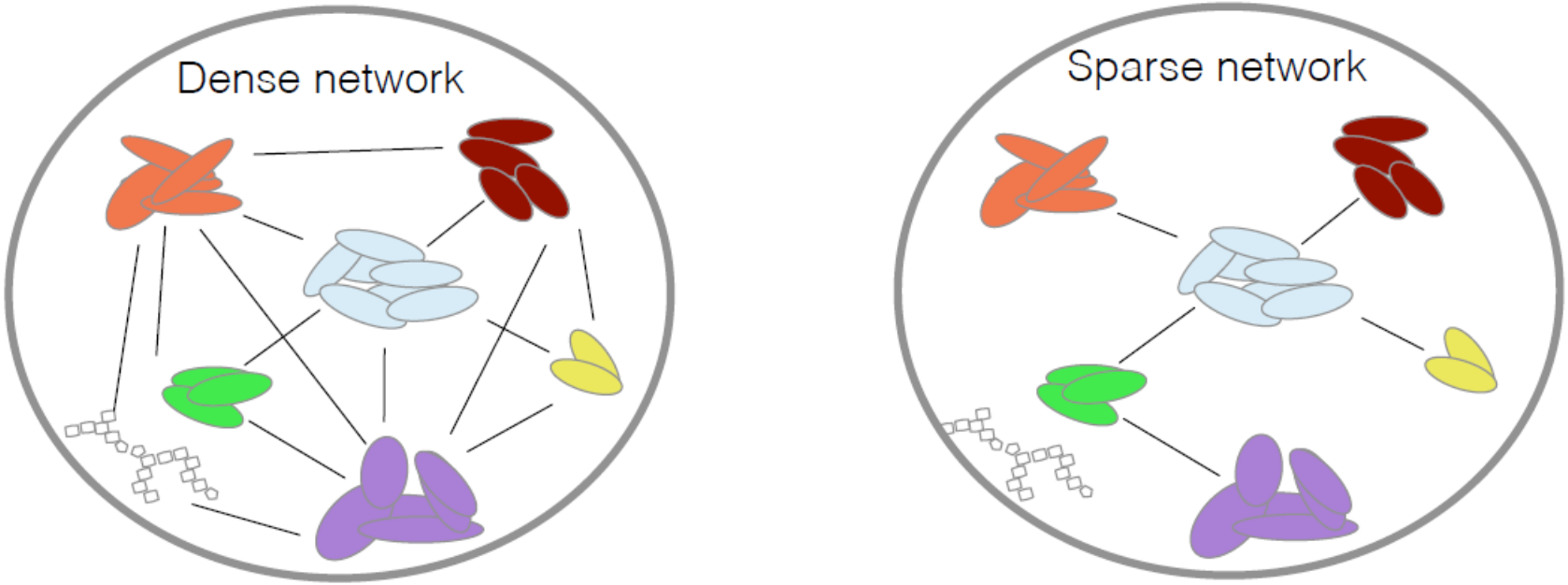
Key hypothesis for network inference

The inverse covariance matrix among transformed microbial relative abundances is sparse.



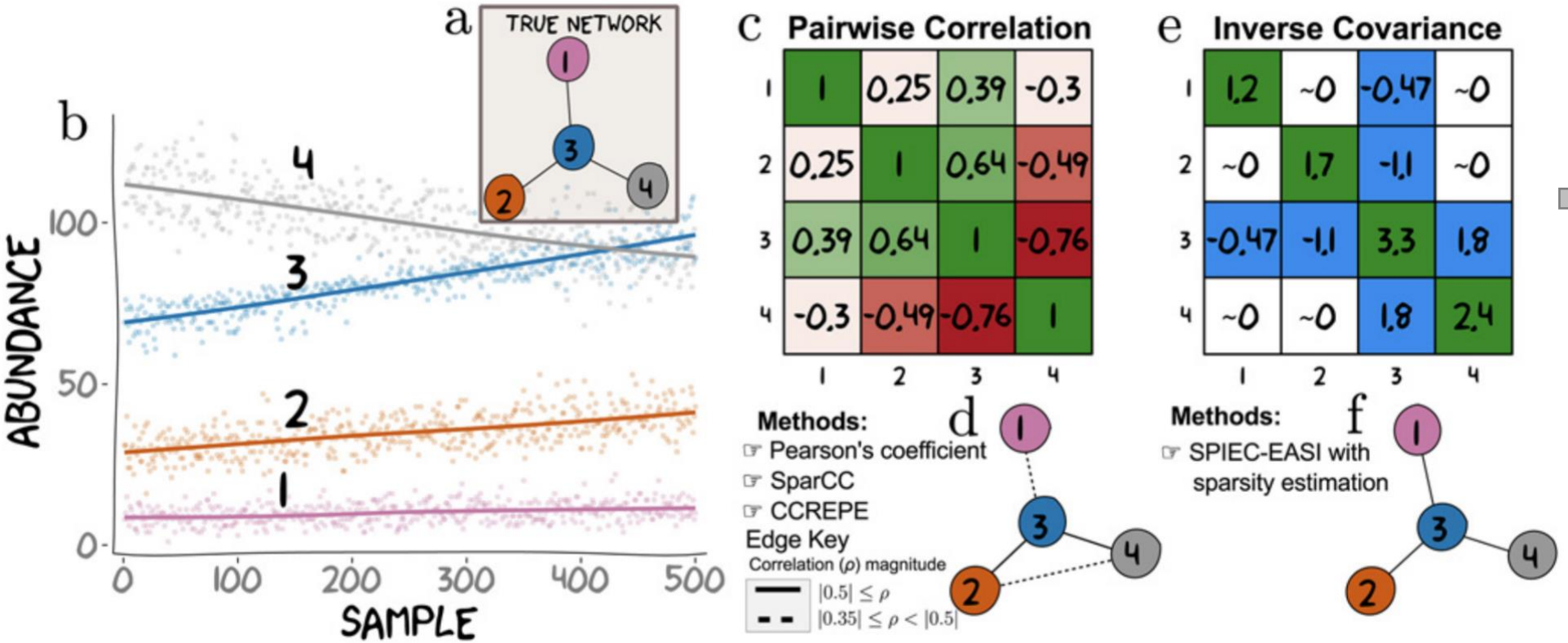
Key hypothesis for network inference

The network of interactions between the different microbes is sparse.



The **SPIEC-EASI** approach and R package

Conditional dependence vs. correlation



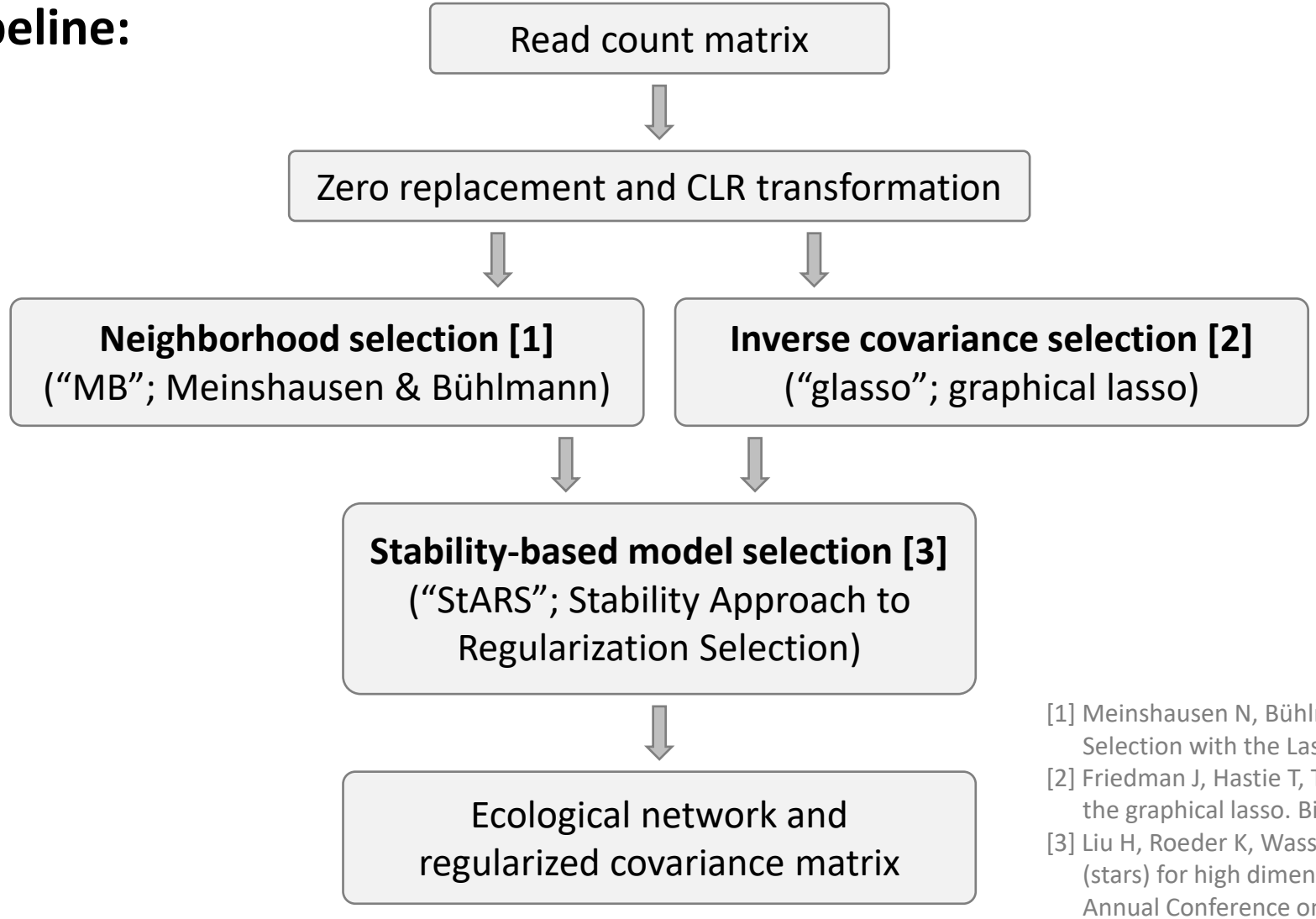
Conditional dependence / Partial correlation

Reference:

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25.

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

Pipeline:



[1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.

[2] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England) 9: 432–441.

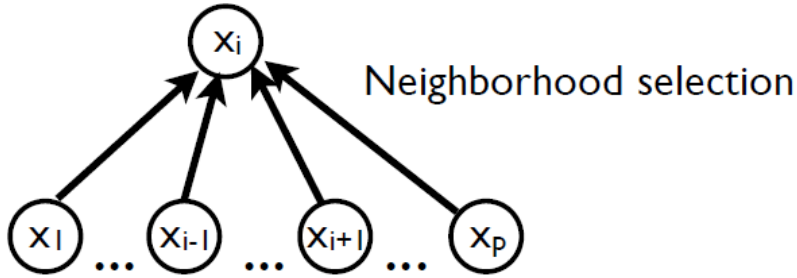
[3] Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS): 1–14.

Sparse neighborhood selection (MB)

- Uses sparse linear regression
- Proposed by Meinshausen and Bühlmann, 2006 (MB) [1]
- Idea: Find a **sparse weighted graph** by **node-wise linear regression:**
  Use each column as response and solve the LASSO problem

OUT/gene *i* all other OTUs/genes

$$n \left\{ X^i = X^{\neg i} \times \beta^* \right\} p\text{-}1 + \sigma \epsilon$$



[1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.

Sparse neighborhood selection (MB)

- Uses sparse linear regression
- Proposed by Meinshausen and Bühlmann, 2006 (MB) [1]
- Idea: Find a **sparse weighted graph** by **node-wise linear regression:** Use each column as response and solve the LASSO problem

LASSO regression [2]:

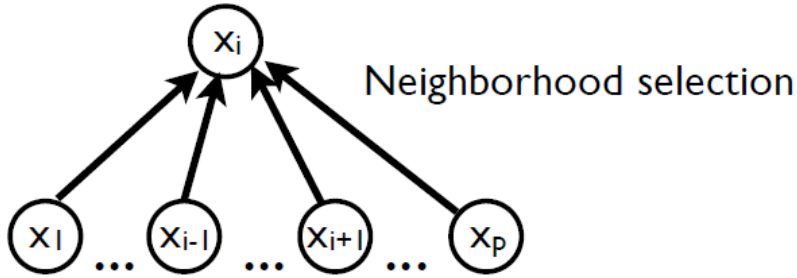
$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}$$

Likelihood term

Sparsity

OUT/gene *i* all other OTUs/genes

$$n \left\{ X^i \right. = X^{\neg i} \times \left. \begin{matrix} \\ \beta^* \\ \\ \end{matrix} \right\} p\text{-}1 + \sigma \epsilon$$

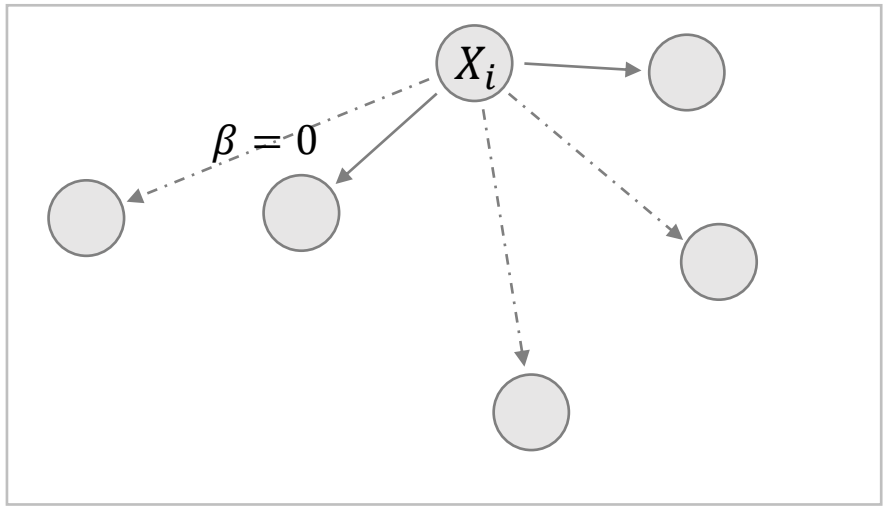


[1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.
[2] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Source: Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Sparse neighborhood selection (MB)

OUT/gene *i* all other OTUs/genes

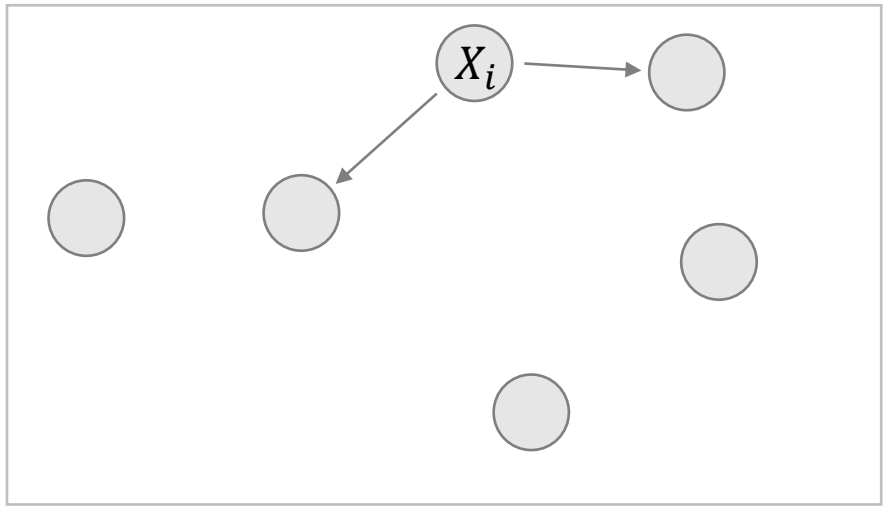
$$n \left\{ X^{i} \right. = X^{\neg i} \times \left. \beta^{*} \right\} p\text{-}1 + \sigma \epsilon$$



Sparse neighborhood selection (MB)

OUT/gene *i* all other OTUs/genes

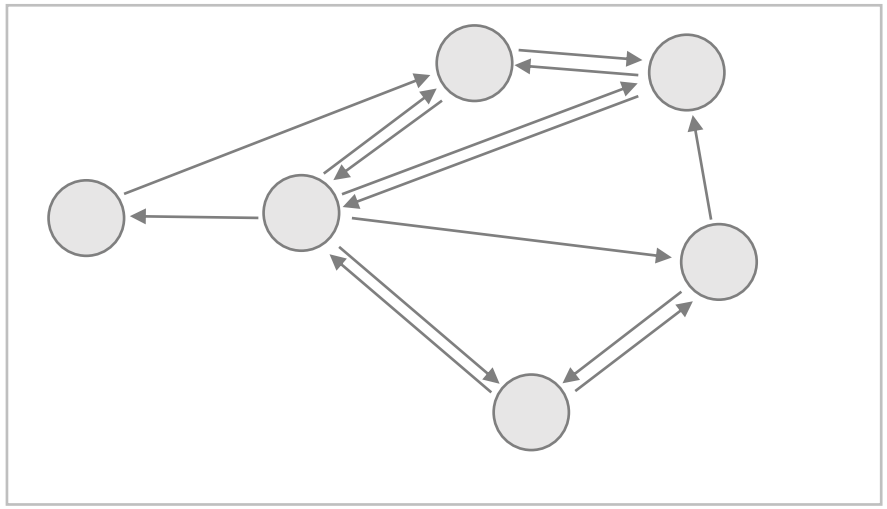
$$n \left\{ X^{i} \right. = X^{\neg i} \times \left. \beta^{*} \right\} p\text{-}1 + \sigma\, \epsilon$$

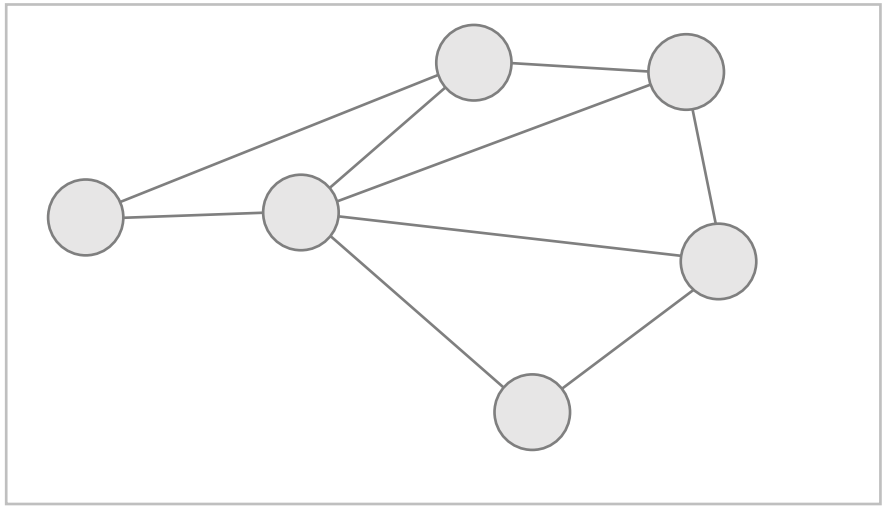


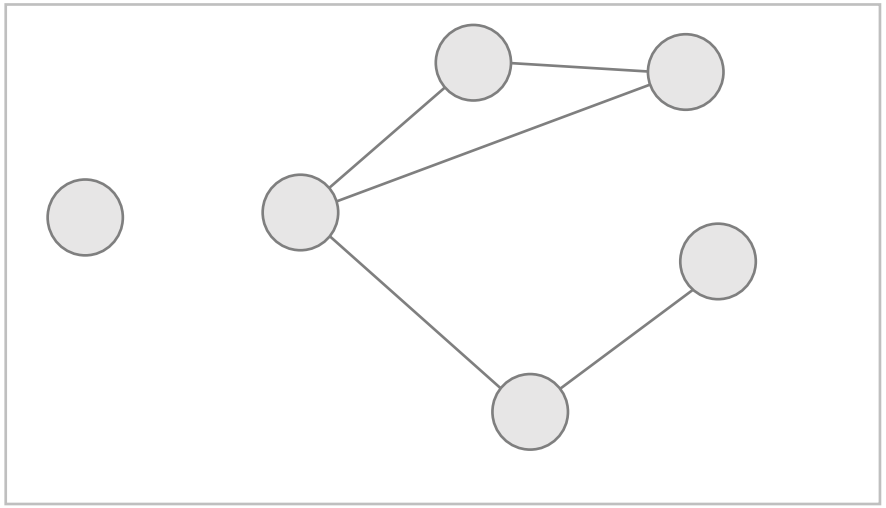
Sparse neighborhood selection (MB)

OUT/gene *i* all other OTUs/genes

$$n \left\{ X^{i} \right. = X^{\neg i} \times \left. \beta^{*} \right\} p\text{-}1 + \sigma \epsilon$$







Sparse graphical model inference (GLASSO)

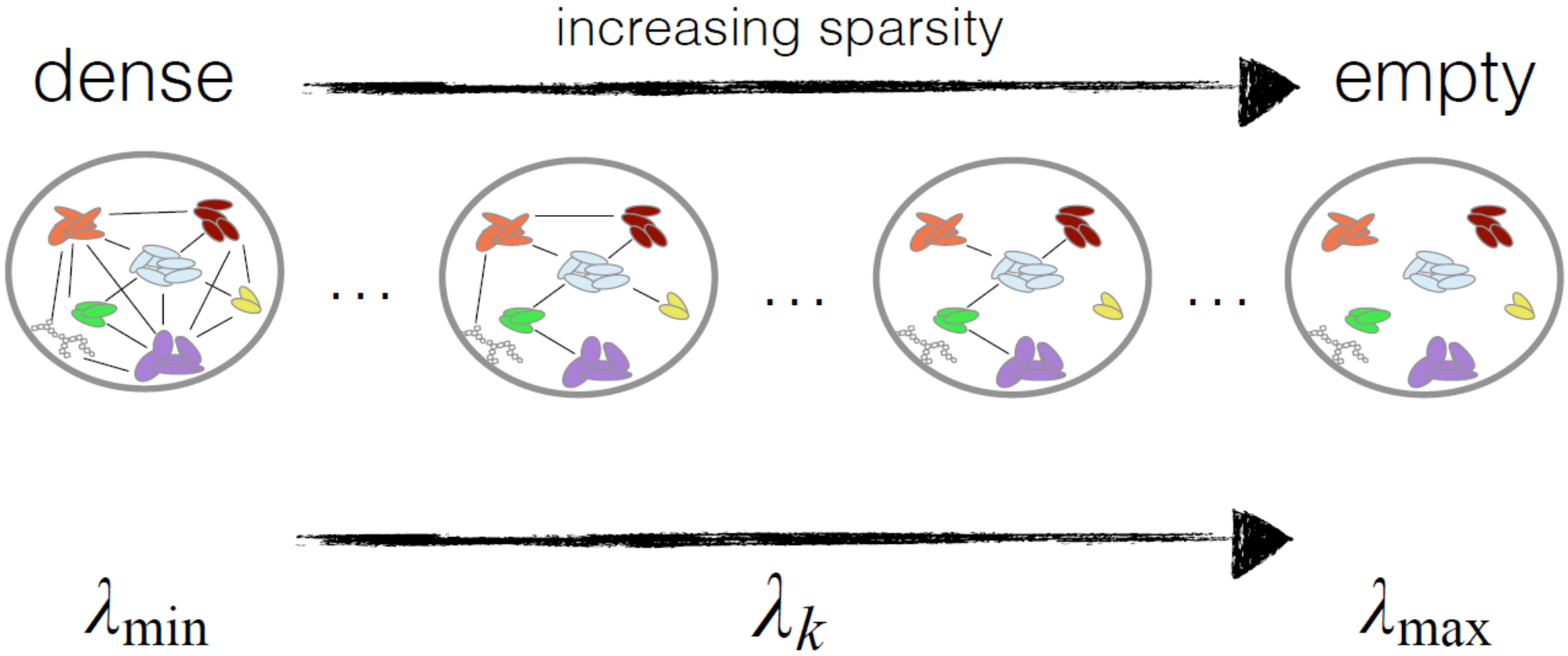
- Sparsity of the underlying network means that the inverse C^{-1} of the correlation (covariance) matrix C is sparse:
  sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C^{-1} by convex optimization

$$C^{-1} = \arg\min_{C^{-1} \in PD} -\log\det(C^{-1}) + \mathrm{tr}(C^{-1}S) + \lambda\|C^{-1}\|_1$$

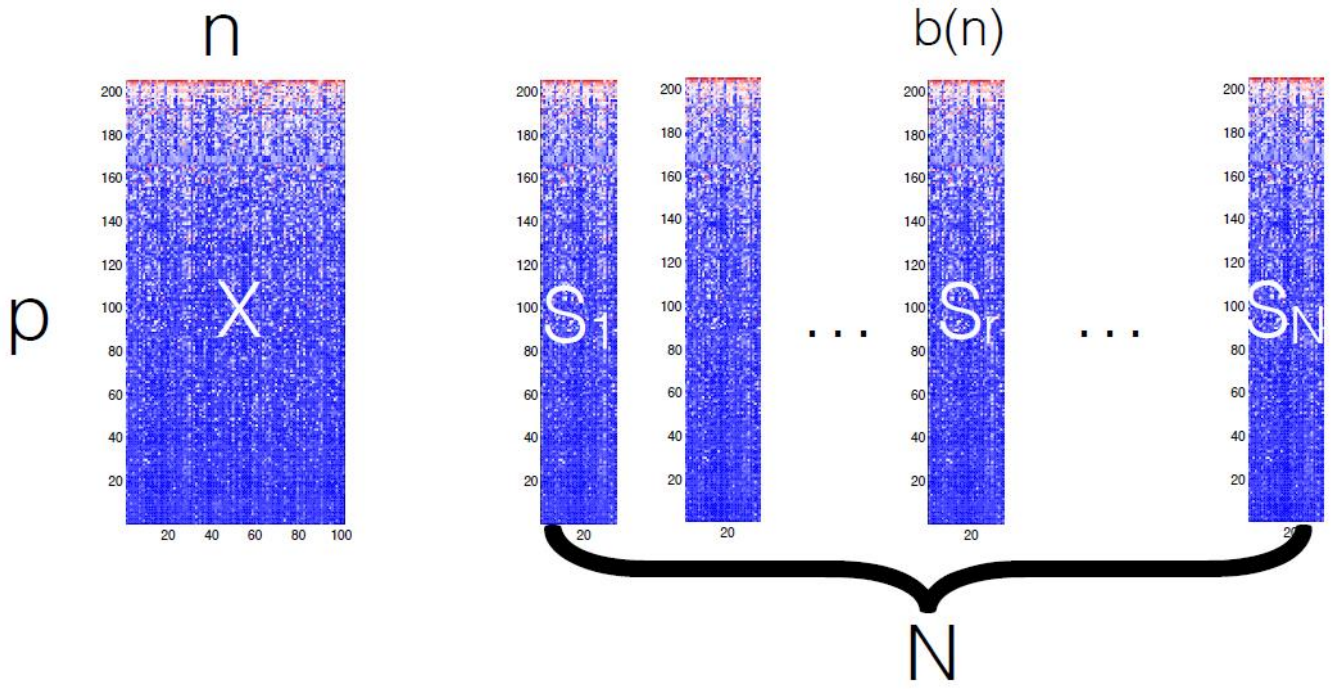
Likelihood term

Sparsity term

How to choose the tuning parameter λ?



Stability-based model selection with StARS



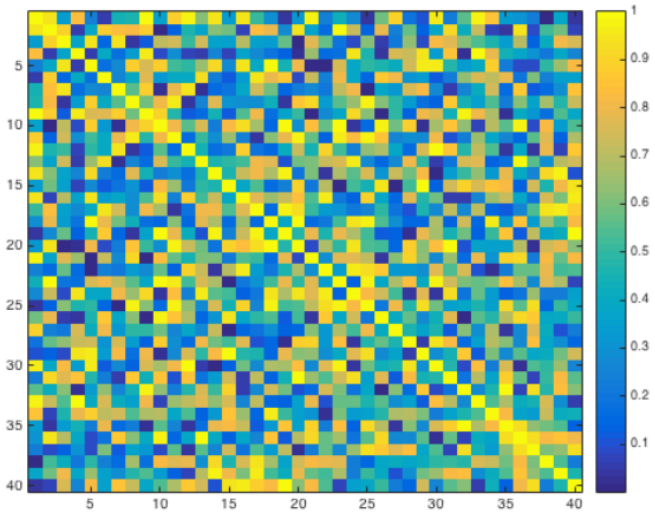
- Draw N random subsamples $S_1, \ldots, S_N$
- For each subsample, estimate the graph using a sequence of λ values: $\{\lambda_1, \ldots, \lambda_K\}$

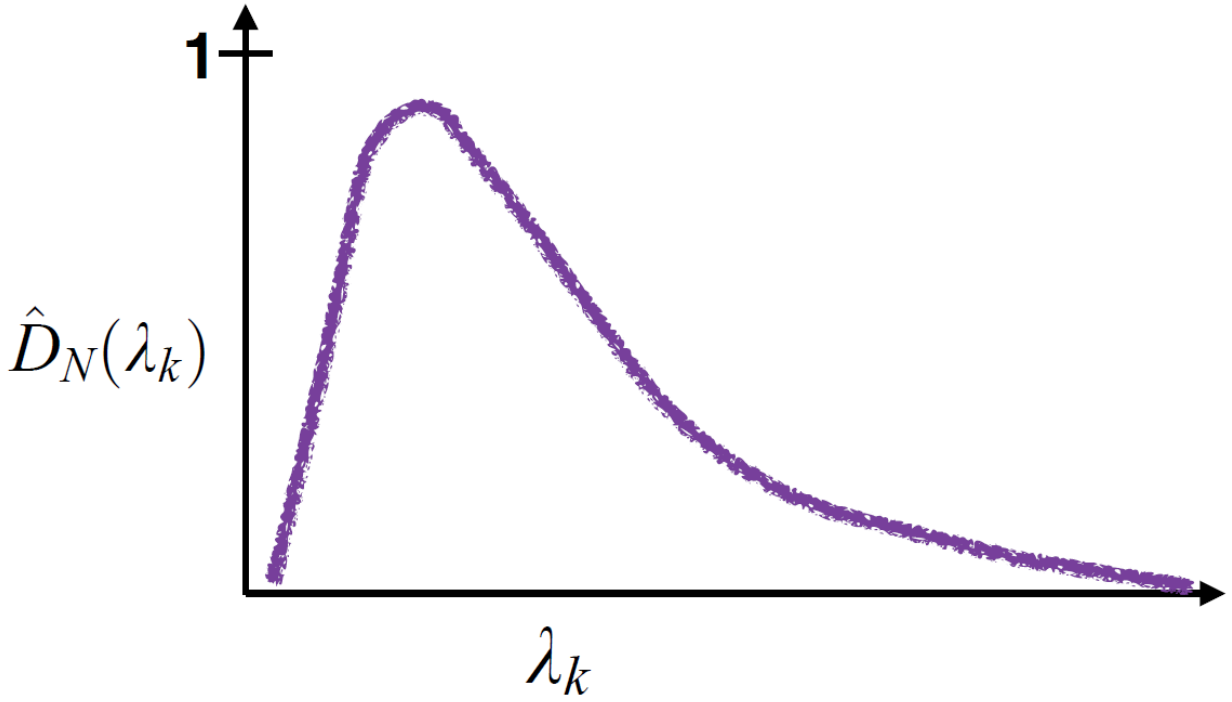
Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010. https://doi.org/10.48550/arxiv.1006.3316

Stability-based model selection with StARS

- For each lambda we estimate the matrix of edge probabilities
- For each lambda value sum up the **variances** of edge probabilities and call it $\hat{D}_N(\lambda_k)$
- Select the lambda where the sum of variances $\hat{D}_N(\lambda_k)$ is below a certain threshold β.

Matrix of edge probabilities

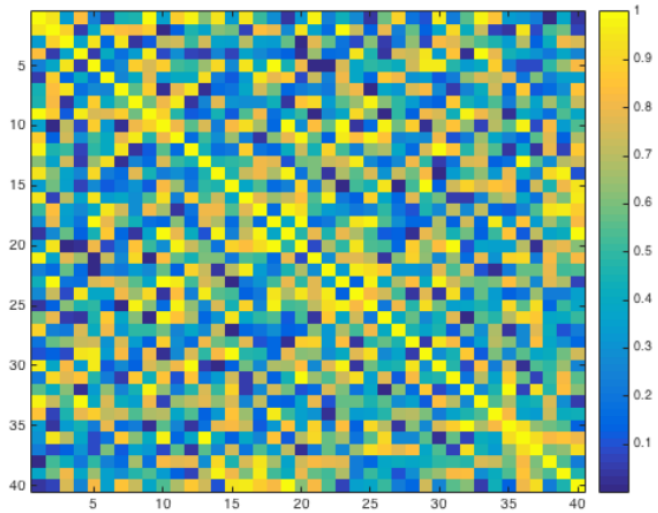


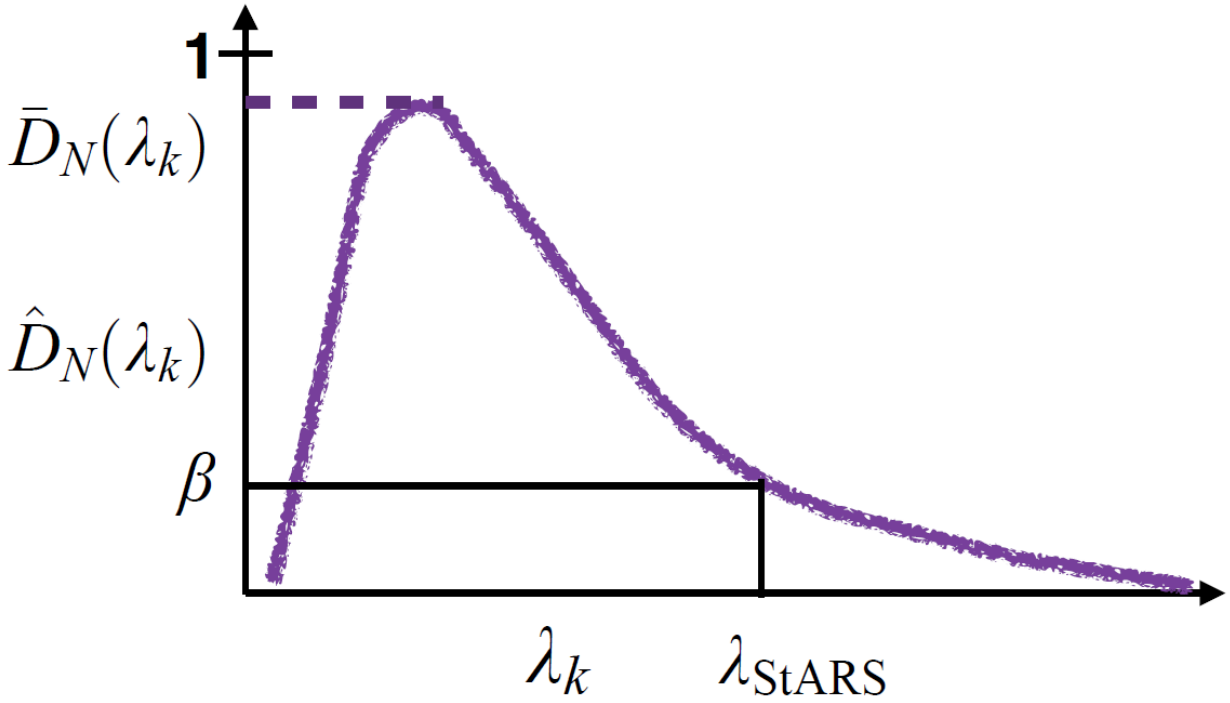


Stability-based model selection with StARS

- For each lambda we estimate the matrix of edge probabilities
- For each lambda value sum up the **variances** of edge probabilities and call it $\hat{D}_N(\lambda_k)$
- Select the lambda where the sum of variances $\hat{D}_N(\lambda_k)$ is below a certain threshold β.

Matrix of edge probabilities





Stability-based model selection with StARS

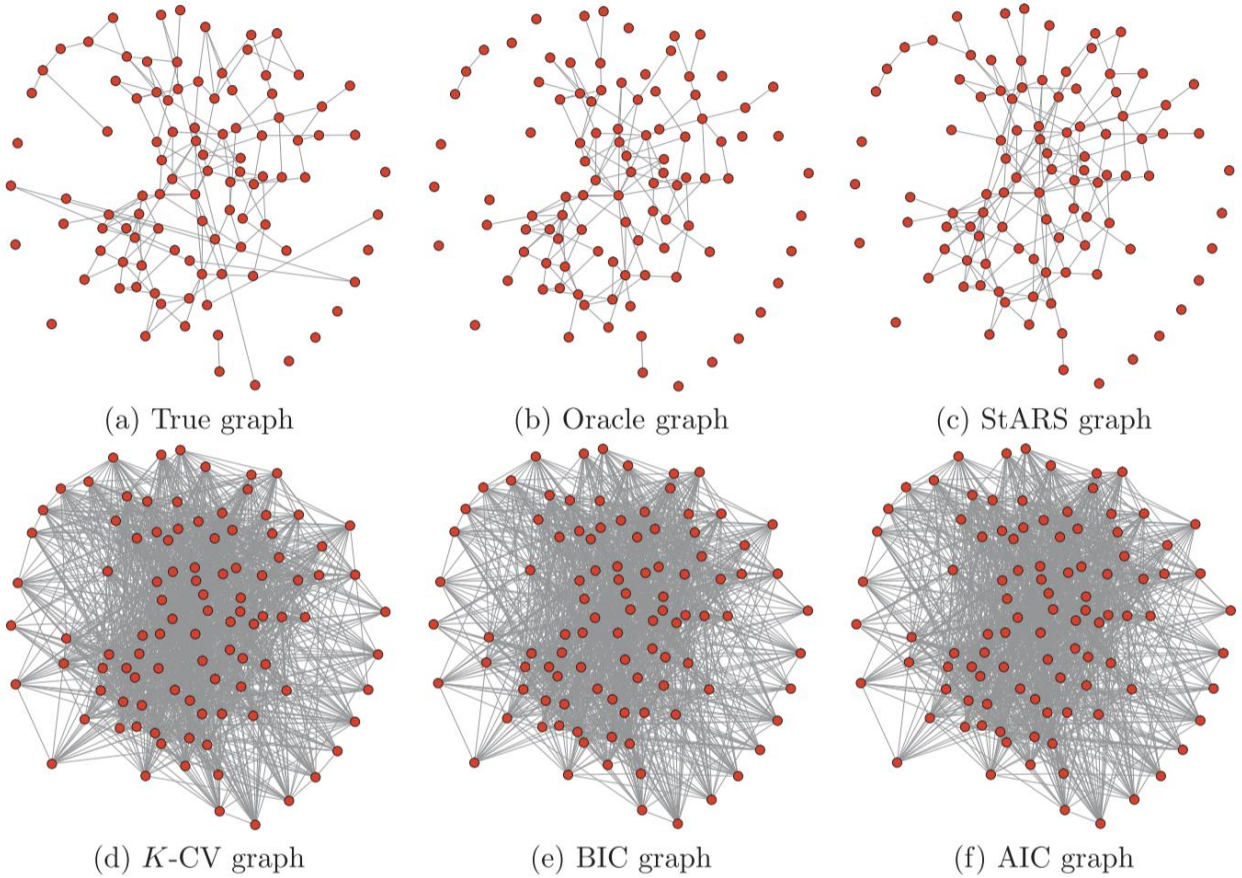
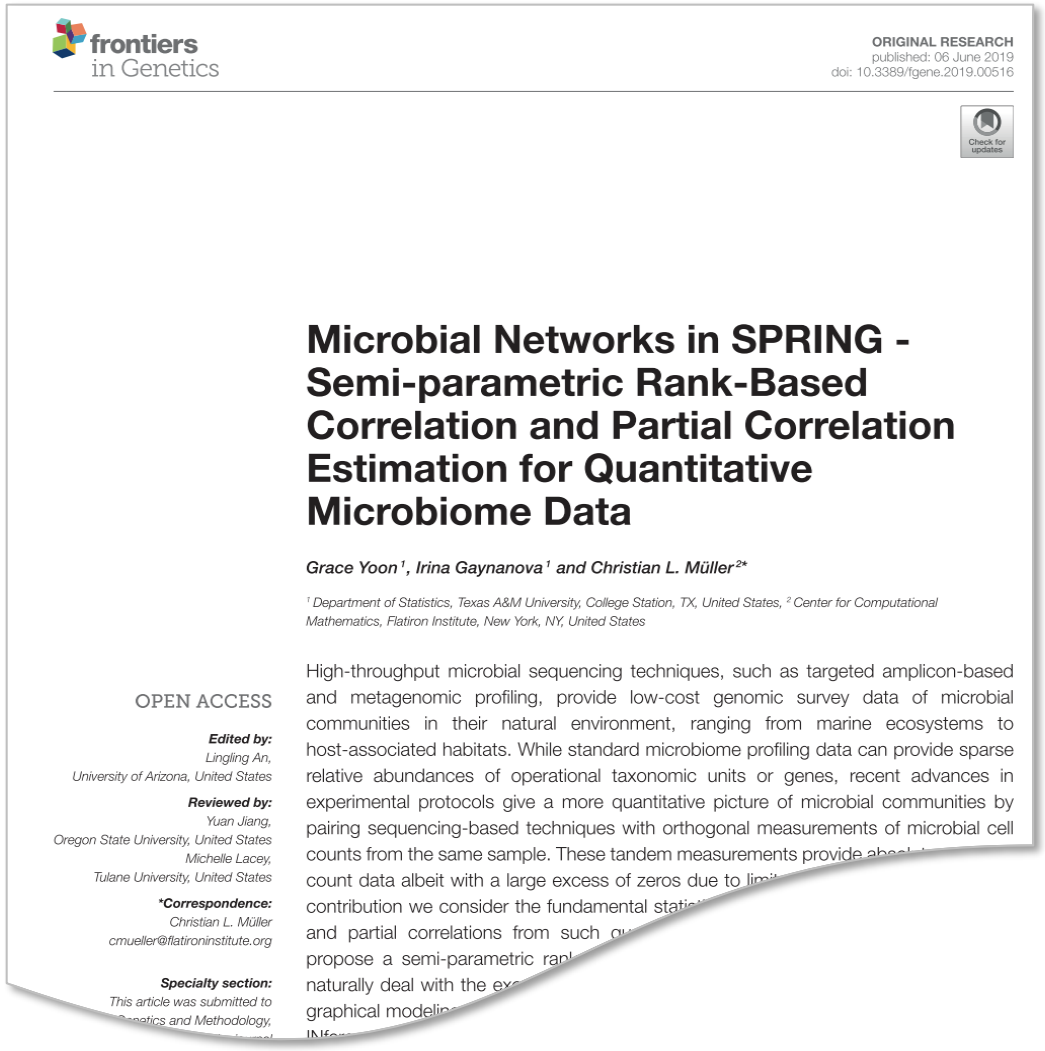


FIG 1. *Comparison of different methods on the data from the neighborhood graphs* ($n = 400, p = 100$).

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010. https://doi.org/10.48550/arxiv.1006.3316

SPRING – An alternative to SPIEC-EASl …

… also applicable to absolute microbial count data.



frontiers in Genetics

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fgene.2019.00516

Check for updates

Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data

Grace Yoon[1], *Irina Gaynanova*[1] *and Christian L. Müller*[2]*

[1] *Department of Statistics, Texas A&M University, College Station, TX, United States,* [2] *Center for Computational Mathematics, Flatiron Institute, New York, NY, United States*

High-throughput microbial sequencing techniques, such as targeted amplicon-based and metagenomic profiling, provide low-cost genomic survey data of microbial communities in their natural environment, ranging from marine ecosystems to host-associated habitats. While standard microbiome profiling data can provide sparse relative abundances of operational taxonomic units or genes, recent advances in experimental protocols give a more quantitative picture of microbial communities by pairing sequencing-based techniques with orthogonal measurements of microbial cell counts from the same sample. These tandem measurements provide abso
count data albeit with a large excess of zeros due to limi
contribution we consider the fundamental statist
and partial correlations from such qu
propose a semi-parametric ran
naturally deal with the exc
graphical modelin
Infer

OPEN ACCESS

Edited by:
Lingling An,
University of Arizona, United States

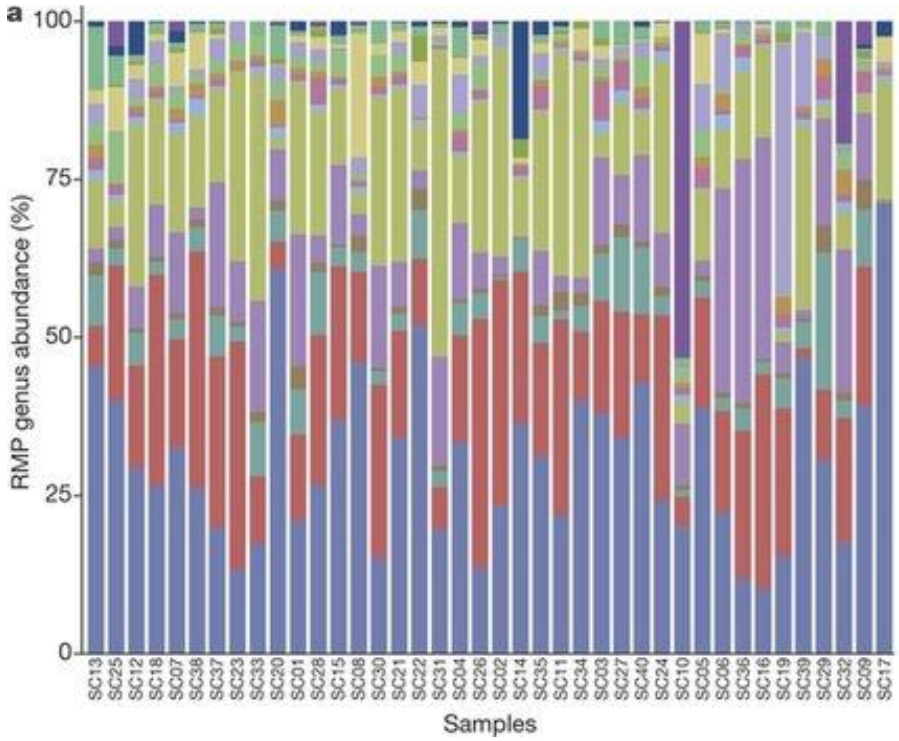
Reviewed by:
Yuan Jiang,
Oregon State University, United States
Michelle Lacey,
Tulane University, United States

***Correspondence:**
Christian L. Müller
cmueller@flatironinstitute.org

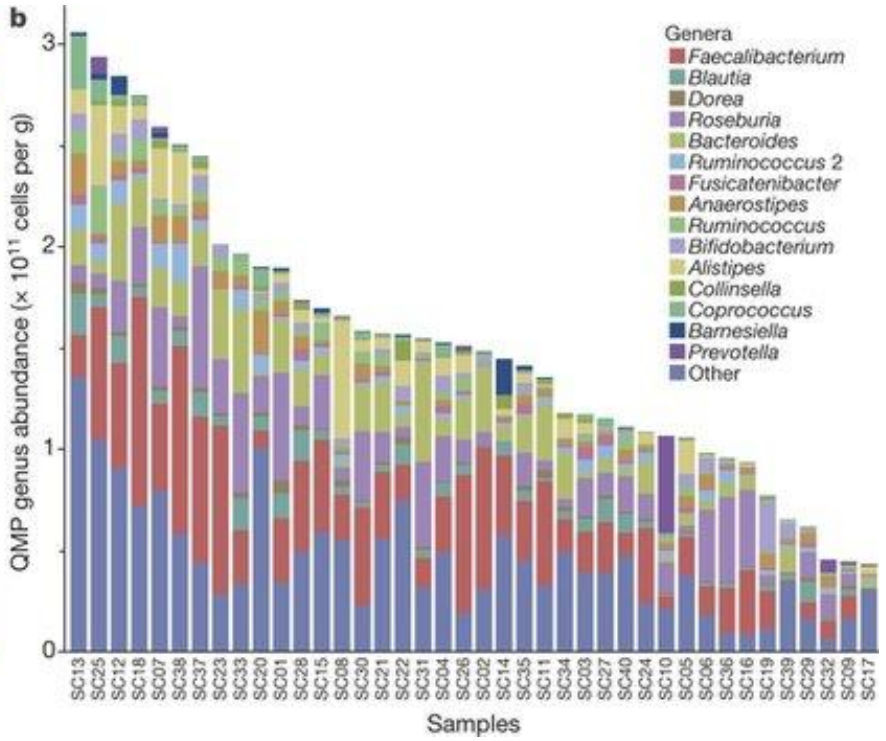
Specialty section:
This article was submitted to
Genetics and Methodology,

Recap: Quantitative microbiome data

Relative abundance

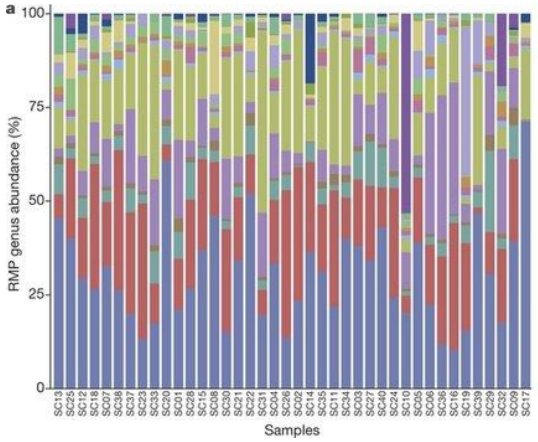


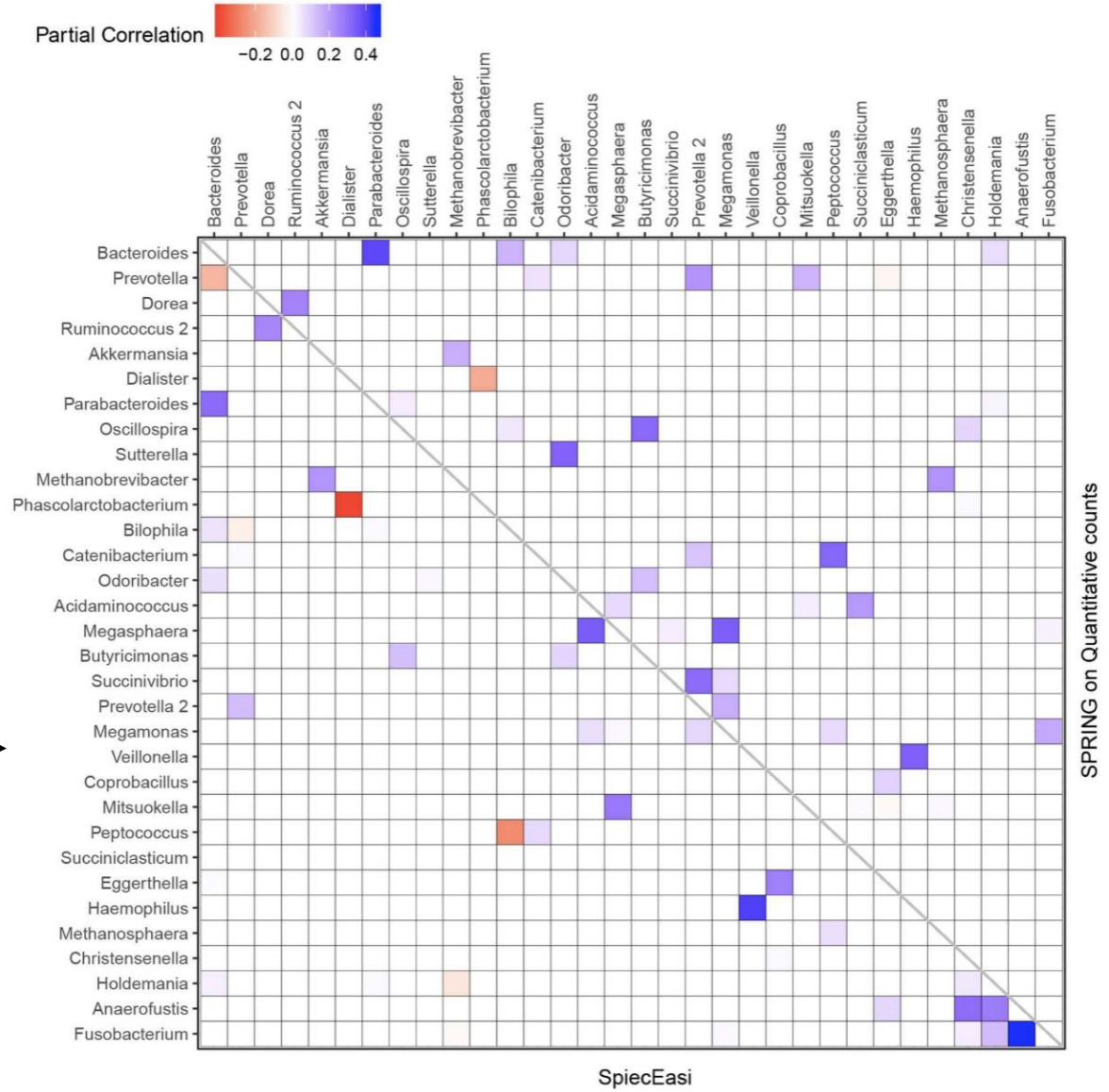
Absolute abundance



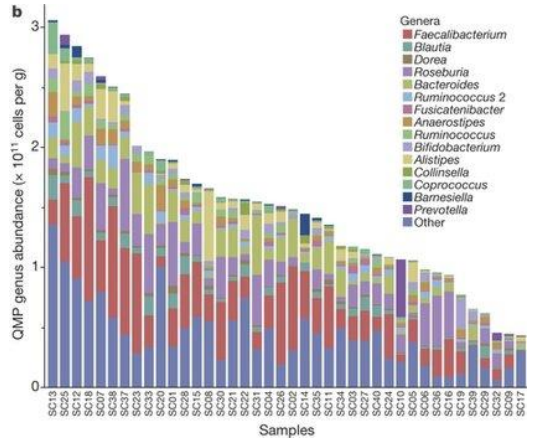
Absolute abundances can be modelled with SPRING

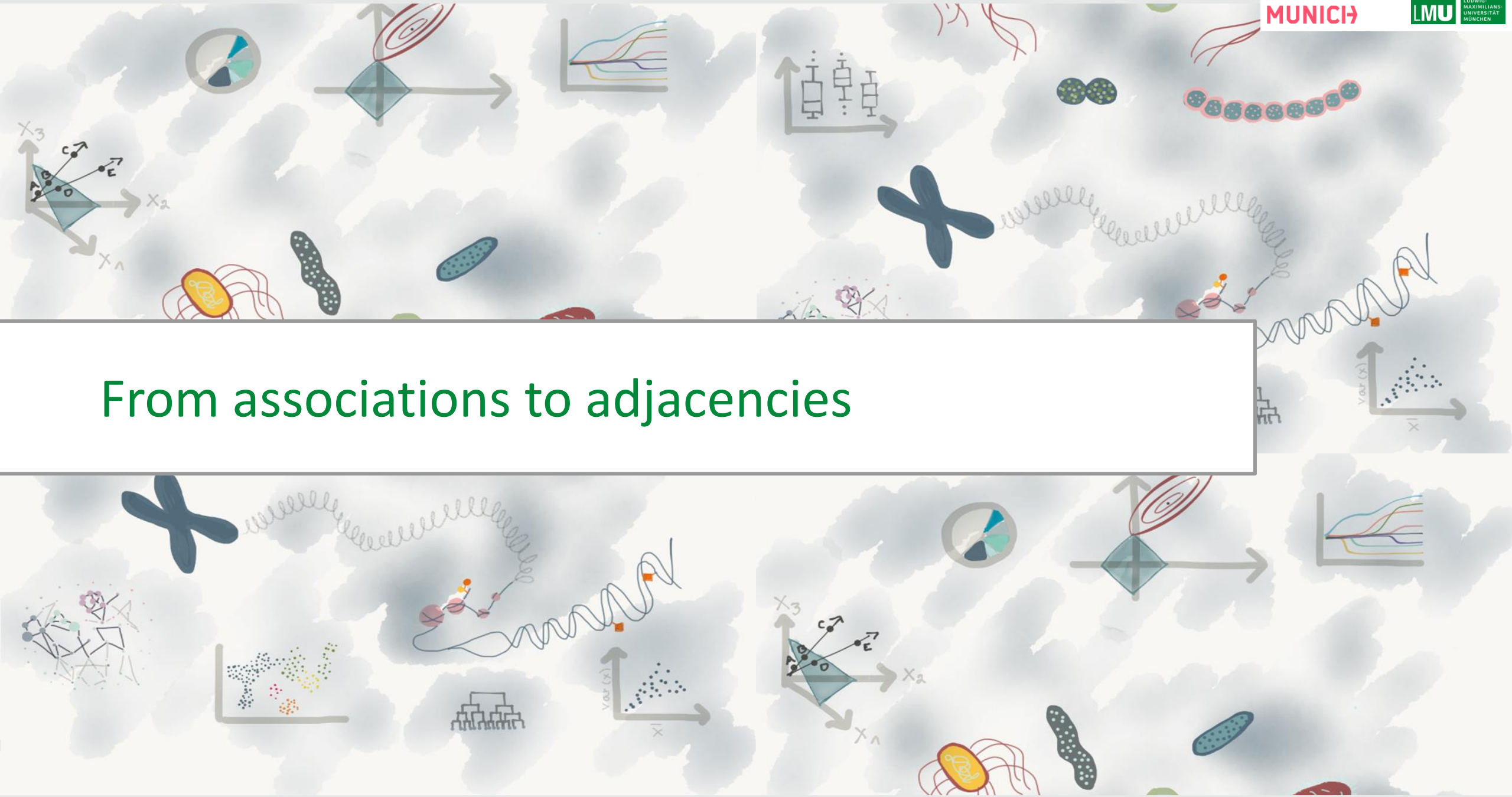
Relative abundance





Absolute abundance

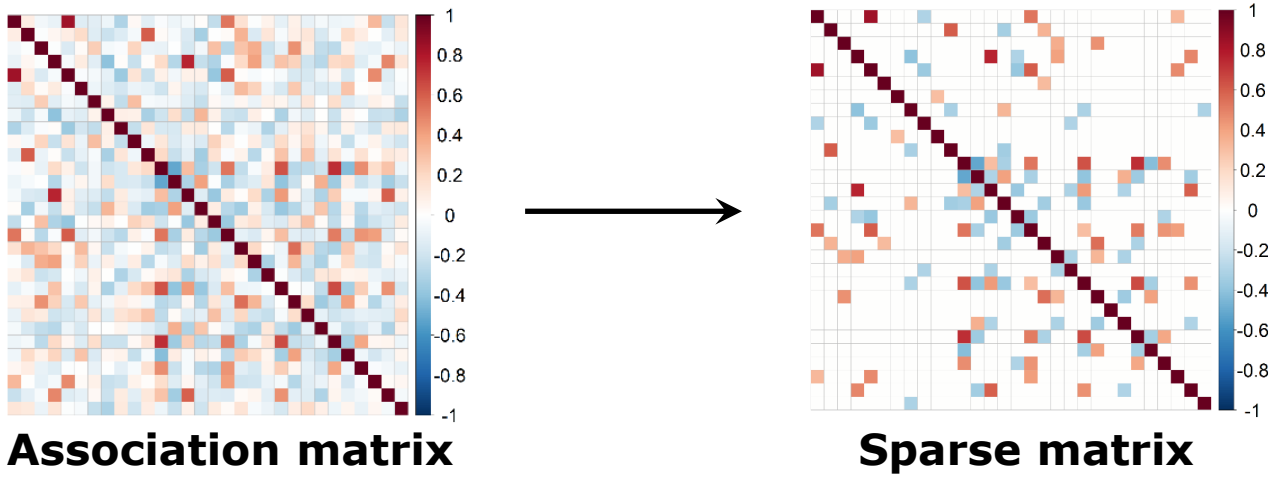




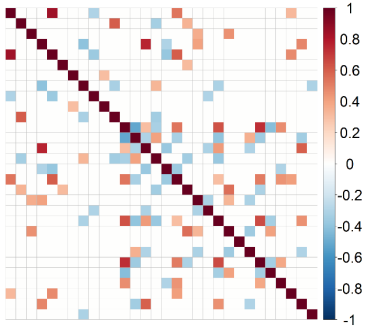
From associations to adjacencies

Further sparsification methods

- Thresholding
- Statistical tests → Applicable to other association measures, e.g., correlation
  - Student‘s t-test
  - Bootstrapping

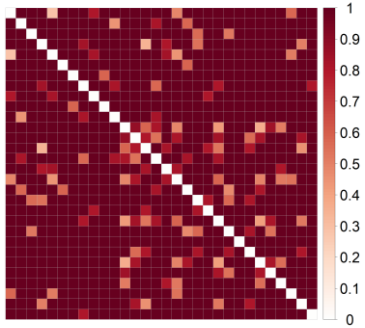


Inferring the adjacency matrix



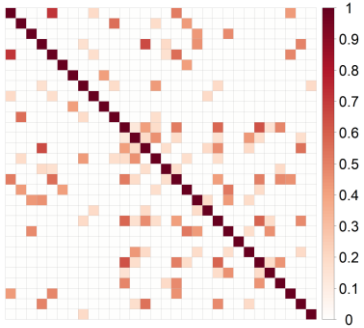
Sparse association matrix with entries r_{ij} in [-1,1]

$d_{ij} = ?$



Dissimilarity matrix (values in [0,1])

$s_{ij} = 1 - d_{ij}$



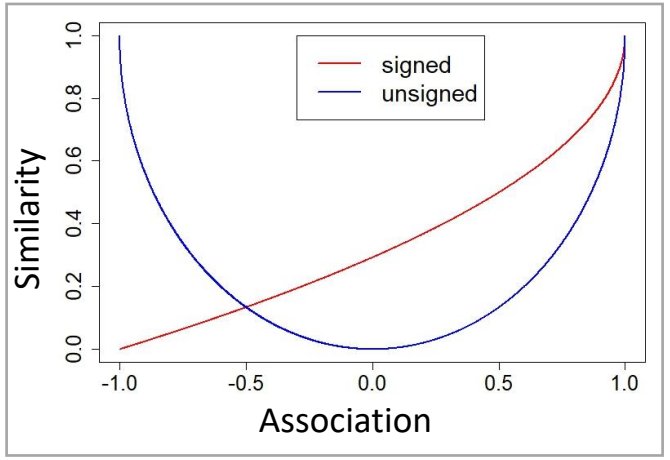
Similarity matrix

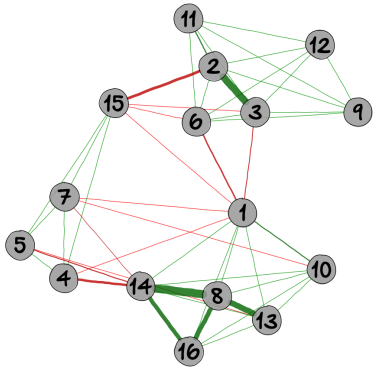
Adjacency matrix

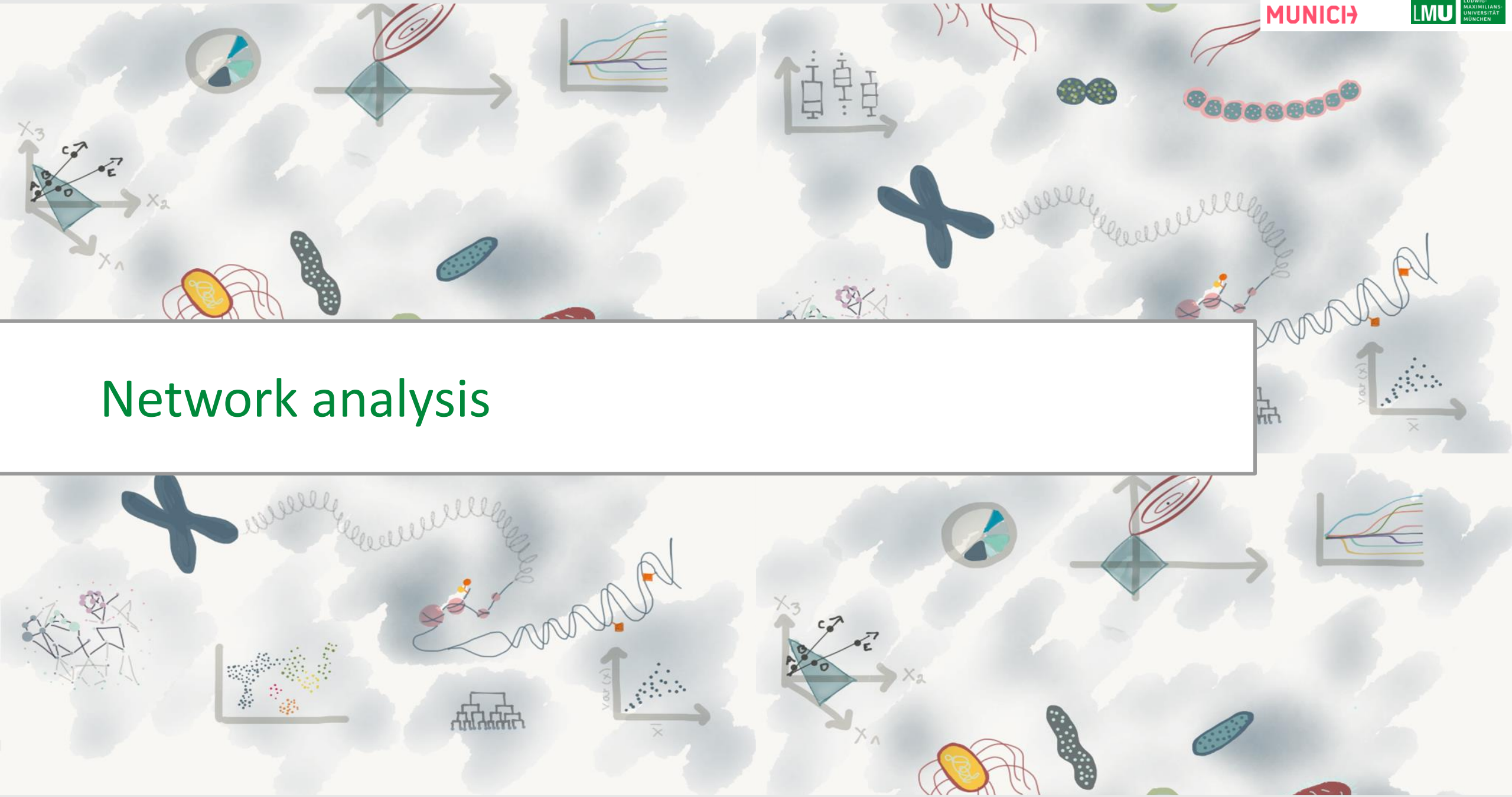
- "unsigned": $d_{ij} = \sqrt{1 - r_{ij}^2}$

  → low distance between strongly associated taxa (positively as well as negatively)
- "signed": $d_{ij} = \sqrt{0.5(1 - r_{ij})}$

  → distance is high for strongly negative associated taxa





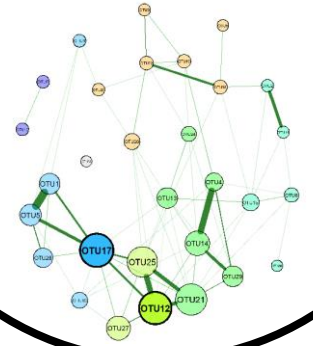


Network analysis

Typical network analysis workflow

Network construction

Network analysis

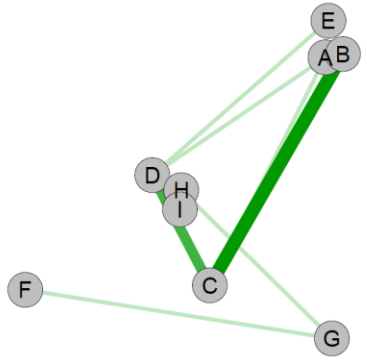


Network comparison

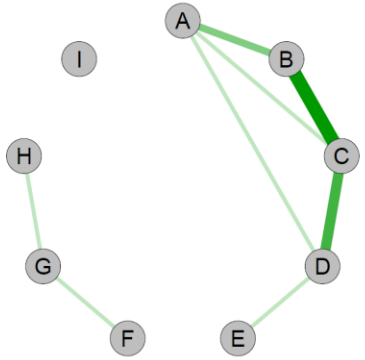
Differential network

Layout

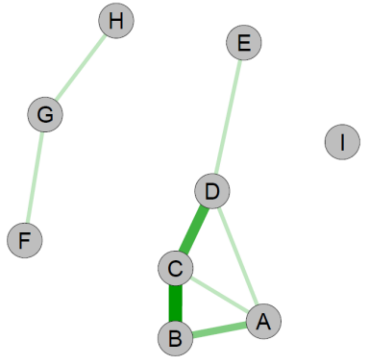
How to place nodes in the two-dimensional space?



random layout



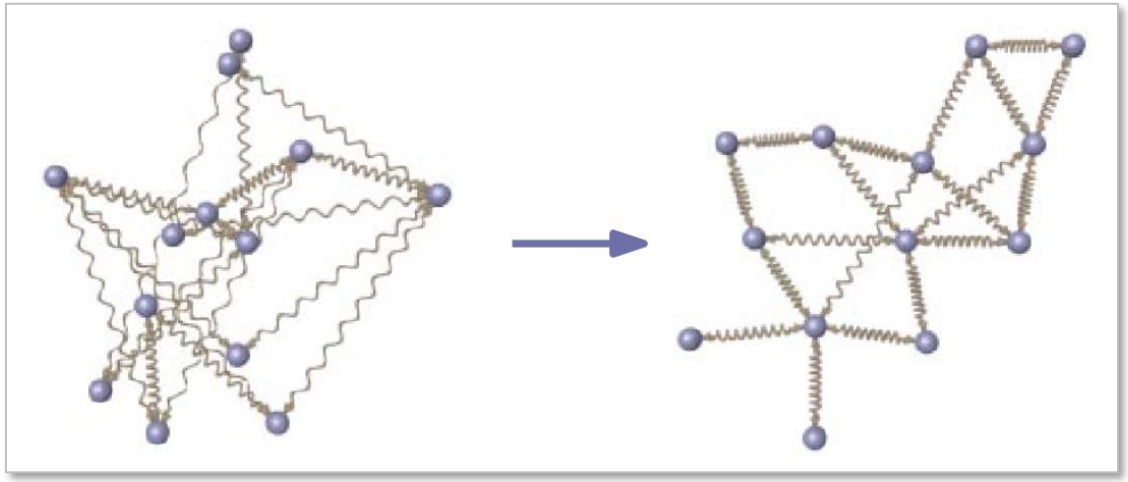
circular layout



force directed layout

Force-directed layout algorithm:

- Based on physical concepts: mechanical springs or electric repulsion
- E.g., Fruchterman-Reingold Layout



Network analysis

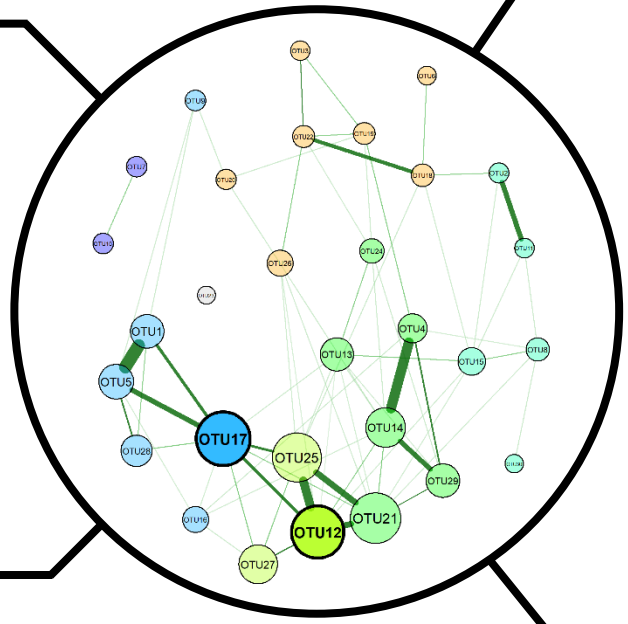
Centrality measures

Degree

Betweenness centrality

Closeness centrality

Eigenvector centrality



Clustering

Hierarchical clustering

Modularity Clustering

Fast greedy modularity optimization

Clustering based on edge betweenness

Global network properties

Modularity

Global clustering coefficient

Average path length

Average dissimilarity

Positive edge percentage

Edge density

Natural connectivity

Edge / vertex connectivity

Number of components

Hubs

Most central nodes (regarding one or more centrality values)