Network reconstruction from count data

Raphaëlle Momal
Supervision: S. Robin[1] and C. Ambroise[2]

[1]UMR AgroParisTech / INRA MIA-Paris
[2]LaMME, Evry

February 1rst, 2019

Context

Rising interest in jointly analysed species abundances:

- Metagenomics
- Microbiology
- Ecology

Ecological network

Tool to better understand species interactions (direct/indirect), eco-systems organizations (hubs?)

Allows for resilience analyses, pathogens control, ecosystem comparison, response prediction...

Example

Data:

- Species: bacteria, fungi...
- Abundances: read counts from Next-Generation Sequencing technologies (metabarcoding) $\Rightarrow$ $n \times p$ matrix Y
- Covariates: temperature, water depth... $\Rightarrow$ $n \times d$ matrix X
- Offsets: species-specific, sample-specific $\Rightarrow$ $p \times p$ matrix O

Goal:

Infer the species interaction network $\widehat{G}$ from count data Y, accounting for X and O :

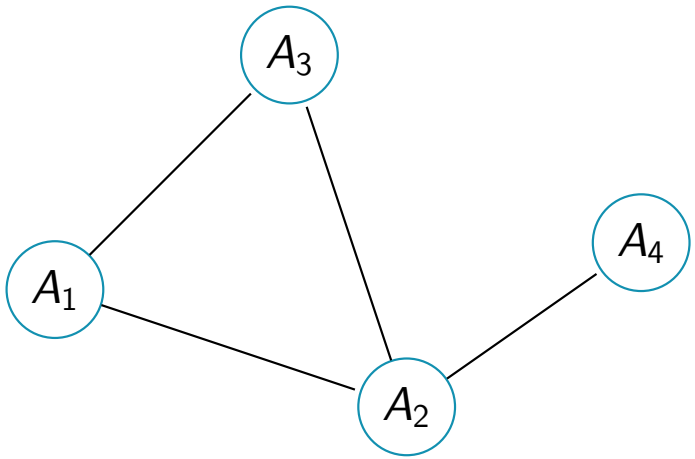
$$\widehat{G} = f(Y, X, O)$$

Challenges

- Statistical network inference
- Count data
- Offsets and covariates

Graphical models: a statistical framework for network inference

Example:

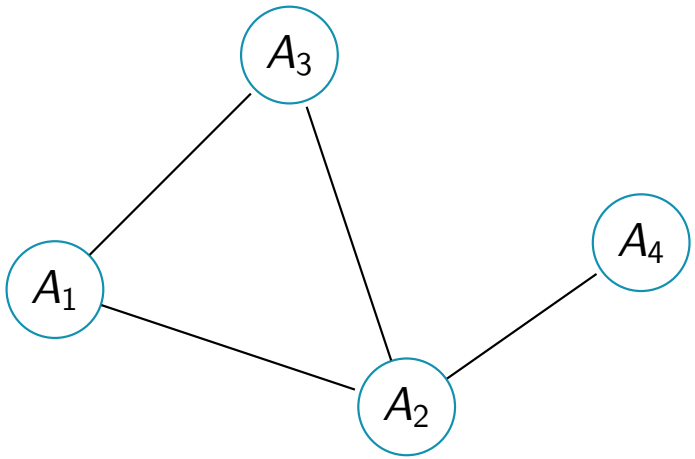


- Connected: all variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

  A_4 is independent from (A_1, A_3) conditionally on A_2

Graphical models: a statistical framework for network inference

Example:



- Connected: all variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

  A_4 is independent from (A_1, A_3) conditionally on A_2

$$P(A_1, \ldots, A_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(A_C)$$

PLN model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

$$\left.\begin{array}{rl} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(0, \Sigma)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)

PLN model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

$$\left.\begin{array}{ll} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets
- Variational estimation algorithm (Chiquet et al., 2017)

PLN model + Graphical model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

$$\left.\begin{array}{rl} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma_G) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma_G)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets
- Variational estimation algorithm (Chiquet et al., 2017)

Proposed method: PLN + Spanning trees

Tree structure on PLN latent layer

EMtree model

$$\left.\begin{array}{rl} T & \sim \prod_{kl} \beta_{kl} / B \\ Z_i | T \ iid & \sim \mathcal{N}_d(0, \Sigma_T) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i | T \\ Y_{ij} | Z_{ij}, T & \sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma_T)$$

Why Spanning trees

Sparse structures:

$$\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}} \text{ reduced to } \#\mathcal{T}_p = p^{(p-2)}$$

Why Spanning trees

Sparse structures:

$$\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}} \text{ reduced to } \#\mathcal{T}_p = p^{(p-2)}$$

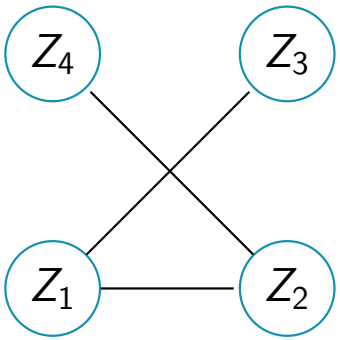
Suitable algebraic tool:

Matrix tree theorem (Chaiken and Kleitman, 1978)

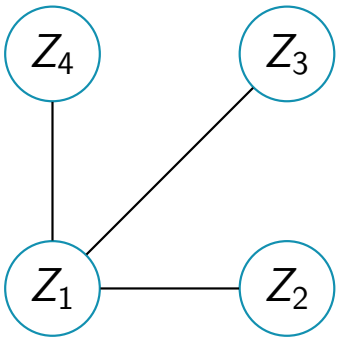
$$\sum_{T \in \mathcal{T}} \prod_{(k,l) \in T} \psi_{k,l}(Y) = \det(L_{\psi(Y)}) \to \Theta(p^3)$$

Approach: infer the network by averaging spanning trees

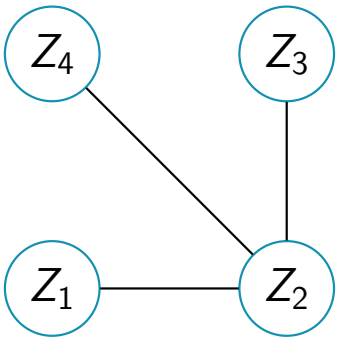
Tree averaging



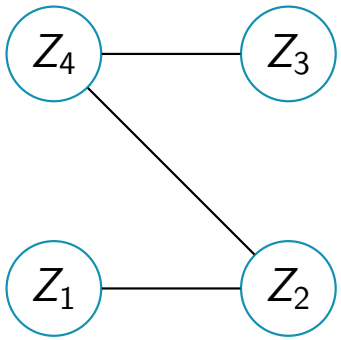
$P\{T = T_1|Z\}$



$P\{T = T_2|Z\}$



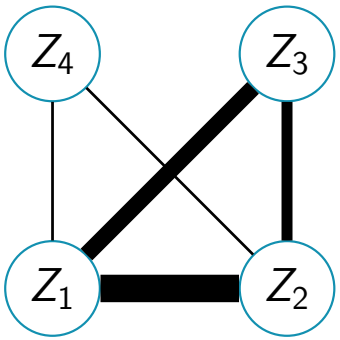
$P\{T = T_3|Z\}$



$P\{T = T_4|Z\}$

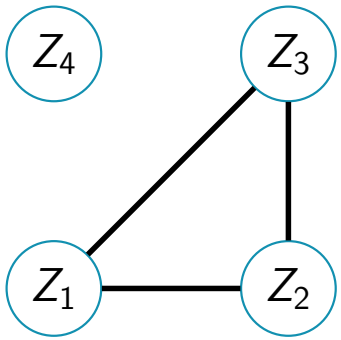
...

Compute edge probabilities:



$P\{(j,k) \in T|Z\}$

Thresholding probabilities:



$P\{(j,k) \in T|Z\}$

EMtree algorithm

Input: Abundance data, covariates, offsets

1rst step: VEM algorithm to fit PLN model $\Rightarrow \hat{\theta}$, $\hat{E}[Z|Y]$, $\hat{E}[Z^2|Y]$

2nd step: EM algorithm to update the $\beta_{jk} \Rightarrow$ conditional probabilities for all edges

EMtree algorithm

Input: Abundance data, covariates, offsets

1rst step: VEM algorithm to fit PLN model $\Rightarrow \hat{\theta}$, $\hat{E}[Z|Y]$, $\hat{E}[Z^2|Y]$

2nd step: EM algorithm to update the $\beta_{jk} \Rightarrow$ conditional probabilities for all edges

Thresholding: Select edges with probability above the probability of edges in a tree drawn uniformly ($2/p$)

Resampling: Strengthen the results: only edges selected in more than 80% of S sub-samples are kept.

Available for download at `https://github.com/Rmomal/EMtree`

Evaluation strategy

Alternatives:

Two methods on transformed counts, no covariates:

- SpiecEasi algorithm Kurtz et al. (2015)
- gCoda Fang et al. (2017)

One taking raw counts and covariates:

- MInt Biswas et al. (2016) (uses PLN model)

Evaluation strategy

Alternatives:

Two methods on transformed counts, no covariates:

- SpiecEasi algorithm Kurtz et al. (2015)
- gCoda Fang et al. (2017)

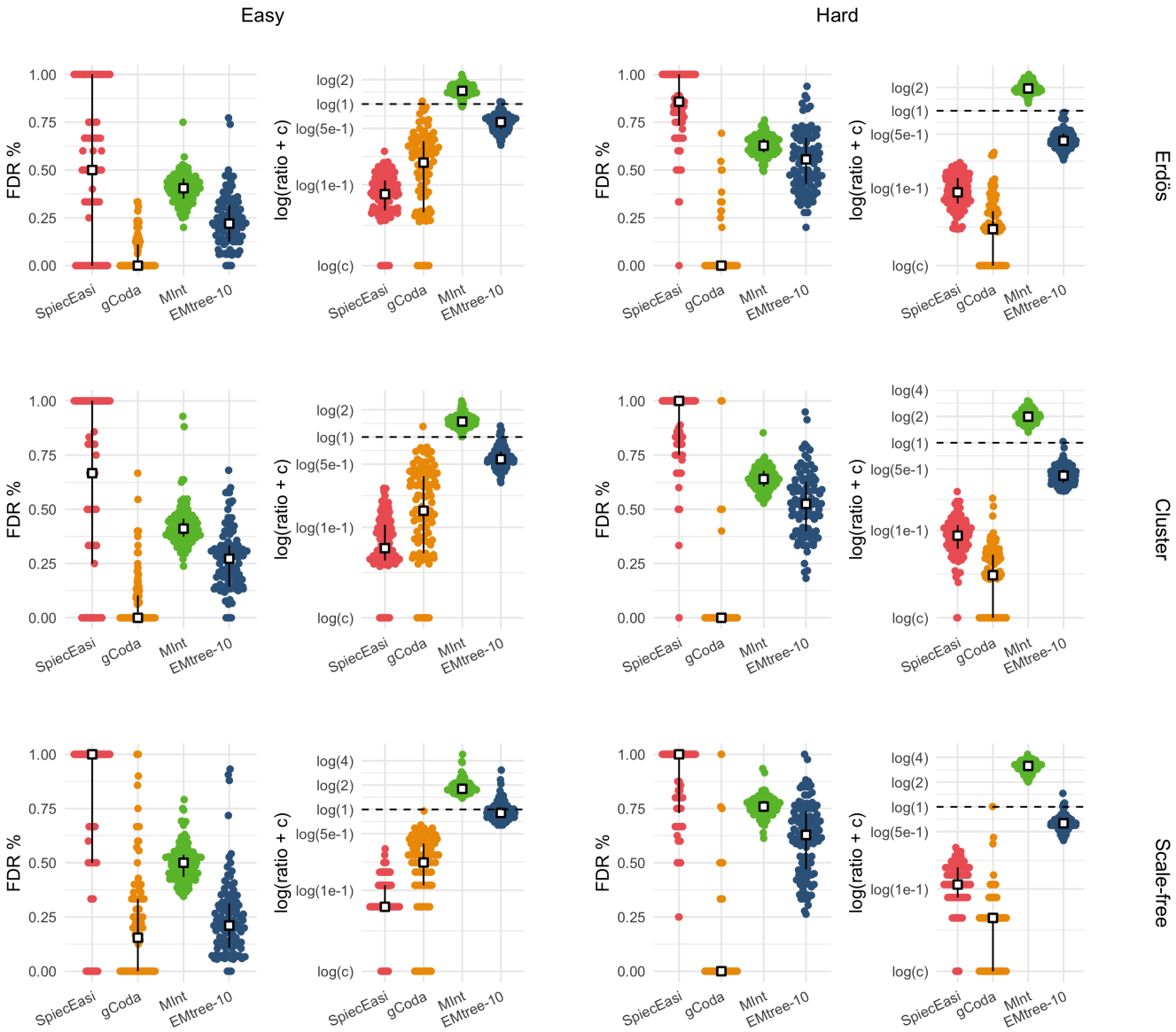
One taking raw counts and covariates:

- MInt Biswas et al. (2016) (uses PLN model)

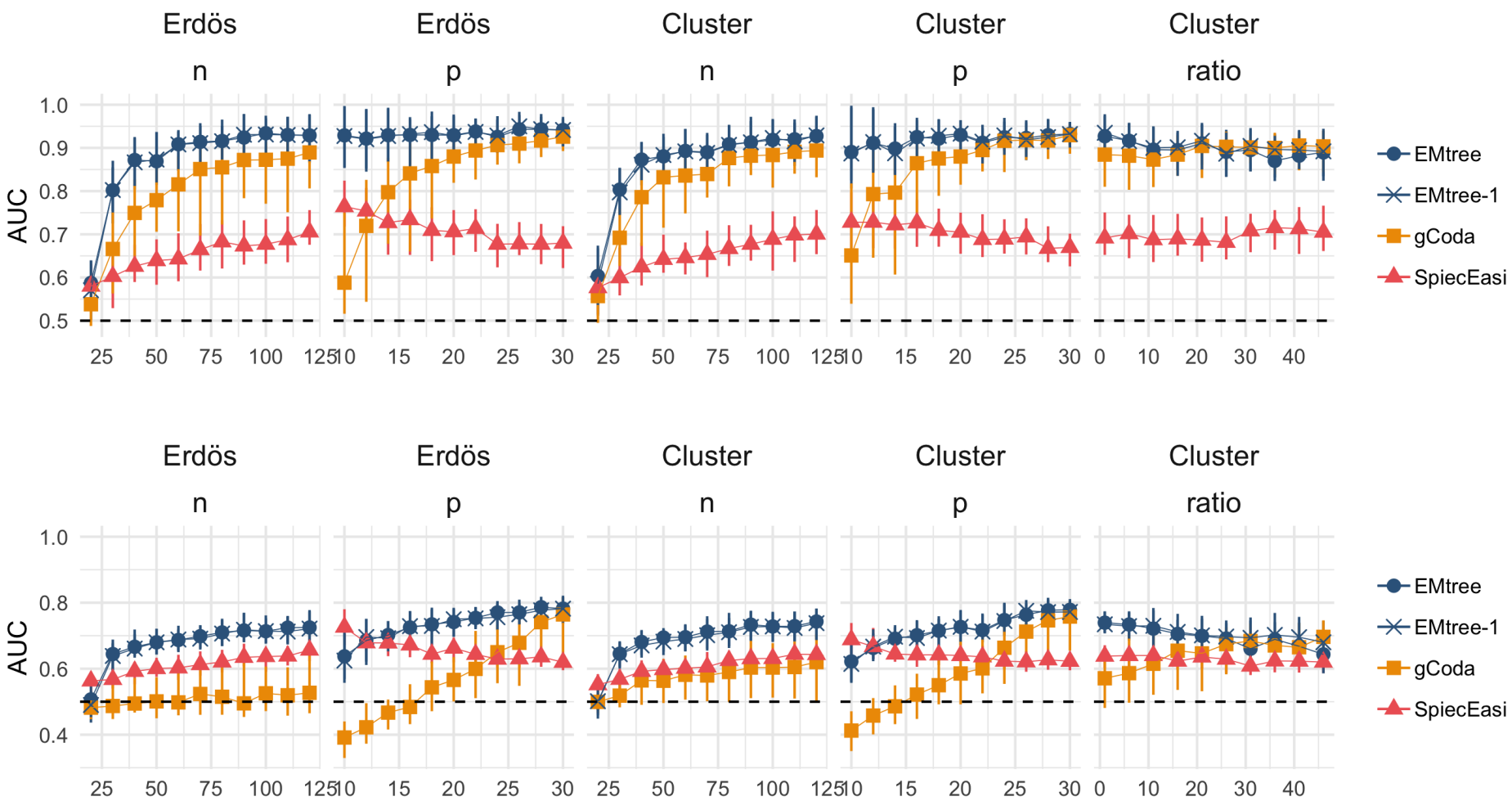
Simulation design:

1. Choose G and define Σ_G accordingly
2. Sample count data Y from $\mathcal{PLN}(X, \Sigma_G)$
3. Infer the network with EMtree, SpiecEasi, gCoda, and MInt
4. Compare results with presence/absence of edges (FDR, AUC)

Difficulty level

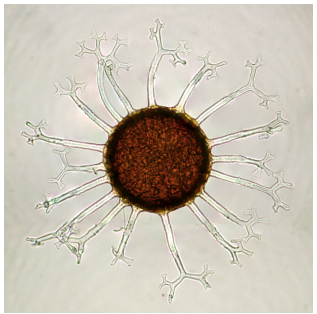


Effects of density



Effect of Erdös and Cluster structures on the evolutions of AUC median and inter-quartile intervals for parameters n, p and *ratio*. *Top*: densities set to $2/p$, *bottom*: densities set to $5/p$.

Oak Mildew



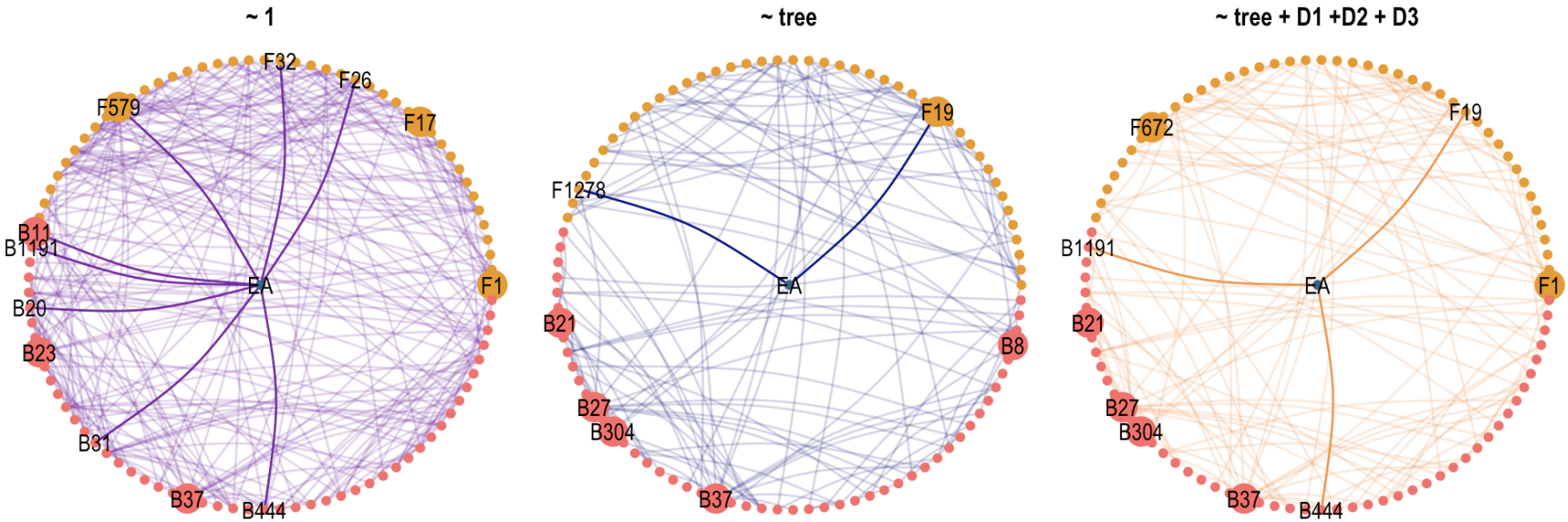
Pathogen Erysiphe alphitoides (EA).

Oak leaf with powdery mildew.

Metabarcoding of oak tree leaves microbiome (Jakuschkin et al., 2016).

- 114 sample of 94 bacterial/fungal-OTUs
- Different read depth for bacteria and fungi
- covariates: tree status; distance to ground, to trunk and to base of the branch.

Inferred networks



Conclusion

Contributions:

- Formal probabilistic model for network inference with count data
- Inclusion of offsets and covariates
- Variational estimation algorithm

Perspectives:

- Network comparison
- Missing major actor (species/covariates)
- Model for the inference in the observed counts layer

Acknowledgments

Special thanks :

Supervisors Stéphane Robin, Christophe Ambroise

PLN team Julien Chiquet (MIA-Paris), Mahendra Mariadassou (INRA Jouy)

Data Corinne Vacher (INRA Bordeaux)

Contact :

email raphaelle.momal@agroparistech.fr

Web Rmomal.github.io

References I

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653.

Biswas, S., McDonald, M., Lundberg, D. S., Dangl, J. L., and Jojic, V. (2016). Learning microbial interaction networks from metagenomic count data. *Journal of Computational Biology*, 23(6):526–535.

Chaiken, S. and Kleitman, D. J. (1978). Matrix tree theorems. *Journal of combinatorial theory, Series A*, 24(3):377–381.

Chiquet, J., Mariadassou, M., and Robin, S. (2017). Variational inference for probabilistic Poisson PCA. Technical report, arXiv:1703.06633. to appear in *Annals of Applied Statistics*.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2017). gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708.

Jakuschkin, B., Fievet, V., Schwaller, L., Fort, T., Robin, C., and Vacher, C. (2016). Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen erysiphe alphitoides. *Microb Ecol*, 72(4):870–880.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226.