Topological analysis of an inferred network

S. Founas, S. Donnet, S. Robin

MIA-Paris, AgroParisTech / INRA / univ. Paris Saclay

NetBio, Oct. 2019, IPS2, Orsay

Introduction

Network analysis. Two distinct statistical problems

- Network inference: species/genes interactions are not observed but reconstructed based on abundance/expression data
  → graphical lasso, tree-based inference, GeneNet, ...

- Network topology: the interaction network is observed and one aim at understanding its organization
  → edge beetweenness, stochastic block model (SBM), ...

Introduction

Network analysis. Two distinct statistical problems

- Network inference: species/genes interactions are not observed but reconstructed based on abundance/expression data
  → graphical lasso, tree-based inference, GeneNet, ...

- Network topology: the interaction network is observed and one aim at understanding its organization
  → edge beetweenness, stochastic block model (SBM), ...

A common situation: Try to understand the organization of the underlying network based on abundance/expression data, i.e. data collected on the nodes only

Introduction

Network analysis. Two distinct statistical problems

- Network inference: species/genes interactions are not observed but reconstructed based on abundance/expression data
  $\rightarrow$ graphical lasso, tree-based inference, GeneNet, ...

- Network topology: the interaction network is observed and one aim at understanding its organization
  $\rightarrow$ edge beetweenness, stochastic block model (SBM), ...

A common situation: Try to understand the organization of the underlying network based on abundance/expression data, i.e. data collected on the nodes only

'Pipeline' approach:

1. Infer the network $\widehat{G}$ based on the available data
2. Analyse $\widehat{G}$ as any observed network

A 'pipeline'

Barents fish [FNA06]: $n = 89$ sites, $p = 30$ species ($+ \ d = 4$ covariates)

A 'pipeline'

Barents fish [FNA06]: $n = 89$ sites, $p = 30$ species ($+ \ d = 4$ covariates)

Abundances Y: $n \times p$

Me.ae	Ra.ra	Mi.po	Ar.at
108	0	325	0
110	0	349	0
788	0	6	0
295	0	2	0
13	2	240	0

⋮

Inferred network $\widehat{G}$: $p \times p$

SBM analysis:

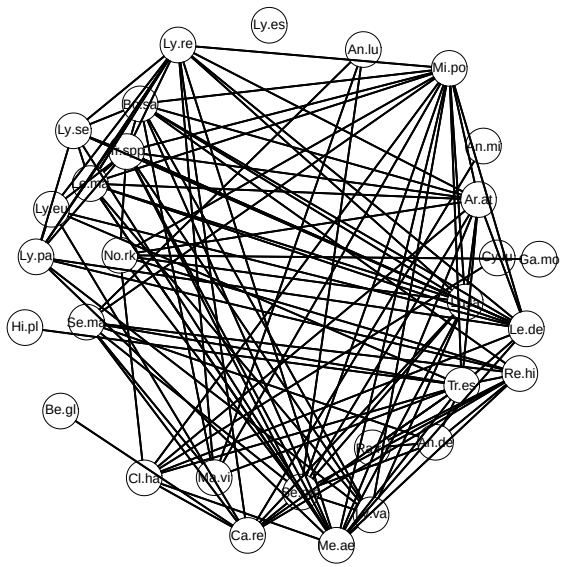
A 'pipeline'

Barents fish [FNA06]: $n = 89$ sites, $p = 30$ species ($+$ $d = 4$ covariates)

Abundances Y: $n \times p$

Me.ae	Ra.ra	Mi.po	Ar.at
108	0	325	0
110	0	349	0
788	0	6	0
295	0	2	0
13	2	240	0
⋮			

Inferred network $\widehat{G}$: $p \times p$



SBM analysis:

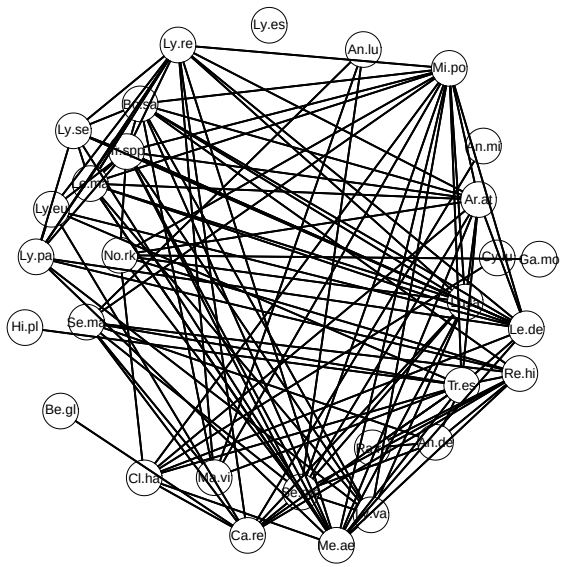
A 'pipeline'

Barents fish [FNA06]: $n = 89$ sites, $p = 30$ species ($+$ $d = 4$ covariates)

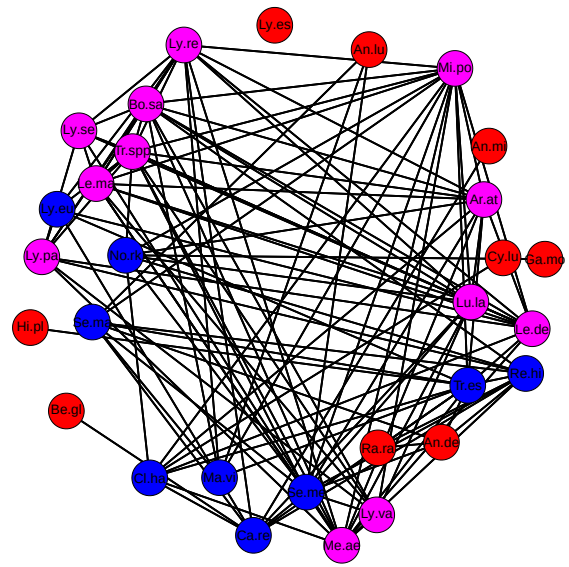
Abundances Y: $n \times p$

Me.ae	Ra.ra	Mi.po	Ar.at
108	0	325	0
110	0	349	0
788	0	6	0
295	0	2	0
13	2	240	0
⋮			

Inferred network $\widehat{G}$: $p \times p$



SBM analysis:



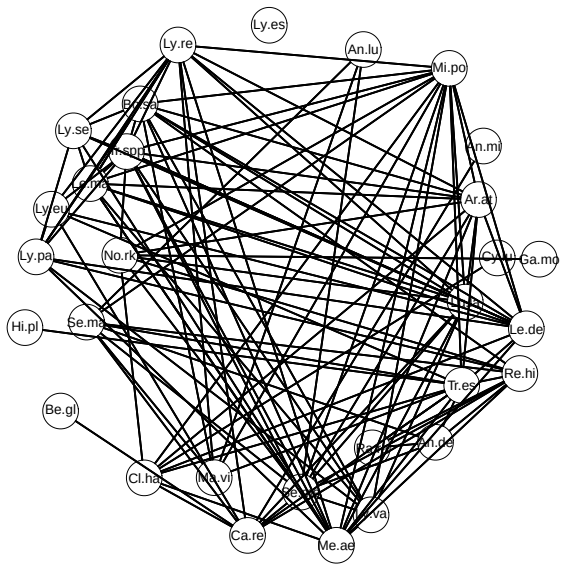
A 'pipeline'

Barents fish [FNA06]: $n = 89$ sites, $p = 30$ species ($+$ $d = 4$ covariates)

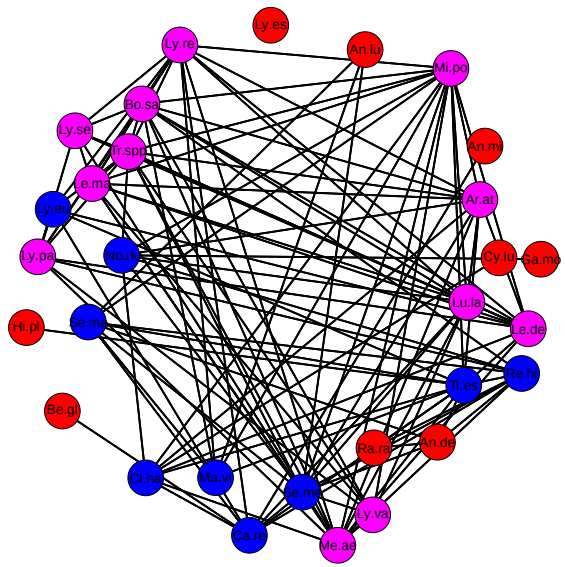
Abundances Y: $n \times p$

Me.ae	Ra.ra	Mi.po	Ar.at
108	0	325	0
110	0	349	0
788	0	6	0
295	0	2	0
13	2	240	0
⋮			

Inferred network $\widehat{G}$: $p \times p$



SBM analysis:



Problem:

- The uncertainty of network inference (step 1)
- is not accounted for in the topological analysis (step 2)

Bridging the gap

Two different definitions of 'network'.

Network inference: the species abundances (or gene expressions) are mutually dependent and the network to be inferred is the graphical model that encodes theses (conditional) dependences (e.g. GGM)

Network topology: the observed network (i.e. the set of observed interactions between the species or genes) is supposed to arise from some random graph model (e.g SBM)

Bridging the gap

Two different definitions of 'network'.

Network inference: the species abundances (or gene expressions) are mutually dependent and the network to be inferred is the graphical model that encodes theses (conditional) dependences (e.g. GGM)

Network topology: the observed network (i.e. the set of observed interactions between the species or genes) is supposed to arise from some random graph model (e.g SBM)

Here,

- The graphical model G itself is supposed to arise from some random graph model
- The observed data are supposed to arise from some joint distribution that is faithful to G

Topological analysis of an inferred network

Rational.

- The observed data are distributed according to some (undirected) graphical model (GM) G
- The GM G itself arise from som random graph model, e.g. $G \sim SBM$

Topological analysis of an inferred network

Rational.

- The observed data are distributed according to some (undirected) graphical model (GM) G
- The GM G itself arise from som random graph model, e.g. $G \sim SBM$

Aim.

- Based on the observed data, say something about the process that produced G
- Case of SBM: say something about the node memberships

Topological analysis of an inferred network

Rational.

- The observed data are distributed according to some (undirected) graphical model (GM) G
- The GM G itself arise from som random graph model, e.g. $G \sim SBM$

Aim.

- Based on the observed data, say something about the process that produced G
- Case of SBM: say something about the node memberships

Versatile approach.

- Be as agnostic as possible about the network inference method
- Just assume that the method provides a score for each edge

Edge scores

Graphical lasso [FHT08]. For Gaussian graphical models, $\Omega = \Sigma^{-1}$ = precision matrix

sparsity assumption: $$\widehat{\Omega}(\lambda) = \arg\max_{\Omega} \log p(Y; \Omega) - \lambda \|\Omega\|_{1,0}$$

inferred network: $$\widehat{G}(\lambda) = \text{support}\left(\widehat{\Omega}(\lambda)\right)$$

[#21] edge score: $$S_{jk} = \max\left\{\lambda : (j,k) \in \widehat{G}(\lambda)\right\}$$

Edge scores

Graphical lasso [FHT08]. For Gaussian graphical models, $\Omega = \Sigma^{-1}$ = precision matrix

sparsity assumption: $$\widehat{\Omega}(\lambda) = \arg\max_{\Omega} \log p(Y; \Omega) - \lambda \|\Omega\|_{1,0}$$

inferred network: $$\widehat{G}(\lambda) = \text{support}\left(\widehat{\Omega}(\lambda)\right)$$

[#21] edge score: $$S_{jk} = \max\left\{\lambda : (j,k) \in \widehat{G}(\lambda)\right\}$$

Tree-based approaches [MJ06,Kir07,SRS19,MRA19]. Random tree-shaped GM T

$$S_{jk} = P\{(j,k) \in T \mid Y\}$$

Edge scores

Graphical lasso [FHT08]. For Gaussian graphical models, $\Omega = \Sigma^{-1}$ = precision matrix

$$\text{sparsity assumption:} \qquad \widehat{\Omega}(\lambda) = \arg\max_{\Omega} \log p(Y; \Omega) - \lambda \|\Omega\|_{1,0}$$

$$\text{inferred network:} \qquad \widehat{G}(\lambda) = \text{support}\left(\widehat{\Omega}(\lambda)\right)$$

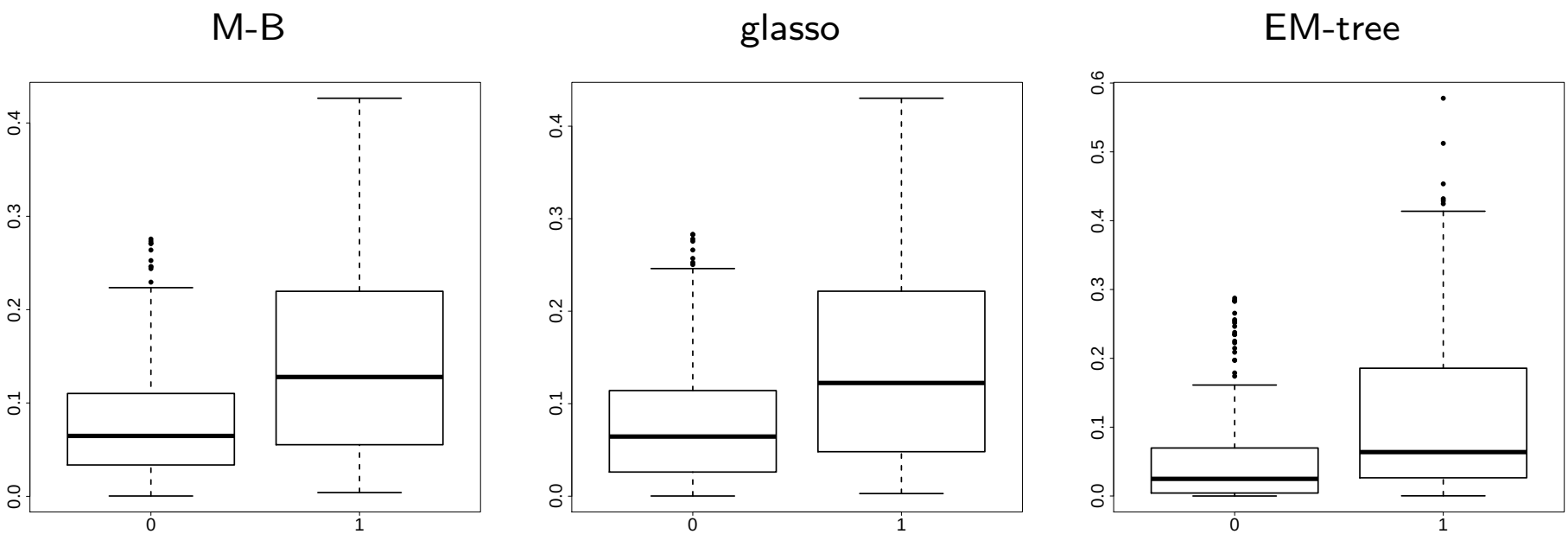
$$\text{[\#21] edge score:} \qquad S_{jk} = \max\left\{\lambda : (j,k) \in \widehat{G}(\lambda)\right\}$$

Tree-based approaches [MJ06,Kir07,SRS19,MRA19]. Random tree-shaped GM T

$$S_{jk} = P\{(j,k) \in T \mid Y\}$$

Assumption 1 (fairly reasonable). The distribution of the scores of present edges is different from the distribution of the scores of absent edges

Edge scores: synthetic data

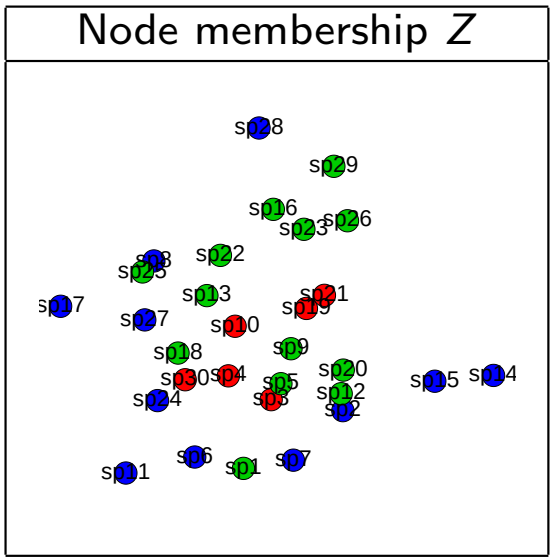


- Any reasonable method provides differentially distributed scores

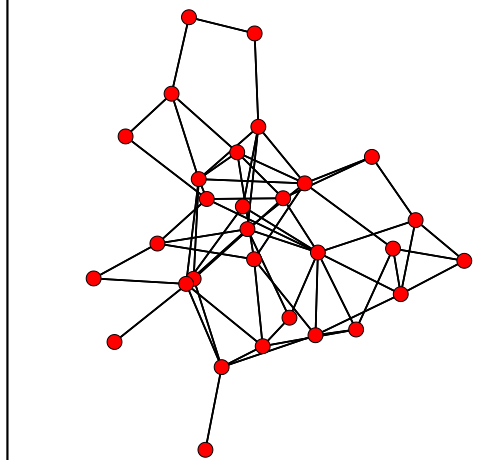
A pictural view

Conceptual (generative) model:

Node membership Z



Graphical model G



Observed data Y

	sp1	sp2	sp3	sp4	sp5
site1	0	2	8	2	0
site2	3	0	9	0	1
site3	1	5	15	0	3
site4	4	1	16	1	2
site5	1	3	104	0	4
site6	1	0	10	1	3
.					
.					
.					

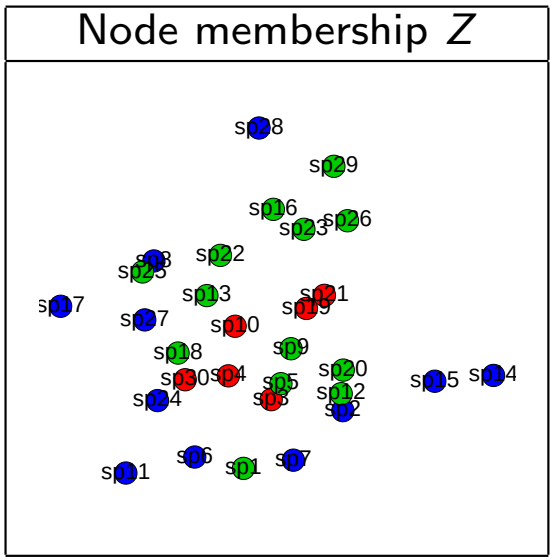
Edge scores S

	sp1	sp2	sp3	sp4	sp5
sp1	-	1.5	0.2	17.7	0.1
sp3		-	26.9	8.9	1.4
sp3			-	1.3	5.2
sp4				-	10.6
sp5					-
.					
.					
.					

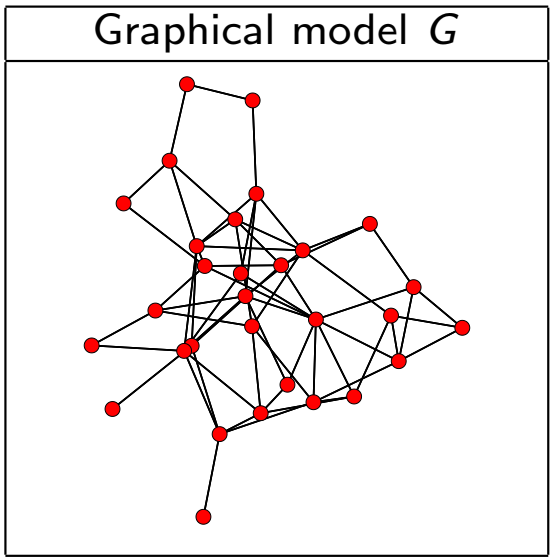
A pictural view

Pipe-line:

Node membership Z



Graphical model G



Edge scores S

	sp1	sp2	sp3	sp4	sp5
sp1	-	1.5	0.2	17.7	0.1
sp3		-	26.9	8.9	1.4
sp3			-	1.3	5.2
sp4				-	10.6
sp5					-
.					
.					
.					

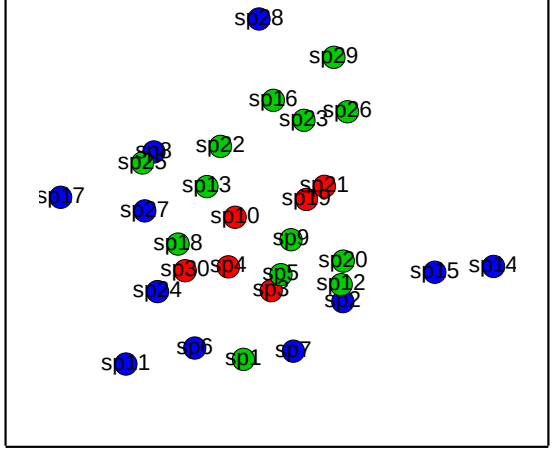
Observed data Y

	sp1	sp2	sp3	sp4	sp5
site1	0	2	8	2	0
site2	3	0	9	0	1
site3	1	5	15	0	3
site4	4	1	16	1	2
site5	1	3	104	0	4
site6	1	0	10	1	3
.					
.					
.					

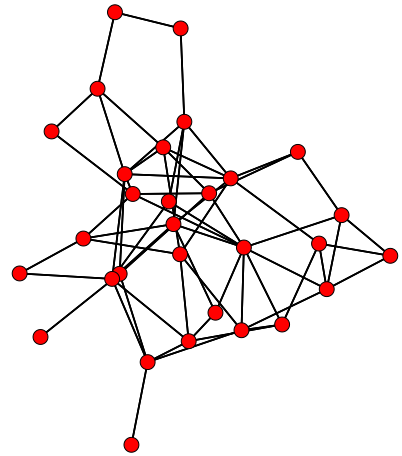
A pictural view

Actual pipe-line:

Node membership Z



Graphical model G



Edge scores S

	sp1	sp2	sp3	sp4	sp5
sp1	-	1.5	0.2	17.7	0.1
sp3		-	26.9	8.9	1.4
sp3			-	1.3	5.2
sp4				-	10.6
sp5					-
.					
.					
.					

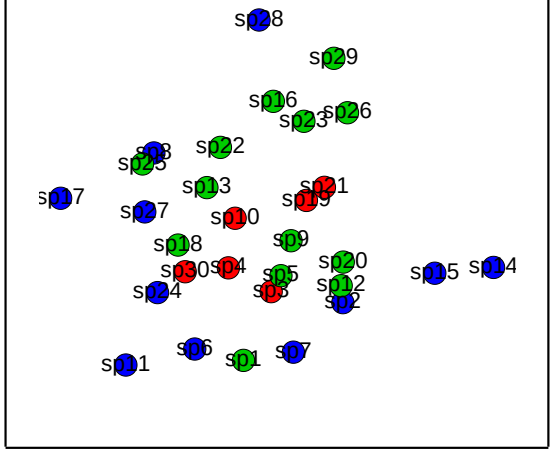
Observed data Y

	sp1	sp2	sp3	sp4	sp5
site1	0	2	8	2	0
site2	3	0	9	0	1
site3	1	5	15	0	3
site4	4	1	16	1	2
site5	1	3	104	0	4
site6	1	0	10	1	3
.					
.					
.					

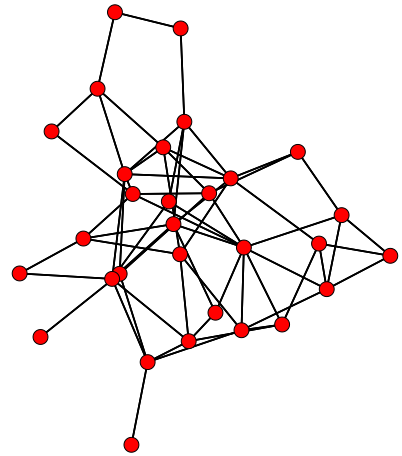
A pictural view

Our aim:

Node membership Z



Graphical model G



Edge scores S

	sp1	sp2	sp3	sp4	sp5
sp1	-	1.5	0.2	17.7	0.1
sp3		-	26.9	8.9	1.4
sp3			-	1.3	5.2
sp4				-	10.6
sp5					-
.					
.					
.					

Observed data Y

	sp1	sp2	sp3	sp4	sp5
site1	0	2	8	2	0
site2	3	0	9	0	1
site3	1	5	15	0	3
site4	4	1	16	1	2
site5	1	3	104	0	4
site6	1	0	10	1	3
.					
.					
.					

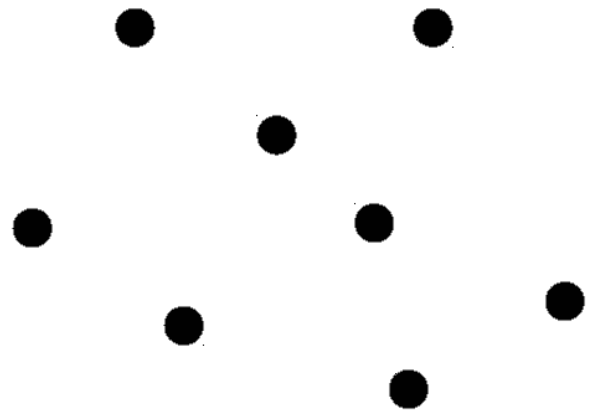
A reminder on (binary) SBM

A mixture model for random graphs. [NS01]

A reminder on (binary) SBM

A mixture model for random graphs. [NS01]

Consider p nodes ($j = 1..p$);



A reminder on (binary) SBM

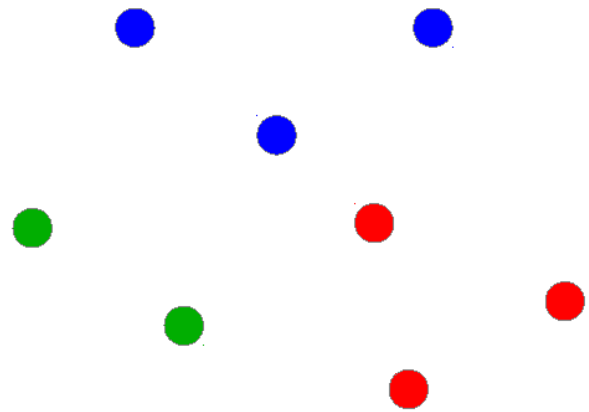
A mixture model for random graphs. [NS01]

Consider p nodes ($j = 1..p$);

Z_j = unobserved label (color)of node i:

$$\pi_q = P(Z_j = q)$$

$\pi = (\pi_1, ...\pi_Q)$;



A reminder on (binary) SBM

A mixture model for random graphs. [NS01]

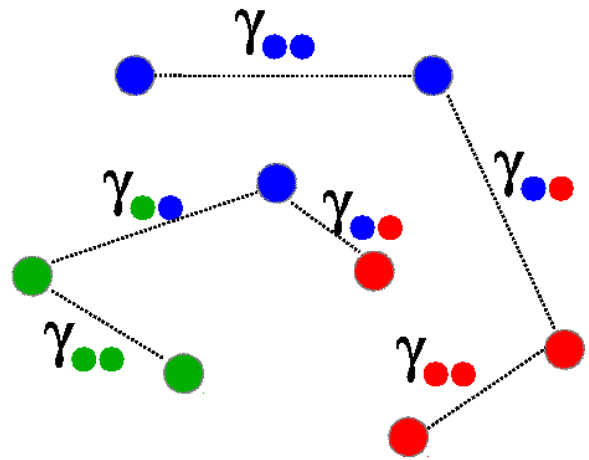
Consider p nodes ($j = 1..p$);

Z_j = unobserved label (color)of node i:

$$\pi_q = P(Z_j = q)$$

$\pi = (\pi_1, ...\pi_Q)$;
Edge G_{jk} depends on the labels:

$$P(G_{jk} = 1 \mid Z_j = q, Z_k = \ell) = \gamma_{q\ell}$$



A reminder on (binary) SBM

A mixture model for random graphs. [NS01]

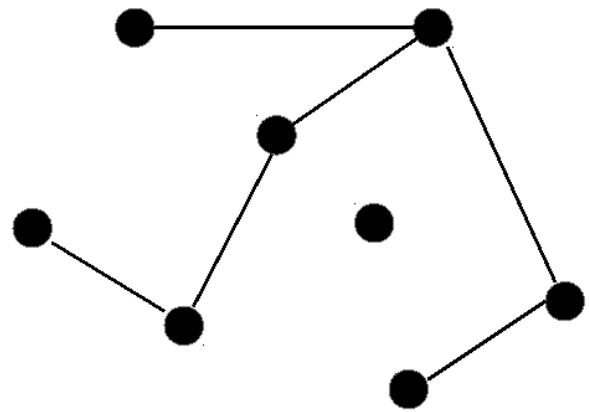
Consider p nodes ($j = 1..p$);

Z_j = unobserved label (color)of node i:

$$\pi_q = P(Z_j = q)$$

$\pi = (\pi_1, ...\pi_Q)$;
Edge G_{jk} depends on the labels:

$$P(G_{jk} = 1 \mid Z_j = q, Z_k = \ell) = \gamma_{q\ell}$$



A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q

A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q
2. Network / GM G: $P\{(j,k) \in G \mid Z_j = q, Z_k = \ell\} = \gamma_{q\ell}$

$$G \sim SBM_{\text{binary}}(p, \pi, \gamma)$$

A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q
2. Network / GM G: $P\{(j,k) \in G \mid Z_j = q, Z_k = \ell\} = \gamma_{q\ell}$

$$G \sim SBM_{\text{binary}}(p, \pi, \gamma)$$

3. Observed data: $\{Y_i\}_{i=1,\dots n}$ iid $\sim \text{GM}_G$ (e.g. $Y_i \sim GGM(\Omega_G^{-1})$)

A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q
2. Network / GM G: $P\{(j,k) \in G \mid Z_j = q, Z_k = \ell\} = \gamma_{q\ell}$

$$G \sim SBM_{\text{binary}}(p, \pi, \gamma)$$

3. Observed data: $\{Y_i\}_{i=1,\dots n}$ iid $\sim \text{GM}_G$ (e.g. $Y_i \sim GGM(\Omega_G^{-1})$)
4. Network inference: score matrix $S = [S_{jk}] =$ `some_network_inference_algorithm`(Y)

$$(S_{jk} \mid G_{jk} = 0) \sim F_0, \qquad (S_{jk} \mid G_{jk} = 1) \sim F_1$$

A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q
2. Network / GM G: $P\{(j,k) \in G \mid Z_j = q, Z_k = \ell\} = \gamma_{q\ell}$

$$G \sim SBM_{\text{binary}}(p, \pi, \gamma)$$

3. Observed data: $\{Y_i\}_{i=1,\dots n}$ iid $\sim \text{GM}_G$ (e.g. $Y_i \sim GGM(\Omega_G^{-1})$)
4. Network inference: score matrix $S = [S_{jk}] =$ `some_network_inference_algorithm`(Y)

$$(S_{jk} \mid G_{jk} = 0) \sim F_0, \qquad (S_{jk} \mid G_{jk} = 1) \sim F_1$$

Assumption 2 (more questionable). The scores (S_{jk}) are independent conditionally on the edge's existence (G_{jk}).

A special instance of SBM

Model. $j, k = 1, \dots p$ nodes = species = genes, Q clusters

1. Node membership Z_j: each node belongs to cluster q with probability π_q
2. Network / GM G: $P\{(j,k) \in G \mid Z_j = q, Z_k = \ell\} = \gamma_{q\ell}$

$$G \sim SBM_{\text{binary}}(p, \pi, \gamma)$$

3. Observed data: $\{Y_i\}_{i=1,\dots n}$ iid $\sim \text{GM}_G$ (e.g. $Y_i \sim GGM(\Omega_G^{-1})$)
4. Network inference: score matrix $S = [S_{jk}] =$ `some_network_inference_algorithm`(Y)

$$(S_{jk} \mid G_{jk} = 0) \sim F_0, \qquad (S_{jk} \mid G_{jk} = 1) \sim F_1$$

Assumption 2 (more questionable). The scores (S_{jk}) are independent conditionally on the edge's existence (G_{jk}).

A mixture distribution for the edge scores:

$$(S_{jk} \mid Z_j = q, Z_k = \ell) \sim (1 - \gamma_{q\ell})F_0 + \gamma_{q\ell}F_1$$

Inference

Aim. Infer the parameter $\theta = (\pi, \gamma, F_0, F_1)$, the node memberships $Z = (Z_j)$ and the graph $G = (G_{jk})$.

Inference

Aim. Infer the parameter $\theta = (\pi, \gamma, F_0, F_1)$, the node memberships $Z = (Z_j)$ and the graph $G = (G_{jk})$.

Incomplete data model.

- Neither the node memberships Z nor the underlying graph G are observed.
- The EM algorithm requires to evaluate the conditional distribution $p_\theta(Z, G \mid S)$.

Inference

Aim. Infer the parameter $\theta = (\pi, \gamma, F_0, F_1)$, the node memberships $Z = (Z_j)$ and the graph $G = (G_{jk})$.

Incomplete data model.

- Neither the node memberships Z nor the underlying graph G are observed.
- The EM algorithm requires to evaluate the conditional distribution $p_\theta(Z, G \mid S)$.

Variational EM (VEM).

- Maximize a lower bound of the log-likelihood $\log p_\theta(S)$
- Using an approximation of the conditional distribution $p_\theta(Z, G \mid S)$:

$$\widetilde{p}(Z, G) = \widetilde{p}(Z) \times \widetilde{p}(G \mid Z)$$

where

$$\widetilde{p}(Z) = \prod_j \widetilde{p}(Z_j) \qquad \text{mean field approximation}$$

$$\widetilde{p}(G \mid Z) = p(G \mid Z, S) \qquad \text{exact form}$$

Some comments

1. When interested in deciphering a cluster structure among species or genes, there is no need to actually infer the network (avoid a delicate thresholding step)

2. The observed data Y do not appear in the model: the information it summarized in the score matrix S

3. The VEM algorithm provides both
   - the classification probabilities $\widetilde{P}\{Z_j = q\}$ for each node,
   - as a by-product: the probability for each edge to be part of the network $\widetilde{P}\{G_{jk} = 1\}$

4. We use Gaussian distributions for the scores: $F_0 = \mathcal{N}(\mu_0, \sigma_0^2)$, $F_1 = \mathcal{N}(\mu_1, \sigma_1^2)$

5. Q can be selected using standard (variational) *BIC* or *ICL* criteria. *ICL* can account for the conditional entropy of Z, or G, or both.

6. Same model as [RRV19], who focus on the control of the rate of false positive edges

Simulation study

Simulation design.

- $K = 3$ node clusters: $\pi = (17\%, 33\%, 50\%)$
- SBM node membership Z and graph G: $(Z, G) \sim SBM(\pi, \gamma)$, $\overline{\gamma} = 1.5 \log(p)/p$
- Gaussian data $Y \sim \mathcal{N}_p(0, \Omega_G^{-1})$
- Sample size $n = 20, 50, 100$
- Edge scores from Meinshausen-Bühlmann (M-B), glasso, tree-based algorithms

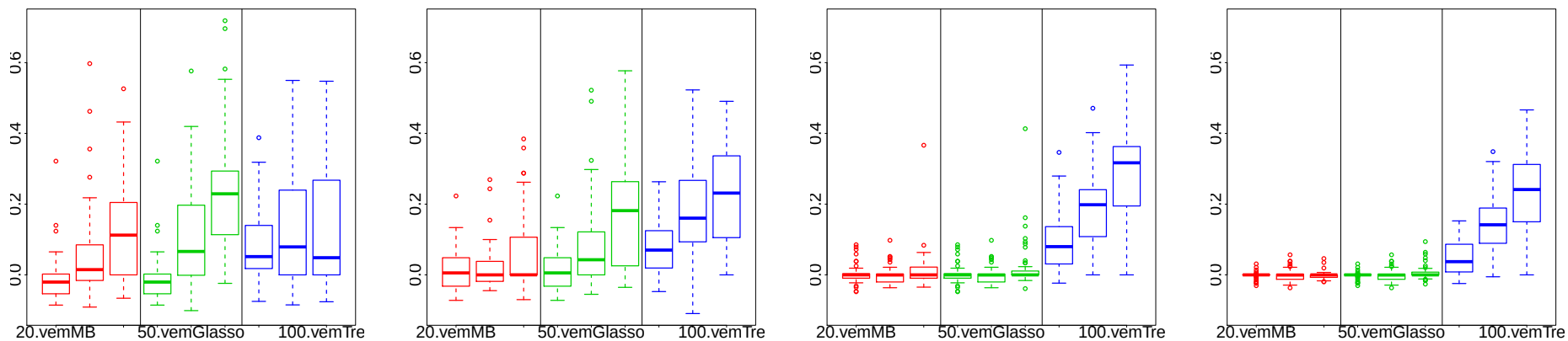
Simulation study

Simulation design.

- $K = 3$ node clusters: $\pi = (17\%, 33\%, 50\%)$
- SBM node membership Z and graph G: $(Z, G) \sim SBM(\pi, \gamma)$, $\overline{\gamma} = 1.5 \log(p)/p$
- Gaussian data $Y \sim \mathcal{N}_p(0, \Omega_G^{-1})$
- Sample size $n = 20, 50, 100$
- Edge scores from Meinshausen-Bühlmann (M-B), glasso, tree-based algorithms

Node classification: ARI = adjusted rand index

$p = 20$ $p = 30$ $p = 50$ $p = 80$



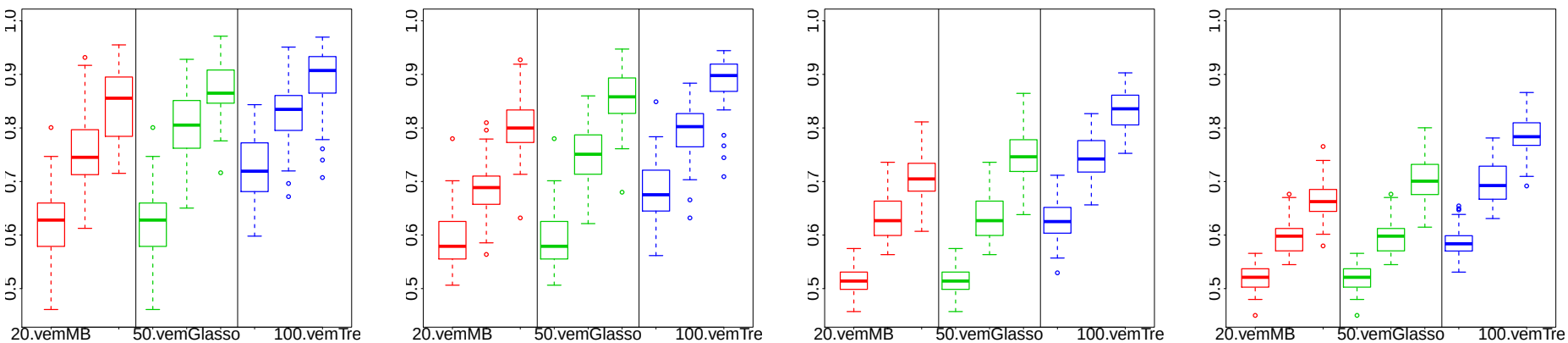
Simulation study

Simulation design.

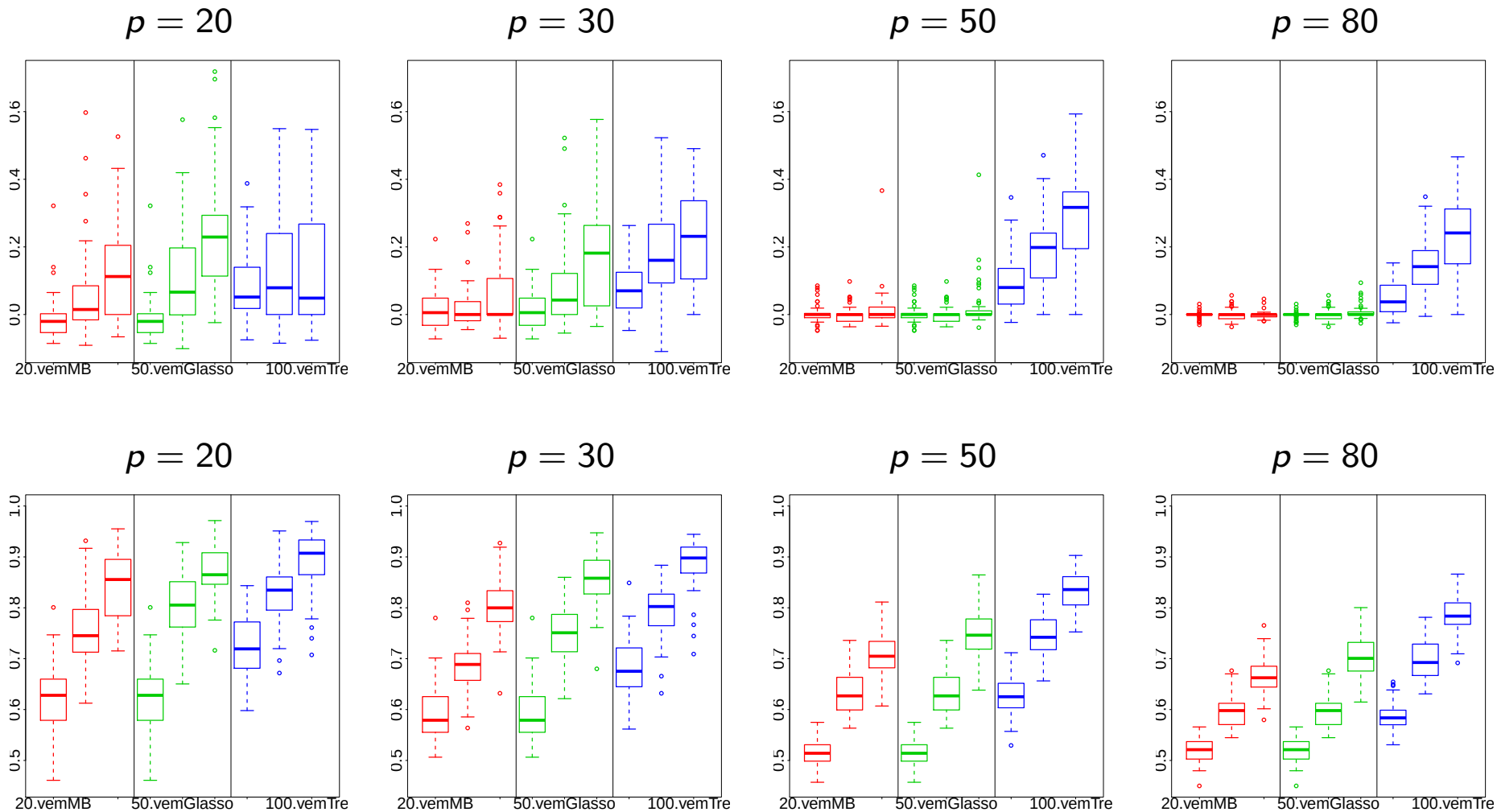
- $K = 3$ node clusters: $\pi = (17\%, 33\%, 50\%)$
- SBM node membership Z and graph G: $(Z, G) \sim SBM(\pi, \gamma)$, $\overline{\gamma} = 1.5 \log(p)/p$
- Gaussian data $Y \sim \mathcal{N}_p(0, \Omega_G^{-1})$
- Sample size $n = 20, 50, 100$
- Edge scores from Meinshausen-Bühlmann (M-B), glasso, tree-based algorithms

Edge recovery: AUC =area under the ROC curve

$p = 20$ $p = 30$ $p = 50$ $p = 80$



Top = node classification, bottom = edge recovery



Scores = M-B, glasso, tree, $n = 20, 50, 100$

- Issue with the choice of the grid of λ in M-B and glasso

Barents fish (1/2)

Dataset: [FNA06]

- $n = 89$ stations, $p = 30$ fish species,
- Y_{ij} = abundance (count) of species j in station i,
- 4 covariates (latitude, longitude, temperature and depth)

Barents fish (1/2)

Dataset: [FNA06]

- $n = 89$ stations, $p = 30$ fish species,
- Y_{ij} = abundance (count) of species j in station i,
- 4 covariates (latitude, longitude, temperature and depth)

Network inference:

- Fit a Poisson log-normal model [AH89] (`PLNmodels` package [CMR18])
- Compute edge scores using a tree-based method (`EMtree` R package [MRA19])

Barents fish (1/2)

Dataset: [FNA06]

- $n = 89$ stations, $p = 30$ fish species,
- Y_{ij} = abundance (count) of species j in station i,
- 4 covariates (latitude, longitude, temperature and depth)

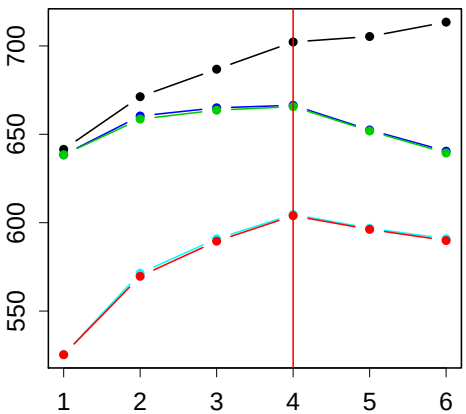
Network inference:

- Fit a Poisson log-normal model [AH89] (`PLNmodels` package [CMR18])
- Compute edge scores using a tree-based method (`EMtree` R package [MRA19])

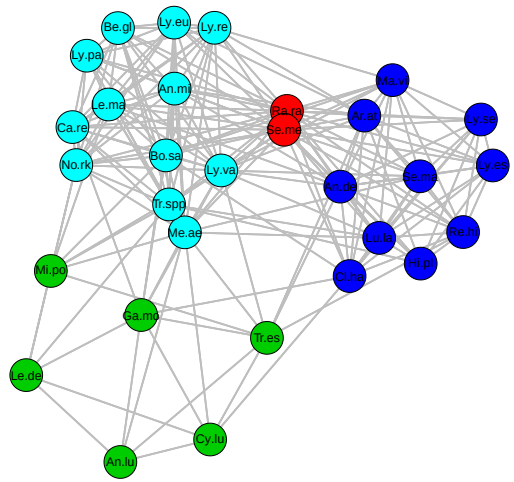
Choosing the number of clusters: $ICL(Z, G)$ criterion

Barents fish (2/2)

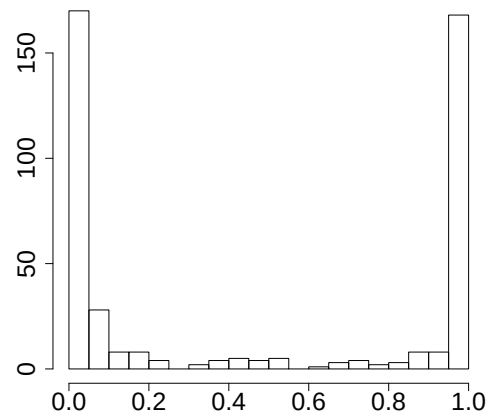
Choosing Q



Species clusters



Edge probabilities



Parameter estimates.

cluster proportions π

6.8	19.5	33.2	40.5

cluster connections γ

100	0.2	100	99.2
0.2	85.6	10	27.8
100	10	88.2	16.1
99.2	27.8	16.1	98.3

- $Q = 4$ node clusters are found, incl. one central cluster
- Low uncertainty for node classification [#22]
- Edge probabilities are highly contrasted
- The network is only drawn for an aesthetic purpose

Oak mildew (1/2)

Dataset: [JFS+16]

- Metabarcoding of $p = 114$ microbial and fungal species, including the mildew pathogen *E. alphitoides*
- Collected on $n = 116$ oak leaves
- Y_{ij} = read count for species j in leaf i
- 3 covariates (tree status, distances to ground and trunk)

Oak mildew (1/2)

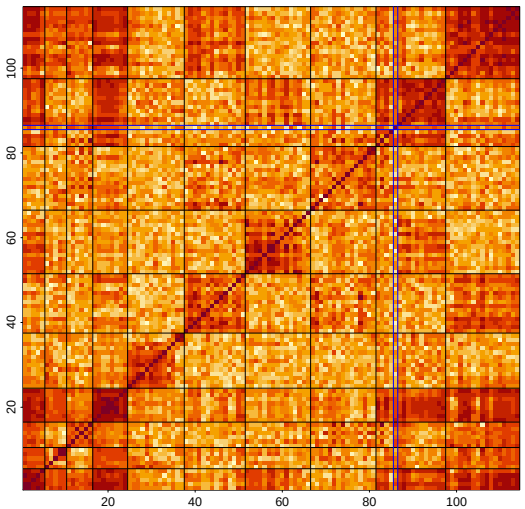
Dataset: [JFS+16]

- Metabarcoding of $p = 114$ microbial and fungal species, including the mildew pathogen *E. alphitoides*
- Collected on $n = 116$ oak leaves
- Y_{ij} = read count for species j in leaf i
- 3 covariates (tree status, distances to ground and trunk)

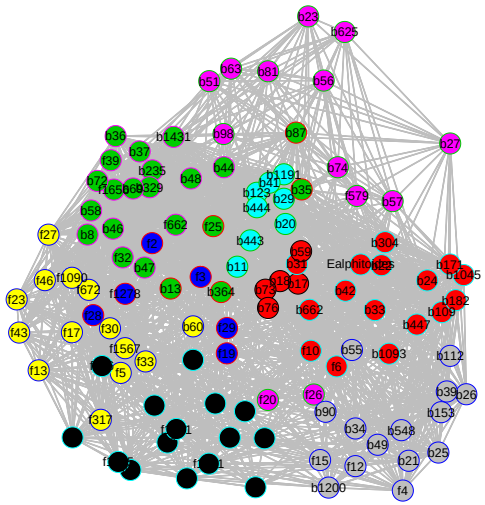
Network inference: same procedure as for Barents fish, accounting for differential sampling depth for fungi and bacteria

Oak mildew (2/2)

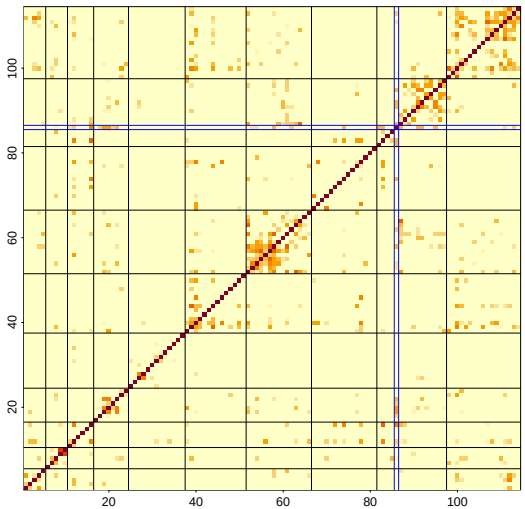
Species correlations



Species clusters



PLNnetwork result



- $Q = 10$ clusters found (max. value)
- Cluster structure in the correlation matrix (corrected for covariate effects)
- Consistent with direct network inference based on PLN/glasso approach [CMR19]
- Low uncertainty for node classification [#22]
- Highly contrasted edge probabilities [#22]
- The pathogens *E. alphitoides* is associated with 2 fungi and 13 bacterias

Discussion

To summarize.

- A formal probabilistic framework to account for network inference uncertainty in network topology analysis via SBM
- An agnostic approach with respect to the network inference procedure
- A new instance of SBM with mixture emission distribution
- A VEM algorithm with *BIC* and *ICL* variational criteria

Discussion

To summarize.

- A formal probabilistic framework to account for network inference uncertainty in network topology analysis via SBM
- An agnostic approach with respect to the network inference procedure
- A new instance of SBM with mixture emission distribution
- A VEM algorithm with *BIC* and *ICL* variational criteria

Further work.

- How to choose the score (i.e. the network inference method) in practice?
- Non-parametric form for the score distribution [#23]

References I

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

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. A variational baysian framework for graphical models. In *International Conference on Machine Learning*, 2019.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

M. Fossheim, E. M Nilssen, and M. Aschan. Fish assemblages in the barents sea. *Marine Biology Research*, 2(4):260–269, 2006.

B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen Erysiphe alphitoides. *Microbial ecology*, pages 1–11, 2016.

S. Kirshner. Learning with tree-averaged densities and distributions. In *NIPS*, pages 761–768, 2007.

M. Meilă and T. Jaakkola. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, March 2006.

R. Momal, S. Robin, and A. Ambroise. Tree-based reconstruction of ecological network from abundance data. Technical Report 1905.02452, arXiv, 2019.

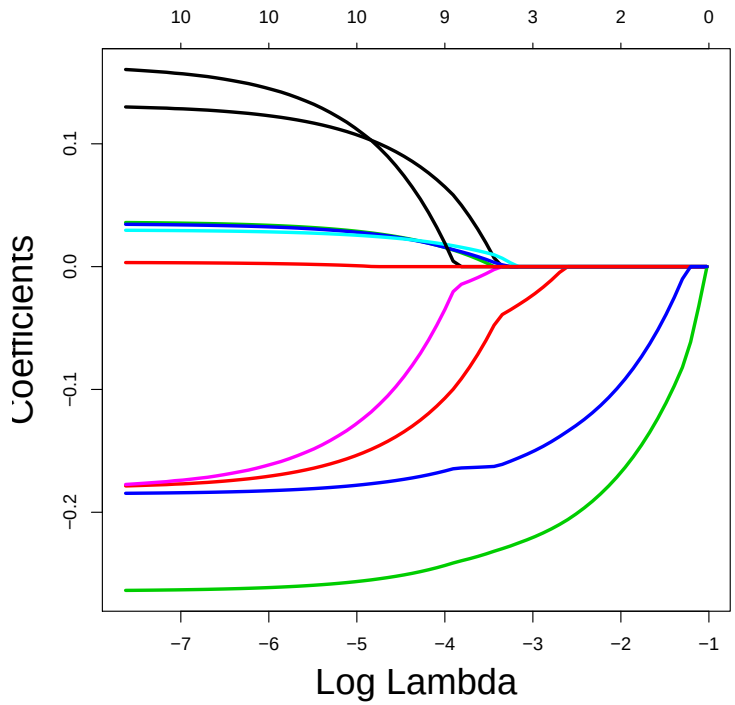
K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

T. Rebafka, E. Roquain, and F. Villers. Graph inference with clustering and false discovery rate control. Technical Report 1907.10176, arXiv, 2019.

L. Schwaller, S. Robin, and M. Stumpf. Bayesian Inference of Graphical Model Structures Using Trees. *J. Soc. Franc. Stat.*, 160(2):1–23, 2019.

Lasso: regularization path

Coefficients become null as λ increases

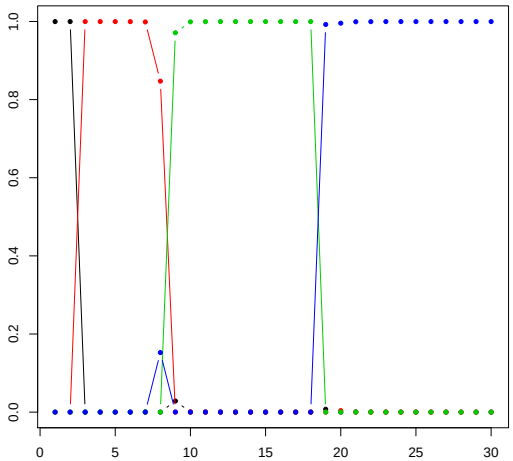


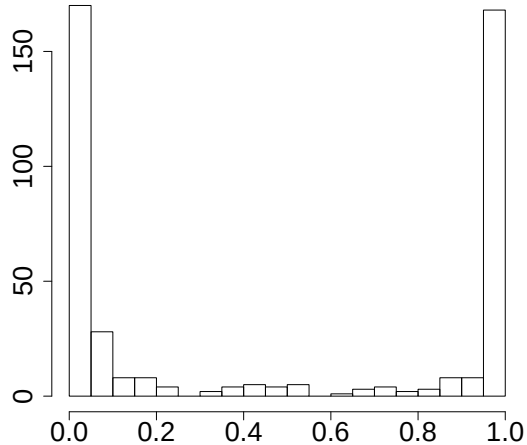
- Regularization path: succession of optimal solutions when λ decreases [#6]

Node membership and edge presence uncertainty

Barents fish.
$Q = 4$

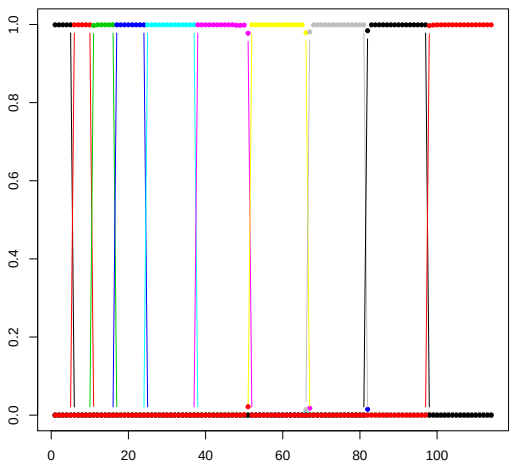
[#16]

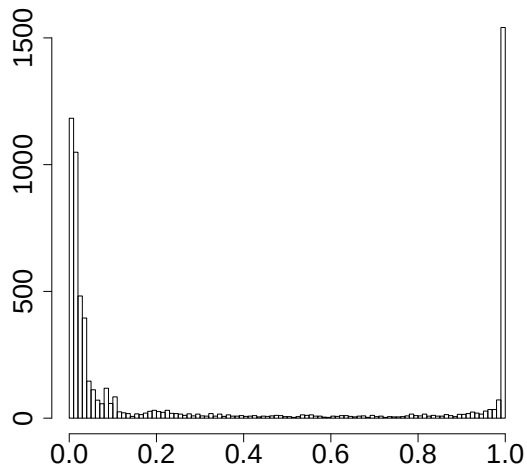




Oak mildew.
$Q = 9$

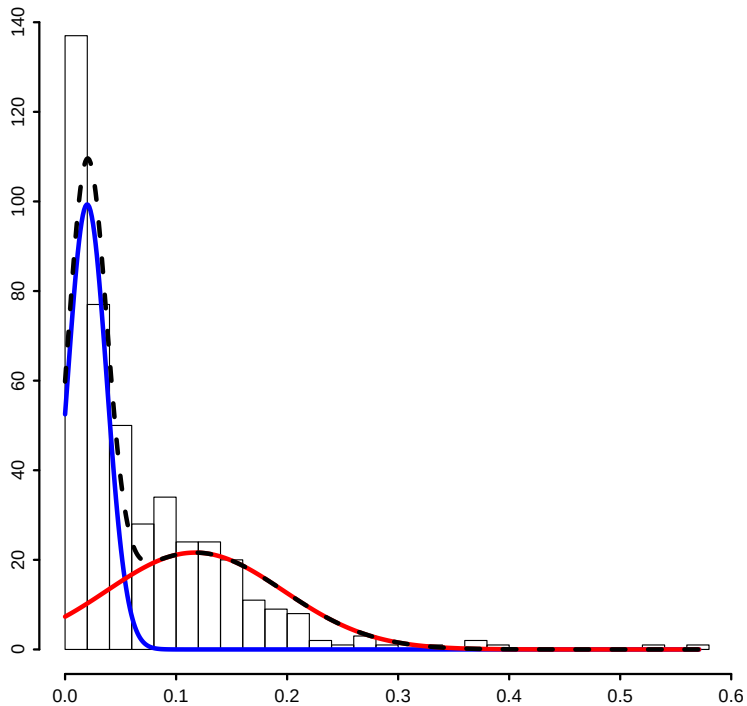
[#18]



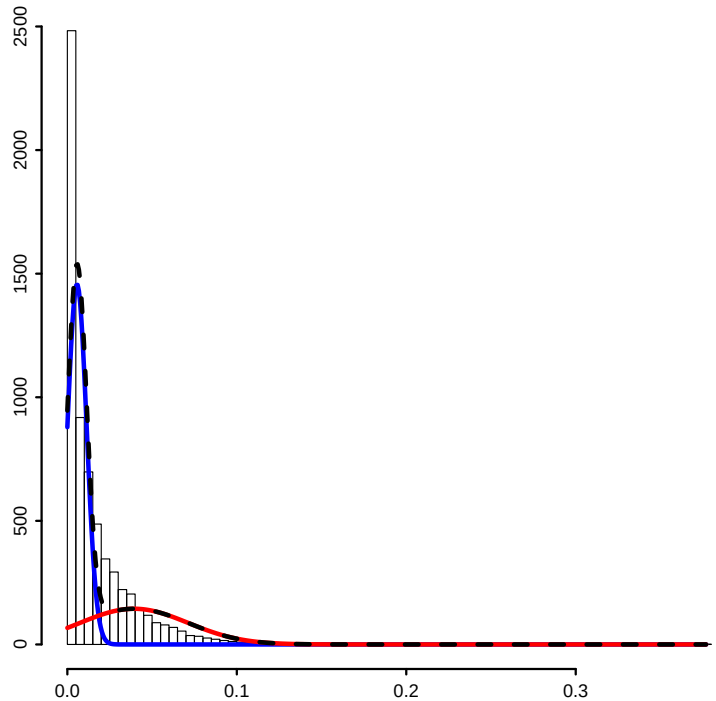


Score distribution

Barents fish.



Oak mildew.



[#19]