Uncovering Latent Structure in Valued Graphs

M. Mariadassou
Joint work with S. Robin and C. Vacher

Laboratoire MIG (UR INRA), Jouy-en-Josas, France.

Singapore, IMS, 10 May 2011

Outline

1. Introduction
2. MixNet: a Mixture Model for Random Graphs
3. Parametric Estimation
4. Simulation Study
5. Ecological Network

Yeast Protein Interaction Network (PIN)

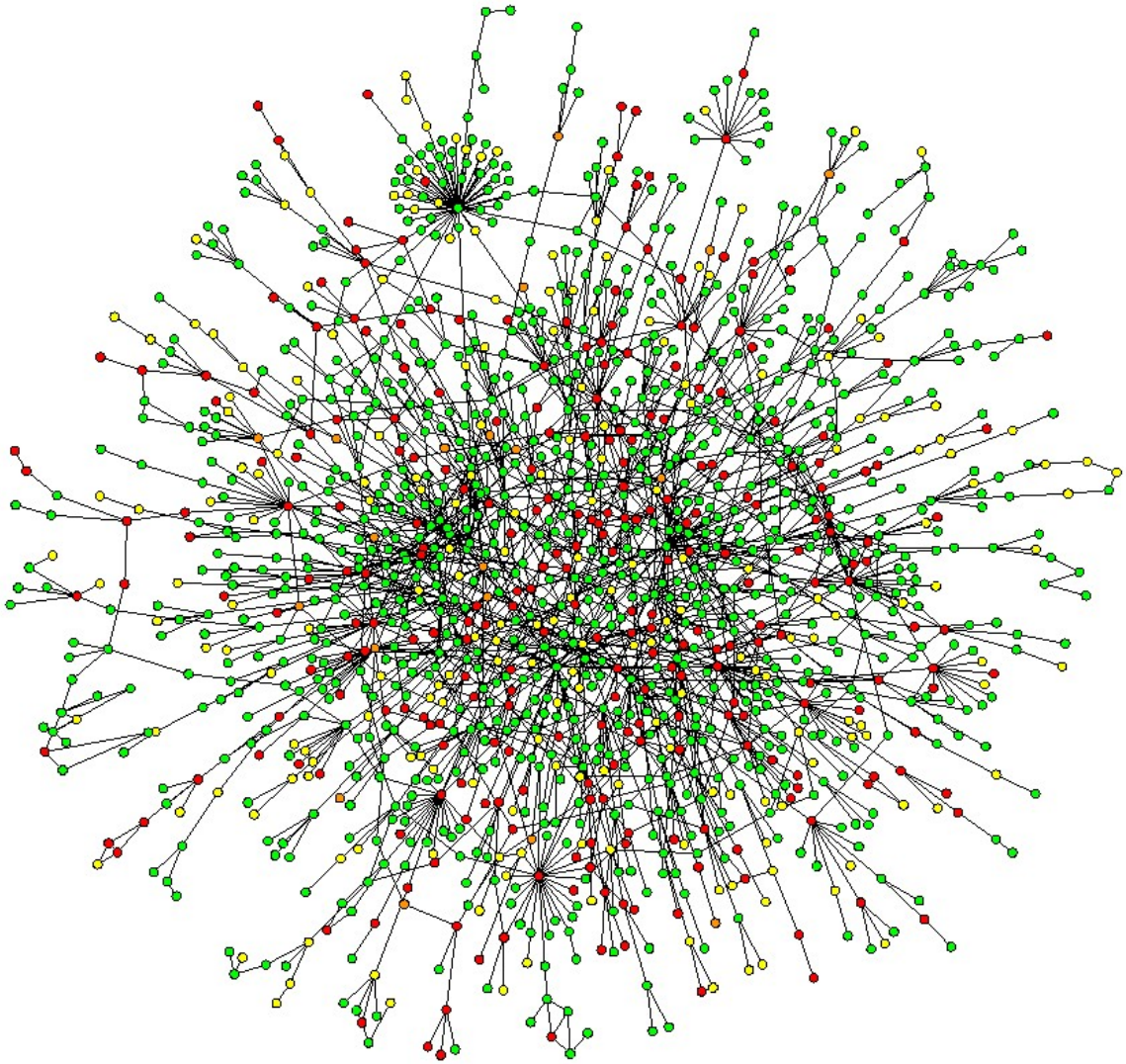


Figure: Yeast PIN. source: www.bordalierinstitute.com/images/yeastProteinInteractionNetwork.jpg

Goal: Simple Representation of the Graph

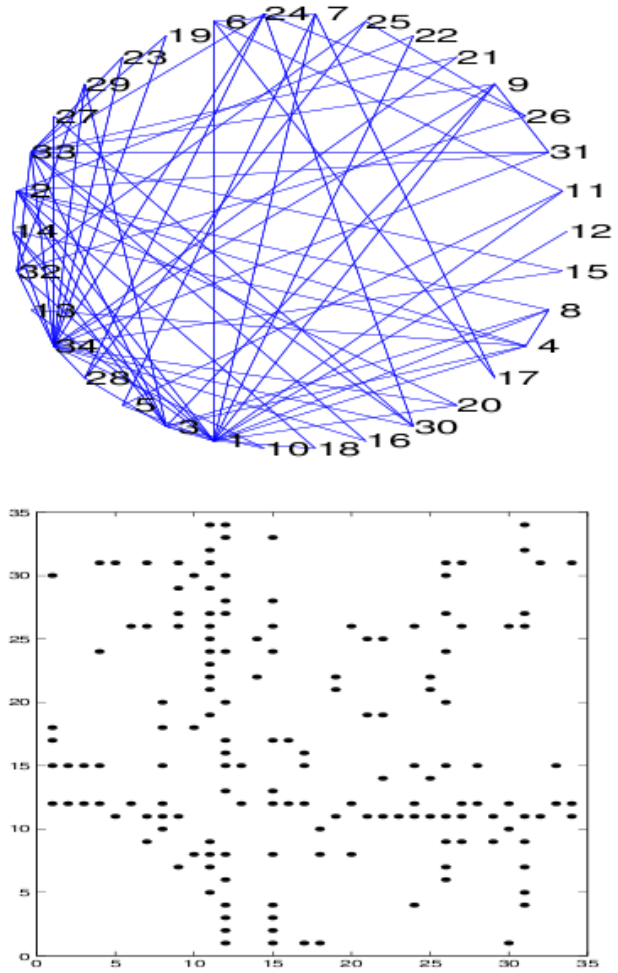


Figure: Zachary's karate club (Zachary 77)

Goal: Simple Representation of the Graph

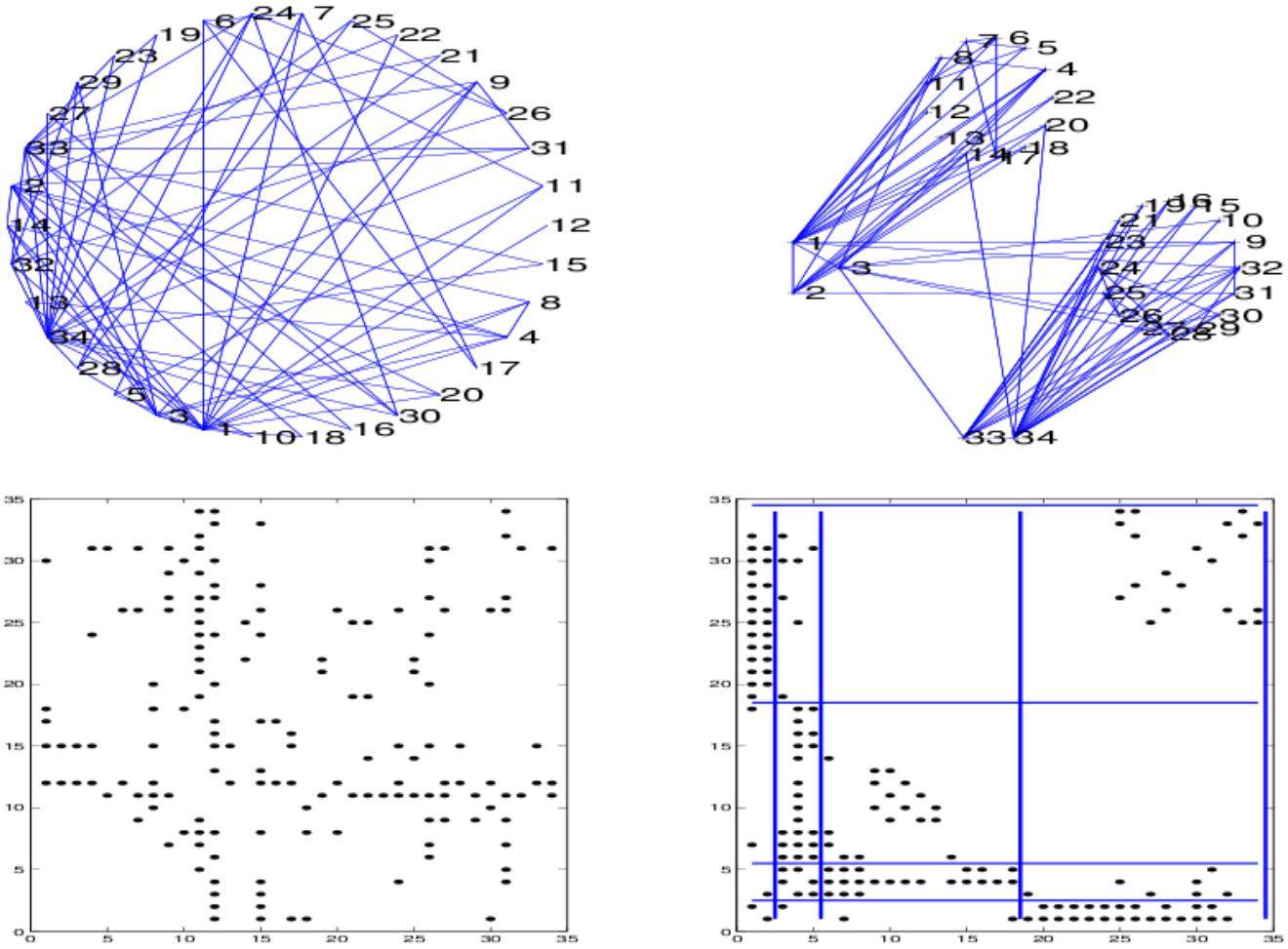


Figure: Zachary's karate club (Zachary 77)

Models for Networks

Classical Models

- Erdos-Renyi random graph (Erdos & Renyi 59);
- Degree distribution (Milo & al 04);
- Preferential Attachment (Barabasi & Albert 99);

Exponential Models

- ERGM (Holland & Leinhardt 81).

→ Local structure induced by relative frequencies of motifs.

Mixture Model

- Stochastic Block Model / MixNet (Holland & al 83, Fienberg & al 85, Snijders & Nowicki 97, Daudin & al 08)

→ Global structure induced by groups of similar nodes.

Models for Networks

Classical Models

- Erdos-Renyi random graph (Erdos & Renyi 59);
- Degree distribution (Milo & al 04);
- Preferential Attachment (Barabasi & Albert 99);

Exponential Models

- ERGM (Holland & Leinhardt 81).

→ Local structure induced by relative frequencies of motifs.

Mixture Model

- Stochastic Block Model / MixNet (Holland & al 83, Fienberg & al 85, Snijders & Nowicki 97, Daudin & al 08)

→ Global structure induced by groups of similar nodes.

Models for Networks

Classical Models

- Erdos-Renyi random graph (Erdos & Renyi 59);
- Degree distribution (Milo & al 04);
- Preferential Attachment (Barabasi & Albert 99);

Exponential Models

- ERGM (Holland & Leinhardt 81).

→ Local structure induced by relative frequencies of motifs.

Mixture Model

- Stochastic Block Model / MixNet (Holland & al 83, Fienberg & al 85, Snijders & Nowicki 97, Daudin & al 08)

→ Global structure induced by groups of similar nodes.

MixNet Probabilistic Model (nodes)

Nodes heterogeneity

- The nodes are distributed among Q different classes (*e.g.* ●,▲,■);
- $\mathbf{Z} = (Z_i)_{i=1..n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ are the group proportions;
- Z_i is not observed.

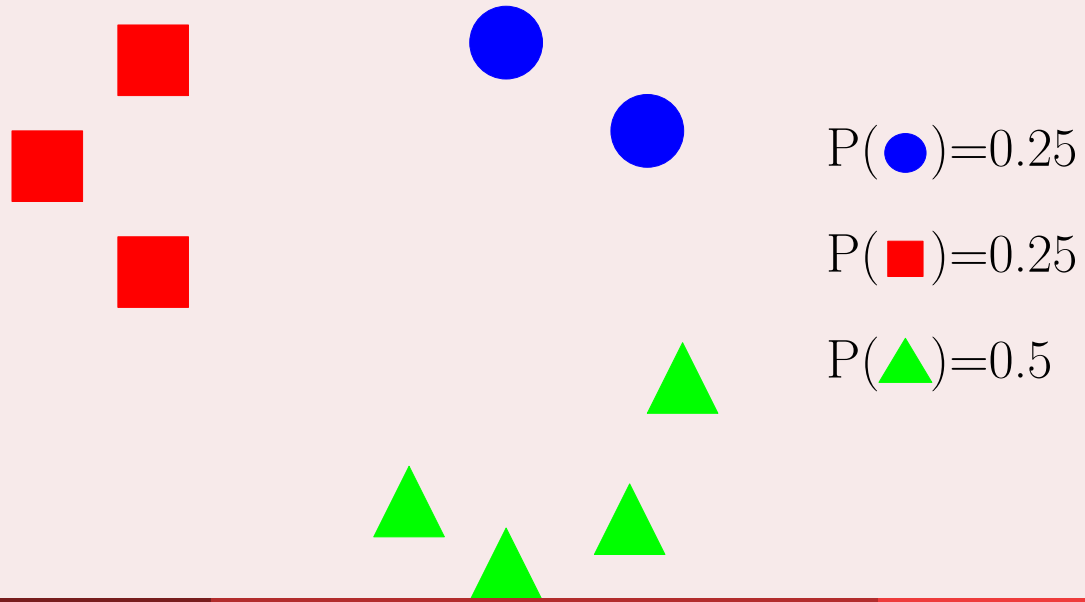
Example: (9 nodes, 3 classes)

MixNet Probabilistic Model (nodes)

Nodes heterogeneity

- The nodes are distributed among Q different classes (*e.g.* ●,▲,■);
- $\mathbf{Z} = (Z_i)_{i=1..n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ are the group proportions;
- Z_i is not observed.

Example: (9 nodes, 3 classes)



MixNet Probabilistic Model (edges)

Observations

- Edges values X_{ij} where $X_{ij} \in \mathbb{R}^s$;
- Conditional on $\mathbf{Z}$, the (X_{ij}) are independent with distribution

$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim f(., \theta_{q\ell})$$

- $\boldsymbol{\theta} = (\theta_{q\ell})_{q,\ell=1..Q}$ is the connectivity parameter.

Example: 3 classes with Poisson-valued edges

MixNet Probabilistic Model (edges)

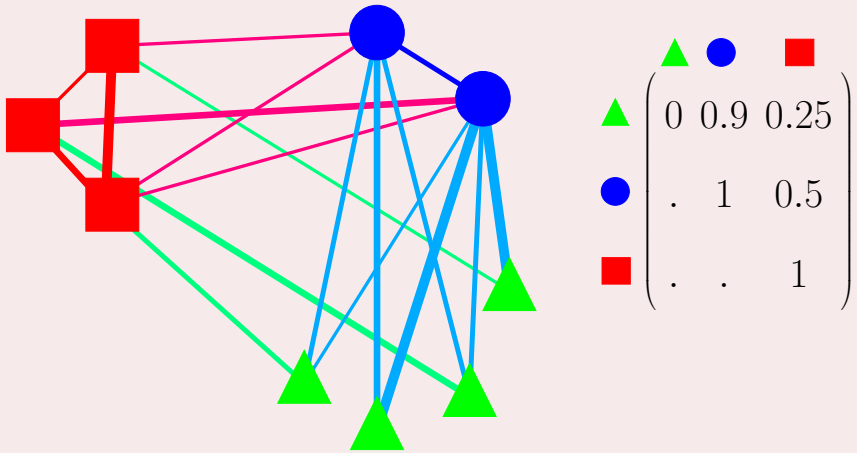
Observations

- Edges values X_{ij} where $X_{ij} \in \mathbb{R}^s$;
- Conditional on $\mathbf{Z}$, the (X_{ij}) are independent with distribution

$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim f(., \theta_{q\ell})$$

- $\boldsymbol{\theta} = (\theta_{q\ell})_{q,\ell=1..Q}$ is the connectivity parameter.

Example: 3 classes with Poisson-valued edges



Flexibility of MixNet

Classical Distributions:

- f_θ can be any probability distribution;

→ Bernoulli (interaction graph): presence/absence of an edge;

$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim \mathcal{B}(\pi_{q\ell})$$

→ Poisson (PM) (count): in coauthorship networks, number of copublished papers;

$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim \mathcal{P}(\lambda_{q\ell})$$

→ Poisson regression with homogeneous effects (PRMH) (counts with covariates): in ecological networks;

$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim \mathcal{P}(\lambda_{q\ell} \exp\{\beta^\intercal \mathbf{y}_{ij}\})$$

(Log)-Likelihood of the Model

- **Complete data likelihood**

$$\begin{aligned}\mathcal{L}(\mathbf{X},\mathbf{Z}) &= \ln \Pr(\mathbf{X},\mathbf{Z}) = \ln \Pr(\mathbf{Z}) P(\mathbf{X}|\mathbf{Z}) \\ &= \sum_i \sum_q Z_{iq} \ln \alpha_q + \sum_{i<j} \sum_{q,l} Z_{iq} Z_{jl} \ln f_{\theta_{ql}}(X_{ij})\end{aligned}$$

- **Observed data likelihood**

$$\mathcal{L}(\mathbf{X}) = \ln \sum_{\mathbf{Z}} \exp \mathcal{L}(\mathbf{X},\mathbf{Z})$$

- Sum over $\mathcal{Q}^n$ is untractable, use EM algorithm instead.

(Log)-Likelihood of the Model

- **Complete data likelihood**

$$\mathcal{L}(\mathbf{X}, \mathbf{Z}) = \sum_i \sum_q Z_{iq} \ln \alpha_q + \sum_{i<j} \sum_{q,l} Z_{iq} Z_{jl} \ln f_{\theta_{ql}}(X_{ij})$$

- **Observed data likelihood**

$$\mathcal{L}(\mathbf{X}) = \ln \sum_{\mathbf{Z}} \exp \mathcal{L}(\mathbf{X}, \mathbf{Z})$$

- Sum over Q^n is untractable, use EM algorithm instead.

But...

- The random variables X_{ij} are not independent;
- The distribution $\Pr(.|\mathbf{X})$ of $\mathbf{Z}$ conditional on $\mathbf{X}$ is not a product distribution;

$\rightarrow$ Exact EM is not possible...

Variational Inference: Pseudo Likelihood

- If $\mathcal{R}_{\mathbf{X}}$ is a distribution over $\mathbf{Z}$, let

$$\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$$

- For $\mathcal{R}_{\mathbf{X}} = \Pr(.|\mathbf{X})$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X})$;
- **Variational approximation:** replace complicated distribution $\Pr(.|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}$ such that $KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$ is minimal to obtain a tight lower bound of $\mathcal{L}(\mathbf{X})$.

$$\begin{aligned} \mathcal{J}(\mathcal{R}_{\mathbf{X}}) &= \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X})) \\ &= \mathcal{H}(\mathcal{R}_{\mathbf{X}}) + \mathbb{E}_{\mathcal{R}_{\mathbf{X}}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})] \end{aligned}$$

where $\mathcal{H}(\mathcal{R}_{\mathbf{X}})$ is the entropy of $\mathcal{R}_{\mathbf{X}}$.

Variational Inference: Pseudo Likelihood

- If $\mathcal{R}_{\mathbf{X}}$ is a distribution over $\mathbf{Z}$, let

$$\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$$

- For $\mathcal{R}_{\mathbf{X}} = \Pr(.|\mathbf{X})$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X})$;
- **Variational approximation:** replace complicated distribution $\Pr(.|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}$ such that $KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$ is minimal to obtain a tight lower bound of $\mathcal{L}(\mathbf{X})$.

$$\begin{aligned} \mathcal{J}(\mathcal{R}_{\mathbf{X}}) &= \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X})) \\ &= \mathcal{H}(\mathcal{R}_{\mathbf{X}}) + \mathbb{E}_{\mathcal{R}_{\mathbf{X}}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})] \end{aligned}$$

where $\mathcal{H}(\mathcal{R}_{\mathbf{X}})$ is the entropy of $\mathcal{R}_{\mathbf{X}}$.

Variational Inference: Pseudo Likelihood

- If $\mathcal{R}_{\mathbf{X}}$ is a distribution over $\mathbf{Z}$, let

$$\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$$

- For $\mathcal{R}_{\mathbf{X}} = \Pr(.|\mathbf{X})$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}) = \mathcal{L}(\mathbf{X})$;
- **Variational approximation:** replace complicated distribution $\Pr(.|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}$ such that $KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X}))$ is minimal to obtain a tight lower bound of $\mathcal{L}(\mathbf{X})$.

$$\begin{aligned} \mathcal{J}(\mathcal{R}_{\mathbf{X}}) &= \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_{\mathbf{X}}, \Pr(.|\mathbf{X})) \\ &= \mathcal{H}(\mathcal{R}_{\mathbf{X}}) + \mathbb{E}_{\mathcal{R}_{\mathbf{X}}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})] \end{aligned}$$

where $\mathcal{H}(\mathcal{R}_{\mathbf{X}})$ is the entropy of $\mathcal{R}_{\mathbf{X}}$.

Variational Inference: Pseudo Likelihood (II)

- Computing $\mathbb{E}_{\mathcal{R}_\mathbf{X}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})]$ is easy, computing $\mathcal{H}(\mathcal{R}_\mathbf{X})$ is hard (in general).
- Restrict $\mathcal{R}_\mathbf{X}$ to a comfortable class of distributions:

$$\mathcal{R}_\mathbf{X}[\mathbf{Z}] = \prod_i h(Z_i; \boldsymbol{\tau}_i)$$

with $h(.; \boldsymbol{\tau}_i)$ the multinomial with paramater $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iQ})$.
Intuitively, $\tau_{iq} \simeq \Pr(Z_{iq} = 1|\mathbf{X})$.

- For such $\mathcal{R}_\mathbf{X}$,

$$\mathcal{J}((\boldsymbol{\tau}_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq}\tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$

Variational Inference: Pseudo Likelihood (II)

- Computing $\mathbb{E}_{\mathcal{R}_\mathbf{X}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})]$ is easy, computing $\mathcal{H}(\mathcal{R}_\mathbf{X})$ is hard (in general).
- Restrict $\mathcal{R}_\mathbf{X}$ to a comfortable class of distributions:

$$\mathcal{R}_\mathbf{X}[\mathbf{Z}] = \prod_i h(Z_i; \boldsymbol{\tau}_i)$$

with $h(.; \boldsymbol{\tau}_i)$ the multinomial with paramater $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iQ})$. Intuitively, $\tau_{iq} \simeq \Pr(Z_{iq} = 1|\mathbf{X})$.

- For such $\mathcal{R}_\mathbf{X}$,

$$\mathcal{J}((\boldsymbol{\tau}_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$

Variational Inference: Pseudo Likelihood (II)

- Computing $\mathbb{E}_{\mathcal{R}_\mathbf{X}}[\mathcal{L}(\mathbf{X}, \mathbf{Z})]$ is easy, computing $\mathcal{H}(\mathcal{R}_\mathbf{X})$ is hard (in general).
- Restrict $\mathcal{R}_\mathbf{X}$ to a comfortable class of distributions:

$$\mathcal{R}_\mathbf{X}[\mathbf{Z}] = \prod_i h(Z_i; \boldsymbol{\tau}_i)$$

with $h(.; \boldsymbol{\tau}_i)$ the multinomial with paramater $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iQ})$. Intuitively, $\tau_{iq} \simeq \Pr(Z_{iq} = 1|\mathbf{X})$.

- For such $\mathcal{R}_\mathbf{X}$,

$$\mathcal{J}((\boldsymbol{\tau}_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq}\tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$

2 Steps Iterative Algorithm

- Maximize pseudo-likelihood:

$$\mathcal{J}((\alpha,\theta),(\tau_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq}\tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$

- Step 1 **Optimize $\mathcal{J}$ w.r.t. (τ_i)**:
  - → Constraint: $\sum_q \tau_{iq} = 1$ for all i;
  - → τ_{iq} variational parameter found via a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{\ell=1}^{Q} f_{\theta_{q\ell}}(X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize $\mathcal{J}$ w.r.t. (α,θ)**:
  - → Constraint: $\sum_q \alpha_q = 1$

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\theta}_{ql} &= arg\max_{\theta} \sum_{i,j} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \log f_{\theta}(X_{ij}) \end{aligned}$$

  - → Simple expression of $\tilde{\theta}_{ql}$ for classical distributions (weighted MLE).

2 Steps Iterative Algorithm

- Maximize pseudo-likelihood:

$$\mathcal{J}((\alpha,\theta),(\tau_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq}\tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$

- Step 1 **Optimize $\mathcal{J}$ w.r.t. (τ_i)**:
  - → Constraint: $\sum_q \tau_{iq} = 1$ for all i;
  - → τ_{iq} variational parameter found via a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{\ell=1}^{Q} f_{\theta_{q\ell}}(X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize $\mathcal{J}$ w.r.t. (α,θ)**:
  - → Constraint: $\sum_q \alpha_q = 1$

$$\tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq}/n$$

$$\tilde{\theta}_{ql} = arg\max_{\theta} \sum_{i,j} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \log f_{\theta}(X_{ij})$$

  - → Simple expression of $\tilde{\theta}_{ql}$ for classical distributions (weighted MLE).

2 Steps Iterative Algorithm

- Maximize pseudo-likelihood:

$$\mathcal{J}((\alpha,\theta),(\tau_i)_{i=1..n}) = -\sum_i\sum_q \tau_{iq}\ln\tau_{iq} + \sum_i\sum_q \tau_{iq}\ln\alpha_q + \sum_{i<j}\tau_{iq}\tau_{j\ell}\ln f_{\theta_{q\ell}}(X_{ij})$$

- Step 1 **Optimize $\mathcal{J}$ w.r.t. (τ_i)**:
  - → Constraint: $\sum_q \tau_{iq} = 1$ for all i;
  - → τ_{iq} variational parameter found via a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j\neq i}\prod_{\ell=1}^{Q} f_{\theta_{q\ell}}(X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize $\mathcal{J}$ w.r.t. (α,θ)**:
  - → Constraint: $\sum_q \alpha_q = 1$

$$\begin{aligned}\tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\theta}_{ql} &= arg\max_\theta \sum_{i,j}\tilde{\tau}_{iq}\tilde{\tau}_{jl}\log f_\theta(X_{ij})\end{aligned}$$

  - → Simple expression of $\tilde{\theta}_{ql}$ for classical distributions (weighted MLE).

Model Selection Criterion

- BIC-like criterion to select the number of classes;
- The likelihood can be split: $\mathcal{L}(\mathbf{X}, \mathbf{Z}|Q) = \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) + \mathcal{L}(\mathbf{Z}|Q)$;
- These terms can be penalized separately:

$$\begin{aligned} \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) &\rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} P_Q \log n(n-1) \\ \mathcal{L}(\mathbf{Z}|Q) &\rightarrow \text{pen}_{\mathbf{Z}} = (Q-1)\log(n) \end{aligned}$$

$$ICL(Q) = \max_{\theta} \mathcal{L}(\mathbf{X}, \tilde{\mathbf{Z}}|\theta, m_Q) - \frac{1}{2}\left(P_Q \log n(n-1) - (Q-1)\log(n)\right)$$

Model Selection Criterion

- BIC-like criterion to select the number of classes;
- The likelihood can be split: $\mathcal{L}(\mathbf{X}, \mathbf{Z}|Q) = \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) + \mathcal{L}(\mathbf{Z}|Q)$;
- These terms can be penalized separately:

$$\begin{aligned} \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) &\rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} = P_Q \log n(n-1) \\ \mathcal{L}(\mathbf{Z}|Q) &\rightarrow \text{pen}_{\mathbf{Z}} = (Q-1)\log(n) \end{aligned}$$

$$ICL(Q) = \max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{X}, \tilde{\mathbf{Z}}|\boldsymbol{\theta}, m_Q) - \frac{1}{2}\left(P_Q \log n(n-1) - (Q-1)\log(n)\right)$$

MixNet Properties

Identifiability

- Identifiablity of Parameters (Allman and al., 2009, 2011);
- Model Selection criteria (Daudin and al., 2008, Latouche and al., 2011)

Quality of Estimates

- VEM algorithm converge to a different optimum than ML in the general case (Gunawardana and Byrne (2005)), except for degenerated models;
- SBM are in a certain sense degenerated: $\Pr(.|\mathbf{X}) \to \delta_{\mathbf{Z}}$ (ongoing work of Celisse and Daudin, Mariadassou and Matias)

MixNet Properties

Identifiability

- Identifiablity of Parameters (Allman and al., 2009, 2011);
- Model Selection criteria (Daudin and al., 2008, Latouche and al., 2011)

Quality of Estimates

- VEM algorithm converge to a different optimum than ML in the general case (Gunawardana and Byrne (2005)), except for degenerated models;
- SBM are in a certain sense degenerated: $\Pr(.|\mathbf{X}) \to \delta_{\mathbf{Z}}$ (ongoing work of Celisse and Daudin, Mariadassou and Matias)

Quality of the Estimates: Simulation Setup

- → Undirected graph with $Q = 3$ classes;
- → Poisson-valued edges;
- → $n = 100,\ 500$ vertices;
- → $\alpha_q \propto a^q$ for $a = 1,\ 0.5,\ 0.2$;
  - $a = 1$: balanced classes;
  - $a = 0.2$: unbalanced classes $(80.6\%,\ 16.1\%,\ 3.3\%)$
- → Connectivity matrix of the form $\begin{pmatrix} \lambda & \gamma\lambda & \gamma\lambda \\ \gamma\lambda & \lambda & \gamma\lambda \\ \gamma\lambda & \gamma\lambda & \lambda \end{pmatrix}$ for $\gamma = 0.1,\ 0.5,\ 0.9,\ 1.5$ and $\lambda = 2,\ 5$.
  - $\gamma = 1$: all classes equivalent (same connectivity pattern);
  - $\gamma \neq 1$: classes are different;
  - λ: mean value of an edge;
- → 100 repeats for each setup.

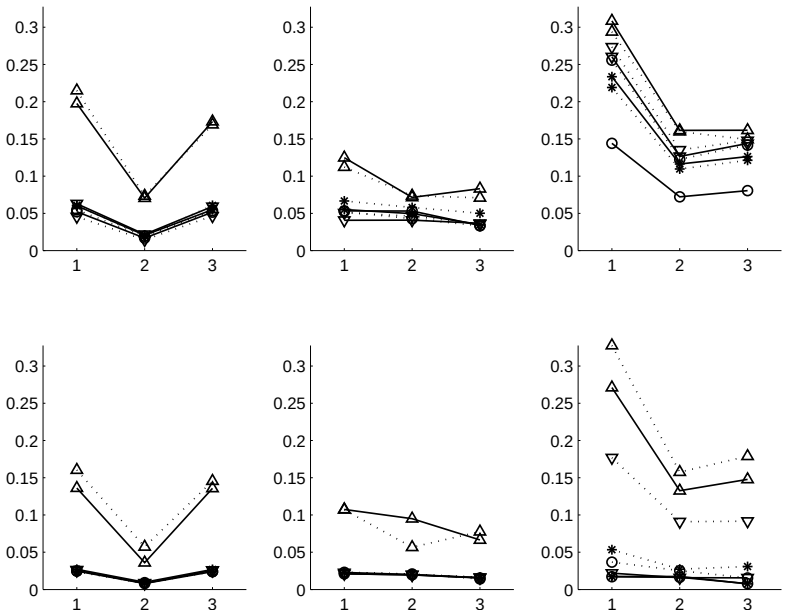
Quality of the Estimates: Results

- Root Mean Square Error (RMSE) = $\sqrt{Bias^2 + Variance}$

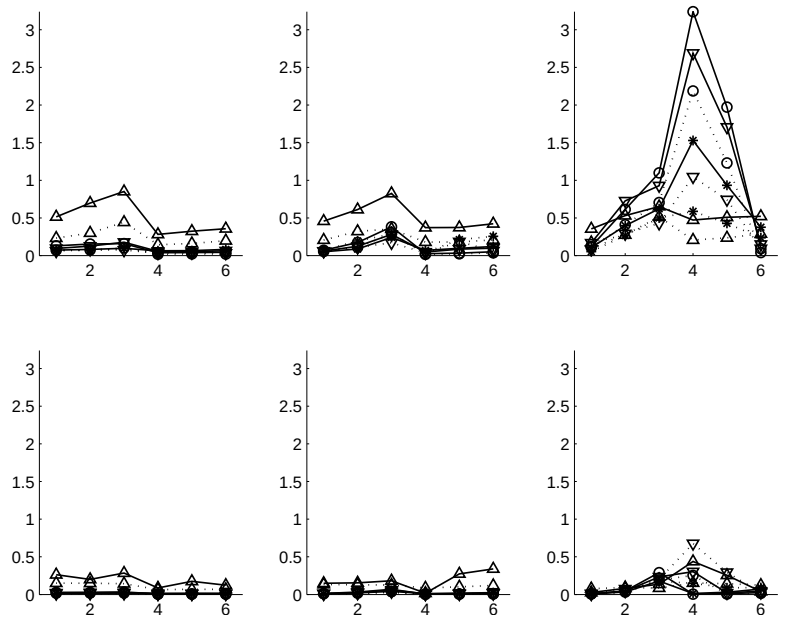
Quality of the Estimates: Results

- Root Mean Square Error (RMSE) = $\sqrt{Bias^2 + Variance}$

RMSE for the α_q



RMSE for the λ_{ql}



x-axis: $\alpha_1, \alpha_2, \alpha_3$

x-axis: $\lambda_{11}, \lambda_{22}, \lambda_{33}, \lambda_{12}, \lambda_{13}, \lambda_{23}$

Top: $n = 100$, Bottom: $n = 500$

Left to right: $a = 1,\ 0.5,\ 0.2$

Solid line: $\lambda = 5$, dashed line: $\lambda = 2$

Symbols depend on γ: $\circ = 0.1$, $\nabla = 0.5$, $\Delta = 0.9$, $* = 1.5$

Number of Classes

- → Undirected graph with $Q^\star = 3$ classes and Poisson edges;
- → $n = 50,\ 100,\ 500,\ 1000$ vertices;
- → $\alpha_q = (57.1\%, 28, 6\%, 14, 3\%)$;
- → Connectivity matrix of the form $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

	Q		
n	2	3	4
50	82	**17**	1
100	7	**90**	3
500	0	**100**	0
1000	0	**100**	0

Table: Frequency of selected Q for various n.

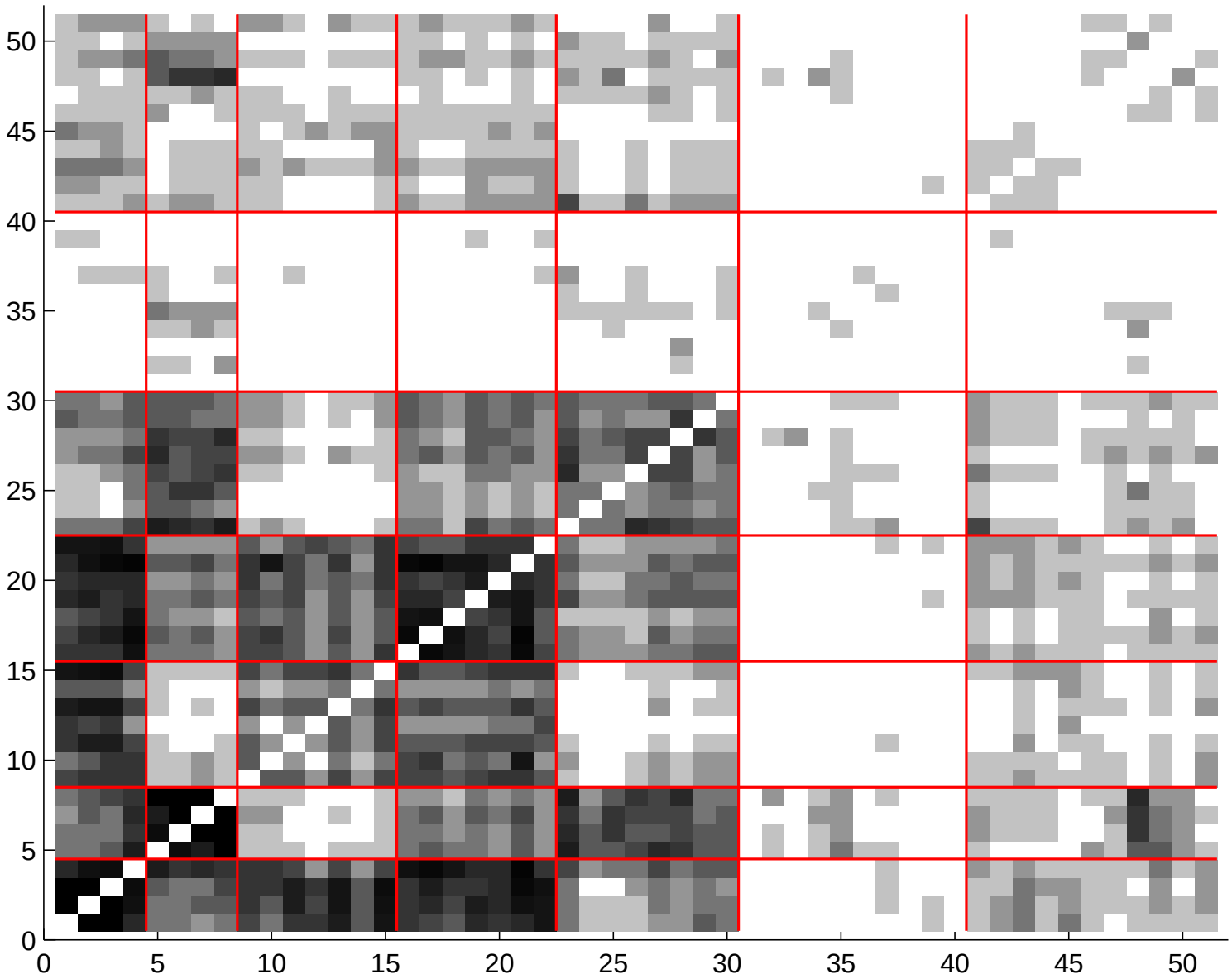
Fungi Trees Interactions

- **Dataset** Parisitic behavior of 154 fungi on 51 trees;
- **Network** Valued Network on trees: $X_{tt'}$ = number of fungis infecting both t and t'.
- **Goal** Identify groups of trees sharing similar interactions: is similarity driven by evolution or geography ?
- **Poisson Model** We assume

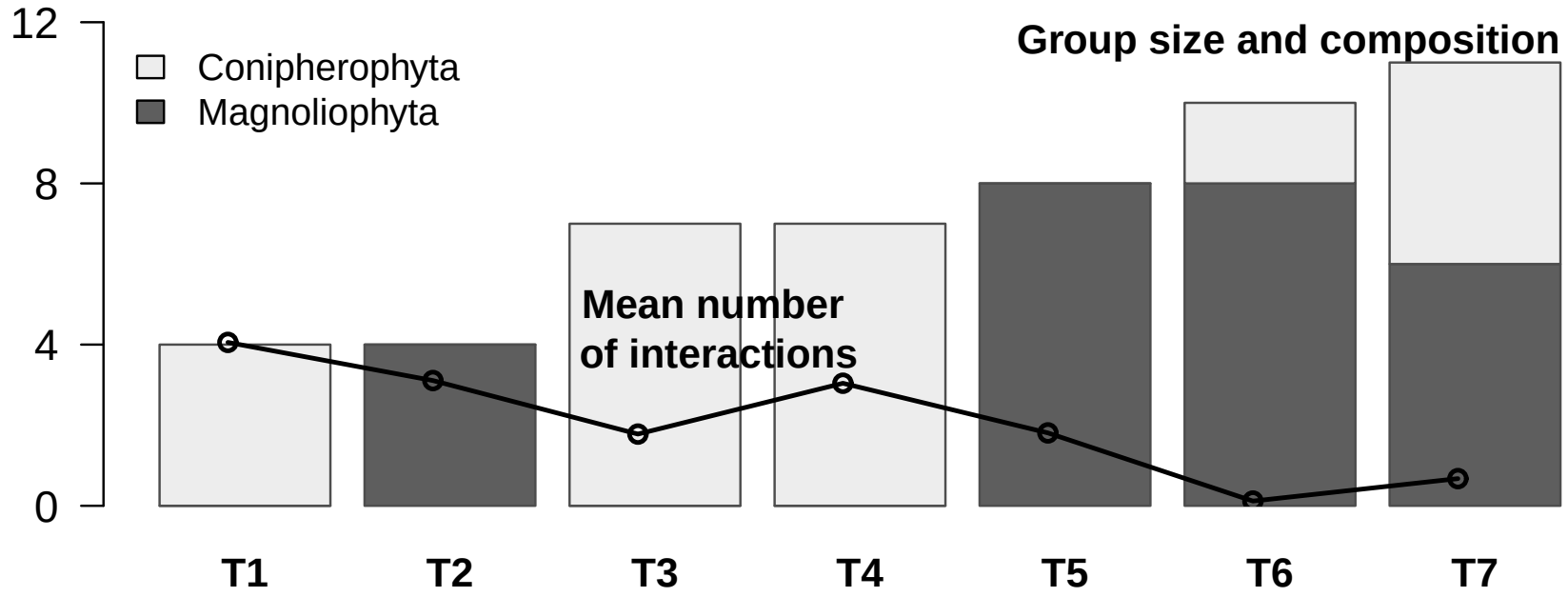
$$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim \mathcal{P}(\lambda_{q\ell})$$

- **Covariate**
  - Phylogenetic relatedness measured by genetic\taxonomic distance;
  - Geographical relatedness measured by Jaccard distance;

With no covariate (7 classes)

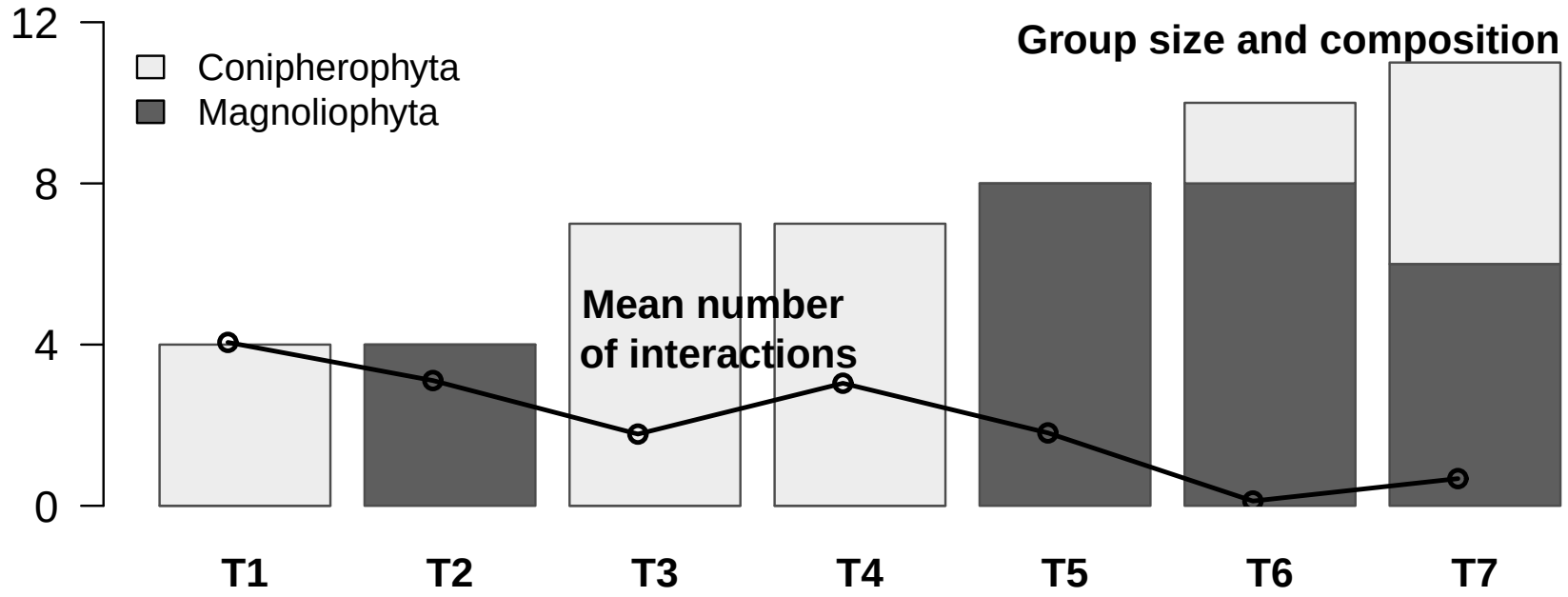


Groups of Trees: No Covariate



- Taxonomic rank: species ¡ genus ¡ family ¡ order ¡ class ¡ phylum;
- Strong effect of taxonomic rank on the group composition;
- Groups T1, T2, T3, T4 are even monofamily;
- Need to account for taxonomic distance.

Groups of Trees: No Covariate



- Taxonomic rank: species ¡ genus ¡ family ¡ order ¡ class ¡ phylum;
- Strong effect of taxonomic rank on the group composition;
- Groups T1, T2, T3, T4 are even monofamily;
- Need to account for taxonomic distance.

Groups of Trees: No Covariate (II)

$\widehat{\lambda}_{q\ell}$	T1	T2	T3	T4	T5	T6	T7
T1	14.46	4.19	5.99	7.67	2.44	0.13	1.43
T2	4.19	14.13	0.68	2.79	4.84	0.53	1.54
T3	5.99	0.68	3.19	4.10	0.66	0.02	0.69
T4	7.67	2.79	4.10	7.42	2.57	0.04	1.05
T5	2.44	4.84	0.66	2.57	3.64	0.23	0.83
T6	0.13	0.53	0.02	0.04	0.23	0.04	0.06
T7	1.43	1.54	0.69	1.05	0.83	0.06	0.27
$\widehat{\alpha}_q$	7.8	7.8	13.7	13.7	15.7	19.6	21.6

- T1, T2, T3, T4, T5: trees sharing lots of parasites;
- T6, T7: Trees with sharing few parasites with any other.

Groups of Trees: With Covariate

Model: $X_{ij} \sim \mathcal{P}(\lambda_{ql} e^{\beta y_{ij}})$ with y_{ij} taxonomic distance

- $\hat{Q} = 4$ classes;
- $\hat{\beta} = -0.317$;

	T'1	T'2	T'3	T'4
T1	0	0	0	4
T2	0	0	0	4
T3	2	5	0	0
T4	0	2	0	5
T5	0	2	0	6
T6	0	0	10	0
T7	7	2	2	0

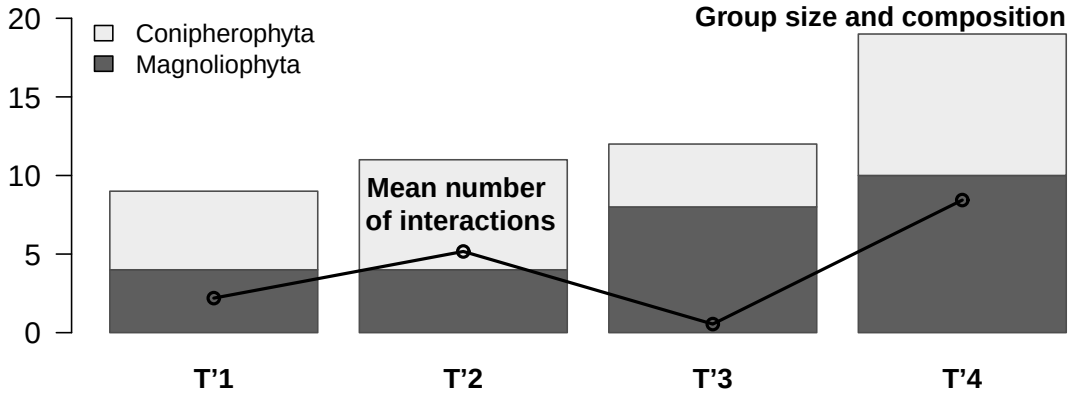
$\widehat{\lambda}_{q\ell}$	T'1	T'2	T'3	T'4
T'1	0.75	2.46	0.40	3.77
T'2	2.46	4.30	0.52	8.77
T'3	0.40	0.52	0.080	1.05
T'4	3.77	8.77	1.05	14.22
$\widehat{\alpha}_q$	17.7	21.5	23.5	37.3

Groups of Trees: With Covariate

Model: $X_{ij} \sim \mathcal{P}(\lambda_{ql} e^{\beta y_{ij}})$ with y_{ij} taxonomic distance

- $\hat{Q} = 4$ classes;
- $\hat{\beta} = -0.317$;

	T'1	T'2	T'3	T'4
T1	0	0	0	4
T2	0	0	0	4
T3	2	5	0	0
T4	0	2	0	5
T5	0	2	0	6
T6	0	0	10	0
T7	7	2	2	0

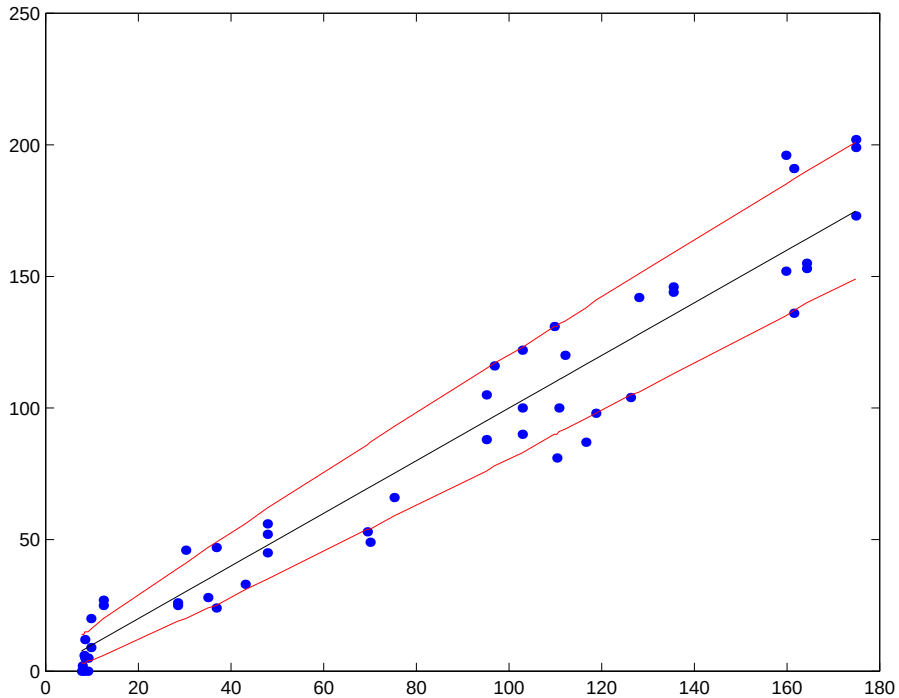


$\widehat{\lambda}_{q\ell}$	T'1	T'2	T'3	T'4
T'1	0.75	2.46	0.40	3.77
T'2	2.46	4.30	0.52	8.77
T'3	0.40	0.52	0.080	1.05
T'4	3.77	8.77	1.05	14.22
$\widehat{\alpha}_q$	17.7	21.5	23.5	37.3

Goodness of fit

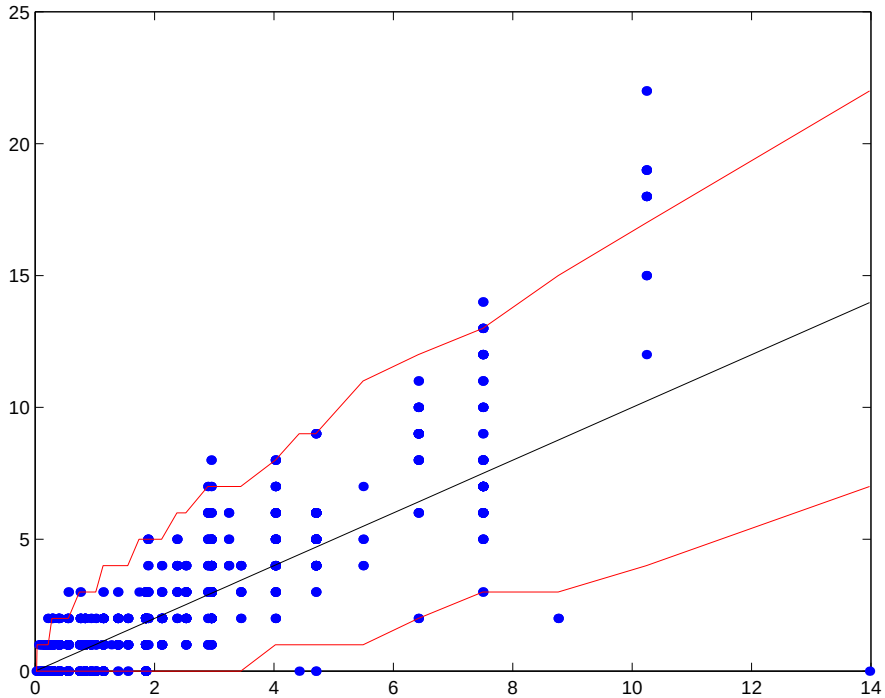
Check predictive power of the model for

Weighted degree



$$K_i = \sum_{j \neq i} X_{ij}$$

Single Edge Value



$$X_{ij}$$

Other covariates

- Genetic distance: same effect than taxonomic distance;
- Jaccard distance: no effect;

→ Main sources of similarity in trees parasitic assemblages are evolutionary processes and not ecological processes.

Tree interaction network				
Factor	Covariate	Q (PM)	Q (PRMH)	Δ ICL
Phylogenetic relatedness	Taxonomic Distance	7	4	116.0
	Genetic distance	7	4	94.8
Geographical overlap	Jaccard distance	7	7	-8.6

Table: Effect of covariates. Δ ICL = gain of switching from PM to PRMH.

Other covariates

- Genetic distance: same effect than taxonomic distance;
- Jaccard distance: no effect;

→ Main sources of similarity in trees parasitic assemblages are evolutionary processes and not ecological processes.

Tree interaction network				
Factor	Covariate	Q (PM)	Q (PRMH)	Δ ICL
Phylogenetic relatedness	Taxonomic Distance	7	4	116.0
	Genetic distance	7	4	94.8
Geographical overlap	Jaccard distance	7	7	-8.6

Table: Effect of covariates. Δ ICL = gain of switching from PM to PRMH.

Summary

MixNet

- Flexible probabilistic model to detect structure in complex valued graphs;
- Pseudo-likelihood estimators computed through variational EM (consistency ?);
- A statistical model selection criteria for the number of classes;
- Package available at `http://pbil.univ-lyon1.fr/software/MixNet`.

Host-Parasite Network

- Similarity in parasitic assemblages of two trees explained by phylogenetic relatedness, not geographical overlap.

Summary

MixNet

- Flexible probabilistic model to detect structure in complex valued graphs;
- Pseudo-likelihood estimators computed through variational EM (consistency ?);
- A statistical model selection criteria for the number of classes;
- Package available at `http://pbil.univ-lyon1.fr/software/MixNet`.

Host-Parasite Network

- Similarity in parasitic assemblages of two trees explained by phylogenetic relatedness, not geographical overlap.

E. Coli Reaction Network

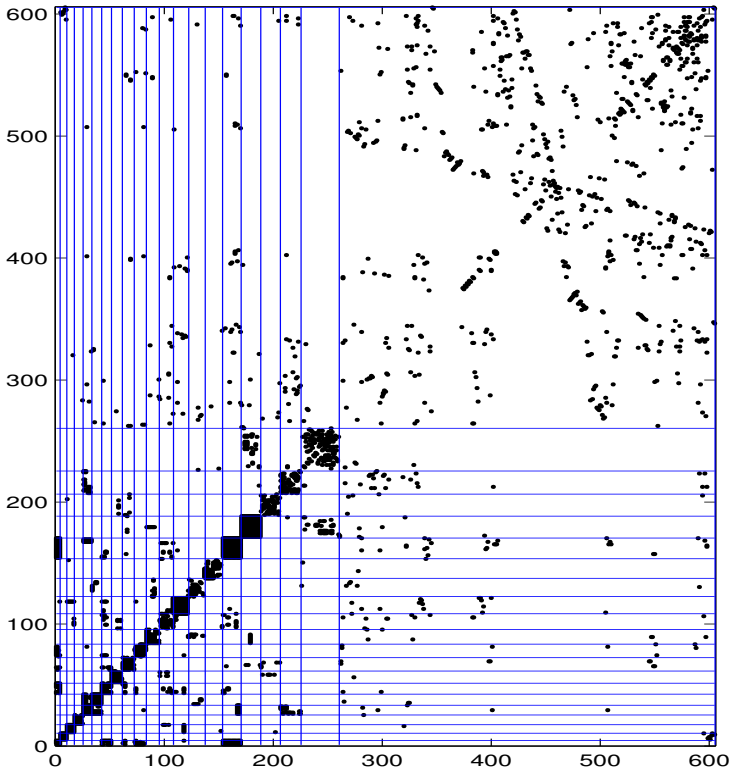
- **Reaction Network of E.Coli**:
  - → data from `http://www.biocyc.org/`,
  - → $n = 605$ vertices (reactions) and $1\,782$ edges.
  - → 2 reactions i and j are connected if the product of i is the substrate of j (cofactors excluded),
  - → V. Lacroix and M.-F. Sagot (INRIA - Hélix).

- **Question**:
  - → Interpretation of the connectivity structure of classes?

- **MixNet results**:
  - → ICL gives $\hat{Q} = 21$ classes,
  - → Most classes correspond to pseudo-cliques,

Biological interpretation of the groups I

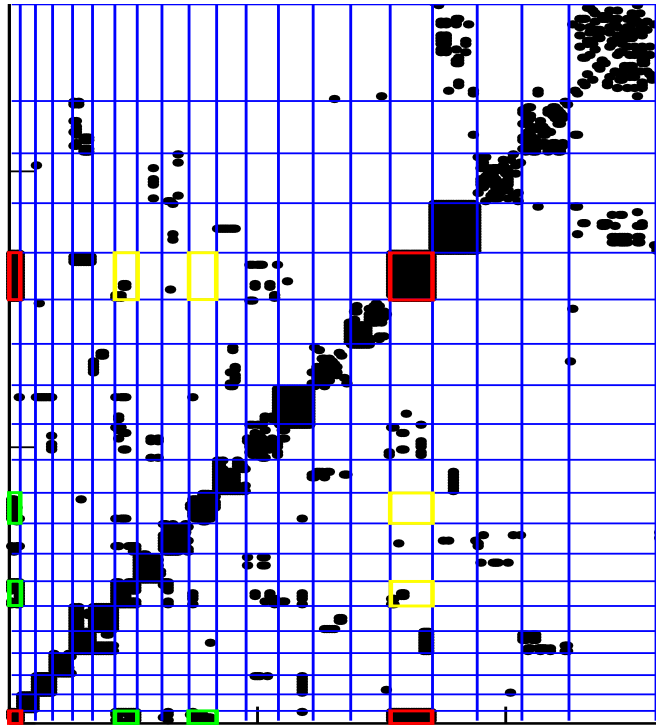
- Dot-plot representation
  - → adjacency matrix (sorted)
- Biological interpretation:
  - → Groups 1 to 20 gather reactions involving all the same compound either as a substrate or as a product,
  - → A compound (chorismate, pyruvate, ATP,*etc*) can be associated to each group.
- The structure of the metabolic network is governed by the compounds.



Biological interpretation of the groups II

- → Classes 1 and 16 constitute s single clique corresponding to a single compound (pyruvate),
- → They are split into two classes because they interact differently with classes 7 (CO2) and 10 (AcetylCoA)
- → Connectivity matrix (sample):

q, l	1	7	10	16
1	1.0			
7	.11	.65		
10	.43		.67	
16	1.0	.01	ϵ	1.0



Adjacency matrix (sample)