Uncovering latent structure in valued graphs: a variational approach
Mahendra Mariadassou, Stephane Robin, Corinne Vacher

To cite this version:

HAL Id: hal-02802815
https://hal.inrae.fr/hal-02802815
Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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
3. Parametric Estimation
4. Simulation Study
5. Ecological Network
Figure: Yeast PIN. source: www.bordalierinstitute.com/images/yeastProteinInteractionNetwork.jpg
Goal: Simple Representation of the Graph

Figure: Zachary’s karate club (Zachary 77)
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. ●, ▲, ■);
- $Z = (Z_i)_{i=1..n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \alpha)$ where $\alpha = (\alpha_1, \ldots, \alpha_Q)$ are the group proportions;
- $Z_i$ is not observed.

Example: (9 nodes, 3 classes)
Nodes heterogeneity

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

Example: (9 nodes, 3 classes)

- $P(●) = 0.25$
- $P(▲) = 0.5$
- $P(■) = 0.25$
MixNet Probabilistic Model (edges)

Observations

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

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

- $\theta = (\theta_{q\ell})_{q,\ell=1..Q}$ 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 $Z$, the $(X_{ij})$ are independent with distribution

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

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

Example: 3 classes with Poisson-valued edges

![Diagram of a network with three classes and corresponding edge values. The classes are represented by different shapes: triangles, circles, and squares. The edge values are shown as matrices: $\begin{pmatrix} 0 & 0.9 & 0.25 \\ . & 1 & 0.5 \\ . & . & 1 \end{pmatrix}$]
Classical Distributions:

- $f_\theta$ 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;
  $$X_{ij}|\{Z_{iq} = 1, Z_{j\ell} = 1\} \sim \mathcal{P}(\lambda_{q\ell} \exp\{\beta^\top y_{ij}\})$$
(Log)-Likelihood of the Model

► Complete data likelihood

\[
L(X, Z) = \ln \Pr(X, Z) = \ln \Pr(Z)P(X|Z) \\
= \sum_i \sum_q Z_{iq} \ln \alpha_q + \sum_{i<j} \sum_{q,l} Z_{iq}Z_{jl} \ln f_{\theta_{ql}}(X_{ij})
\]

► Observed data likelihood

\[
L(X) = \ln \sum_Z \exp L(X, Z)
\]

► Sum over \(Q^n\) is untractable, use EM algorithm instead.
(Log)-Likelihood of the Model

- Complete data likelihood

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

- Observed data likelihood

\[ \mathcal{L}(X) = \ln \sum_Z \exp \mathcal{L}(X, Z) \]

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

But...

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

→ Exact EM is not possible...
If $R_X$ is a distribution over $Z$, let

$$J(R_X) = L(X) - KL(R_X, Pr(.|X))$$

For $R_X = Pr(.|X)$, $J(R_X) = L(X)$;

Variational approximation: replace complicated distribution $Pr(.|X)$ by a simple $R_X$ such that $KL(R_X, Pr(.|X))$ is minimal to obtain a tight lower bound of $L(X)$.

$$J(R_X) = L(X) - KL(R_X, Pr(.|X))$$
$$= H(R_X) + E_{R_X}[L(X, Z)]$$

where $H(R_X)$ is the entropy of $R_X$. 
Variational Inference: Pseudo Likelihood

If $\mathcal{R}_X$ is a distribution over $\mathbf{Z}$, let

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

For $\mathcal{R}_X = \Pr(.|\mathbf{X})$, $\mathcal{J}(\mathcal{R}_X) = \mathcal{L}(\mathbf{X})$;

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

$$ \mathcal{J}(\mathcal{R}_X) = \mathcal{L}(\mathbf{X}) - KL(\mathcal{R}_X, \Pr(.|\mathbf{X})) \\
= \mathcal{H}(\mathcal{R}_X) + E_{\mathcal{R}_X}[\mathcal{L}(\mathbf{X}, Z)] $$

where $\mathcal{H}(\mathcal{R}_X)$ is the entropy of $\mathcal{R}_X$. 
If $R_X$ is a distribution over $Z$, let

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

For $R_X = Pr(.|X)$, $J(R_X) = \mathcal{L}(X)$;

**Variational approximation**: replace complicated distribution $Pr(.|X)$ by a simple $R_X$ such that $KL(R_X, Pr(.|X))$ is minimal to obtain a tight lower bound of $\mathcal{L}(X)$.

$$J(R_X) = \mathcal{L}(X) - KL(R_X, Pr(.|X))$$
$$= \mathcal{H}(R_X) + E_{R_X}[\mathcal{L}(X, Z)]$$

where $\mathcal{H}(R_X)$ is the entropy of $R_X$. 
Computing \( \mathbb{E}_{R_X}[L(X, Z)] \) is easy, computing \( \mathcal{H}(R_X) \) is hard (in general).

Restrict \( R_X \) to a comfortable class of distributions:

\[ R_X[Z] = \prod_i h(Z_i; \tau_i) \]

with \( h(.; \tau_i) \) the multinomial with parameter \( \tau_i = (\tau_{i1}, \ldots, \tau_{iQ}) \).

Intuitively, \( \tau_{iq} \approx \Pr(Z_{iq} = 1|X) \).

For such \( R_X \),

\[ J((\tau_i)_{i=1..n}) = - \sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{j\ell} \ln f_{\theta_q\ell}(X_{ij}) \]
Computing $\mathbb{E}_{\mathcal{R}_X}[\mathcal{L}(X, Z)]$ is easy, computing $\mathcal{H}(\mathcal{R}_X)$ is hard (in general).

Restrict $\mathcal{R}_X$ to a comfortable class of distributions:

$$\mathcal{R}_X[Z] = \prod_i h(Z_i; \tau_i)$$

with $h(.; \tau_i)$ the multinomial with parameter $\tau_i = (\tau_{i1}, \ldots, \tau_{iQ})$. Intuitively, $\tau_{iq} \approx \Pr(Z_{iq} = 1|X)$.

For such $\mathcal{R}_X$,

$$\mathcal{J}((\tau_i)_{i=1..n}) = - \sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$
Computing $\mathbb{E}_{\mathcal{R}_X}[\mathcal{L}(X, Z)]$ is easy, computing $\mathcal{H}(\mathcal{R}_X)$ is hard (in general).

Restrict $\mathcal{R}_X$ to a comfortable class of distributions:

$$\mathcal{R}_X[Z] = \prod_i h(Z_i; \tau_i)$$

with $h(.; \tau_i)$ the multinomial with parameter $\tau_i = (\tau_{i1}, \ldots, \tau_{iQ})$.

Intuitively, $\tau_{iq} \approx \Pr(Z_{iq} = 1|X)$.

For such $\mathcal{R}_X$, 

$$\mathcal{I}((\tau_i)_{i=1..n}) = - \sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{j\ell} \ln f_{\theta_{q\ell}}(X_{ij})$$
2 Steps Iterative Algorithm

- Maximize pseudo-likelihood:

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

- **Step 1** Optimize \( J \) w.r.t. \( (\tau_i) \):
  - Constraint: \( \sum_q \tau_{iq} = 1 \) for all \( i \);
  - \( \tau_{iq} \) 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 \( J \) w.r.t. \( (\alpha, \theta) \):
  - Constraint: \( \sum_q \alpha_q = 1 \)

\[ \tilde{\alpha}_q = \frac{\sum_i \tilde{\tau}_{iq}}{n} \]

\[ \tilde{\theta}_{q\ell} = \arg \max_{\theta} \sum_{i,j} \tilde{\tau}_{iq} \tilde{\tau}_{jl} \log f_{\theta}(X_{ij}) \]

- Simple expression of \( \tilde{\theta}_{q\ell} \) for classical distributions (weighted MLE).
2 Steps Iterative Algorithm

Maximize pseudo-likelihood:

\[ J((\alpha, \theta), (\tau_i)_{i=1..n}) = -\sum_{i} \sum_{q} \tau_{iq} \ln \tau_{iq} + \sum_{i} \sum_{q} \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{jl} \ln f_{\theta_{ql}}(X_{ij}) \]

Step 1 Optimize \( J \) w.r.t. \( (\tau_i) \):

→ Constraint: \( \sum_{q} \tau_{iq} = 1 \) for all \( i \);
→ \( \tau_{iq} \) 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 \( J \) w.r.t. \( (\alpha, \theta) \):

→ 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:

\[ J((\alpha, \theta), (\tau_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{jl} \ln f_{\theta_{ql}}(X_{ij}) \]

**Step 1 Optimize** \( J \) **w.r.t.** \((\tau_i)\):

→ Constraint: \( \sum_q \tau_{iq} = 1 \) for all \( i \);
→ \( \tau_{iq} \) variational parameter found via a fixed point algorithm:

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

**Step 2 Optimize** \( J \) **w.r.t.** \((\alpha, \theta)\):

→ 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).
**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:
  
  \[
  \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) \rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} = P_Q \log n(n - 1) \\
  \mathcal{L}(\mathbf{Z}|Q) \rightarrow \text{pen}_{\mathbf{Z}} = (Q - 1) \log(n)
  \]

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

- BIC-like criterion to select the number of classes;

- The likelihood can be split: \( \mathcal{L}(X, Z|Q) = \mathcal{L}(X|Z, Q) + \mathcal{L}(Z|Q) \);

- These terms can be penalized separately:
  \[
  \mathcal{L}(X|Z, Q) \rightarrow \text{pen}_{X|Z} P_Q \log n(n - 1)
  \]
  \[
  \mathcal{L}(Z|Q) \rightarrow \text{pen}_Z = (Q - 1) \log(n)
  \]

\[
ICL(Q) = \max_{\theta} \mathcal{L}(X, \hat{Z}|\theta, m_Q) - \frac{1}{2} \left( P_Q \log n(n - 1) - (Q - 1) \log(n) \right)
\]
## MixNet Properties

### Identifiability
- Identifiability 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: \( \Pr(\cdot|X) \to \delta_Z \) (ongoing work of Celisse and Daudin, Mariadassou and Matias)
MixNet Properties

Identifiability
- Identifiability 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: $\Pr(\cdot | X) \rightarrow \delta_Z$ (ongoing work of Celisse and Daudin, Mariadassou and Matias)
→ 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;
  - $\lambda$: 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) = $\sqrt{\text{Bias}^2 + \text{Variance}}$

RMSE for the $\alpha_q$  

RMSE for the $\lambda_{ql}$

$x$-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 $\gamma$: $\circ = 0.1$, $\triangledown = 0.5$, $\triangle = 0.9$, $\star = 1.5$
Undirected graph with \(Q^* = 3\) classes and Poisson edges;

\(n = 50, 100, 500, 1000\) vertices;

\(\alpha_q = (57.1\%, 28, 6\%, 14, 3\%)\);

Connectivity matrix of the form

\[
\begin{pmatrix}
2 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>(n)</th>
<th>(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>82</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
</tr>
</tbody>
</table>

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: $X_{tt'} = \text{number of fungis infecting both } t \text{ and } t'$.

- **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)

<table>
<thead>
<tr>
<th>$\hat{\lambda}_{ql}$</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>14.46</td>
<td>4.19</td>
<td>5.99</td>
<td>7.67</td>
<td>2.44</td>
<td>0.13</td>
<td>1.43</td>
</tr>
<tr>
<td>T2</td>
<td>4.19</td>
<td>14.13</td>
<td>0.68</td>
<td>2.79</td>
<td>4.84</td>
<td>0.53</td>
<td>1.54</td>
</tr>
<tr>
<td>T3</td>
<td>5.99</td>
<td>0.68</td>
<td>3.19</td>
<td>4.10</td>
<td>0.66</td>
<td>0.02</td>
<td>0.69</td>
</tr>
<tr>
<td>T4</td>
<td>7.67</td>
<td>2.79</td>
<td>4.10</td>
<td>7.42</td>
<td>2.57</td>
<td>0.04</td>
<td>1.05</td>
</tr>
<tr>
<td>T5</td>
<td>2.44</td>
<td>4.84</td>
<td>0.66</td>
<td>2.57</td>
<td>3.64</td>
<td>0.23</td>
<td>0.83</td>
</tr>
<tr>
<td>T6</td>
<td>0.13</td>
<td>0.53</td>
<td>0.02</td>
<td>0.04</td>
<td>0.23</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>T7</td>
<td>1.43</td>
<td>1.54</td>
<td>0.69</td>
<td>1.05</td>
<td>0.83</td>
<td>0.06</td>
<td>0.27</td>
</tr>
</tbody>
</table>

| $\hat{\alpha}_q$     | 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: $X_{ij} \sim P(\lambda_{q\ell}e^{\beta y_{ij}})$ with $y_{ij}$ taxonomic distance

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

<table>
<thead>
<tr>
<th></th>
<th>T'1</th>
<th>T'2</th>
<th>T'3</th>
<th>T'4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>T2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>T3</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>T5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>T6</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>T7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>T'1</th>
<th>T'2</th>
<th>T'3</th>
<th>T'4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T'1</td>
<td>0.75</td>
<td>2.46</td>
<td>0.40</td>
<td>3.77</td>
</tr>
<tr>
<td>T'2</td>
<td>2.46</td>
<td>4.30</td>
<td>0.52</td>
<td>8.77</td>
</tr>
<tr>
<td>T'3</td>
<td>0.40</td>
<td>0.52</td>
<td>0.080</td>
<td>1.05</td>
</tr>
<tr>
<td>T'4</td>
<td>3.77</td>
<td>8.77</td>
<td>1.05</td>
<td>14.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>T'1</th>
<th>T'2</th>
<th>T'3</th>
<th>T'4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\lambda}_{q\ell}$</td>
<td>17.7</td>
<td>21.5</td>
<td>23.5</td>
<td>37.3</td>
</tr>
</tbody>
</table>
Groups of Trees: With Covariate

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

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

<table>
<thead>
<tr>
<th></th>
<th>T'1</th>
<th>T'2</th>
<th>T'3</th>
<th>T'4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>T2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>T3</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>T5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>T6</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>T7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\lambda}_{q\ell}$</th>
<th>T'1</th>
<th>T'2</th>
<th>T'3</th>
<th>T'4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T'1</td>
<td>0.75</td>
<td>2.46</td>
<td>0.40</td>
<td>3.77</td>
</tr>
<tr>
<td>T'2</td>
<td>2.46</td>
<td>4.30</td>
<td>0.52</td>
<td>8.77</td>
</tr>
<tr>
<td>T'3</td>
<td>0.40</td>
<td>0.52</td>
<td>0.080</td>
<td>1.05</td>
</tr>
<tr>
<td>T'4</td>
<td>3.77</td>
<td>8.77</td>
<td>1.05</td>
<td>14.22</td>
</tr>
</tbody>
</table>

| $\hat{\alpha}_q$       | 17.7  | 21.5  | 23.5  | 37.3  |
Goodness of fit

Check predictive power of the model for

Weighted degree

Single Edge Value

\[ K_i = \sum_{j \neq i} X_{ij} \]
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.

<table>
<thead>
<tr>
<th>Tree interaction network</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
</tr>
<tr>
<td>Phylogenetic relatedness</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Geographical overlap</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Covariate</th>
<th>Q (PM)</th>
<th>Q (PRMH)</th>
<th>Δ ICL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenetic relatedness</td>
<td>Taxonomic Distance</td>
<td>7</td>
<td>4</td>
<td>116.0</td>
</tr>
<tr>
<td></td>
<td>Genetic distance</td>
<td>7</td>
<td>4</td>
<td>94.8</td>
</tr>
<tr>
<td>Geographical overlap</td>
<td>Jaccard distance</td>
<td>7</td>
<td>7</td>
<td>-8.6</td>
</tr>
</tbody>
</table>

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;

Host-Parasite Network

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

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:**
- \( n = 605 \) vertices (reactions) and \( 1782 \) 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.
Classes 1 and 16 constitute a 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):

<table>
<thead>
<tr>
<th>q, l</th>
<th>1</th>
<th>7</th>
<th>10</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.11</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.43</td>
<td></td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
<td>.01</td>
<td></td>
<td>1.0</td>
</tr>
</tbody>
</table>