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
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.
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)
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 $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \sim M(1, \alpha)$ where $\alpha = (\alpha_1, \ldots, \alpha_Q)$ 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 $Z$, 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
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
Flexibility of MixNet

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

► Complete data likelihood

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

► Observed data likelihood

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

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

- **Complete data likelihood**

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

- **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...
Variational Inference: Pseudo Likelihood

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

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

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

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

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

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

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

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

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

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

where $\mathcal{H}(\mathcal{R}_X)$ is the entropy of $\mathcal{R}_X$. 
Variational Inference: Pseudo Likelihood

- 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 $\mathcal{L}(X)$.

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

$$= H(R_X) + E_{R_X}[\mathcal{L}(X, Z)]$$

where $H(R_X)$ is the entropy of $R_X$. 
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, $\tau_{iq} \simeq \Pr(Z_{iq} = 1|X)$.

For such $\mathcal{R}_X$,

$$\mathcal{I}((\tau_i)_{i=1..n}) = - \sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{j\ell} \ln f_{\theta_q}(X_{ij})$$
Computing $E_{R_X}[\mathcal{L}(X,Z)]$ 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 $R_X$,

$$\mathcal{J}((\tau_i)_{i=1..n}) = -\sum_i \sum_q \tau_{iq} \ln \tau_{iq} + \sum_i \sum_q \tau_{iq} \ln \alpha_q + \sum_{i<j} \tau_{iq} \tau_{jq} \ln f_{\theta_{q\ell}}(X_{ij})$$
Computing $\mathbb{E}_{R_X}[\mathcal{L}(X, Z)]$ 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 $h(\cdot; \tau_i)$ the multinomial with parameter $\tau_i = (\tau_{i1}, \ldots, \tau_{iQ})$.

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

For such $R_X$,

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

- Maximize pseudo-likelihood:

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

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

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

- **Step 2** Optimize \( \mathcal{J} \) w.r.t. \( (\alpha, \theta) \):
  - Constraint: \( \sum_q \alpha_q = 1 \)
  - 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:

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

- **Step 2** Optimize \( J \) w.r.t. \((\alpha, \theta)\):
  - Constraint: \( \sum_q \alpha_q = 1 \)

\[
\tilde{\alpha}_q = \frac{\sum_i \tilde{\tau}_{iq}}{n} \\
\tilde{\theta}_{ql} = \arg\max_{\theta} \sum_{i,j} \tilde{\tau}_{iq} \tilde{\tau}_{jl} \log f_{\theta}(X_{ij})
\]

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

- **Maximize pseudo-likelihood:**

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

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

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

- **Step 2 Optimize \( \mathcal{J} \) 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).
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:

\[
\begin{align*}
\mathcal{L}(X|Z, Q) & \rightarrow \text{pen}_{X|Z} P_Q \log n(n - 1) \\
\mathcal{L}(Z|Q) & \rightarrow \text{pen}_Z = (Q - 1) \log(n)
\end{align*}
\]

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

\[
\begin{align*}
\mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) & \rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} P Q \log n(n - 1) \\
\mathcal{L}(\mathbf{Z}|Q) & \rightarrow \text{pen}_{\mathbf{Z}} = (Q - 1) \log(n)
\end{align*}
\]

\[
ICL(Q) = \max_{\theta} \mathcal{L}(\mathbf{X}, \tilde{\mathbf{Z}}|\theta, m_Q) - \frac{1}{2} \left( P Q \log n(n - 1) - (Q - 1) \log(n) \right)
\]
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)
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: $\Pr(.|X) \rightarrow \delta_{Z}$ (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;
  - $\lambda$: 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{Bias^2 + 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 \( \gamma \): \( \circ = 0.1, \nabla = 0.5, \triangle = 0.9, \star = 1.5 \)
Number of Classes

→ 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

$$
\begin{bmatrix}
2 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2 \\
\end{bmatrix}
$$

<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: $X_{tt'} = \text{number of fungi infecting both } t \text{ and } t'$.
- **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

![Bar chart showing mean number of interactions for different groups (T1 to T7).](chart)

- 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}_{q\ell} )</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>

\[
\hat{\lambda}_{q\ell} \quad T'1 \quad T'2 \quad T'3 \quad T'4 \\
T'1 | 0.75 | 2.46 | 0.40 | 3.77 \\
T'2 | 2.46 | 4.30 | 0.52 | 8.77 \\
T'3 | 0.40 | 0.52 | 0.080 | 1.05 \\
T'4 | 3.77 | 8.77 | 1.05 | 14.22 \\
\]

\[
\hat{\alpha}_q \quad 17.7 \quad 21.5 \quad 23.5 \quad 37.3
\]
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>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>
</tr>
<tr>
<td>T2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T3</td>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>T4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>T5</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>T6</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>T7</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Mean number of interactions

| Group size and composition |

<table>
<thead>
<tr>
<th>T’1</th>
<th>T’2</th>
<th>T’3</th>
<th>T’4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.75</td>
<td>2.46</td>
<td>0.40</td>
</tr>
<tr>
<td>T2</td>
<td>2.46</td>
<td>4.30</td>
<td>0.52</td>
</tr>
<tr>
<td>T3</td>
<td>0.40</td>
<td>0.52</td>
<td>0.080</td>
</tr>
<tr>
<td>T4</td>
<td>3.77</td>
<td>8.77</td>
<td>1.05</td>
</tr>
</tbody>
</table>

$\hat{\lambda}_{q\ell}$

<table>
<thead>
<tr>
<th>T’1</th>
<th>T’2</th>
<th>T’3</th>
<th>T’4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.75</td>
<td>2.46</td>
<td>0.40</td>
</tr>
<tr>
<td>T2</td>
<td>2.46</td>
<td>4.30</td>
<td>0.52</td>
</tr>
<tr>
<td>T3</td>
<td>0.40</td>
<td>0.52</td>
<td>0.080</td>
</tr>
<tr>
<td>T4</td>
<td>3.77</td>
<td>8.77</td>
<td>1.05</td>
</tr>
</tbody>
</table>

$\hat{\alpha}_q$

<table>
<thead>
<tr>
<th>T’1</th>
<th>T’2</th>
<th>T’3</th>
<th>T’4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>17.7</td>
<td>21.5</td>
<td>23.5</td>
</tr>
</tbody>
</table>

Mariadassou (INRA) Uncovering Structure in Valued Graphs May 11
Goodness of fit

Check predictive power of the model for

Weighted degree

Single Edge Value

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

\[ X_{ij} \]
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>Factor</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>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. $\Delta$ ICL = gain of switching from PM to PRMH.
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.
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.
Reaction Network of E.Coli:

- data from http://www.biocyc.org/,
- \( n = 605 \) 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):

\[
\begin{array}{c|cccc}
q, l & 1 & 7 & 10 & 16 \\
\hline
1 & 1.0 & & & \\
7 & .11 & .65 & & \\
10 & .43 & .67 & & \\
16 & 1.0 & .01 & \epsilon & 1.0 \\
\end{array}
\]