Catching up with big fish in the big pond? Multi-level network analysis through linked design
Emmanuel Lazega, Marie-Thérèse Jourda, Lise Mounier, Rafaël Stofer
*Social Networks*, **30**:2, 159-176 (2008)

- "Elite", of French cancer researchers at the end of 1990s
- Among the 168 researchers, 128 persons (76%) accepted an interview
- Description of researchers: age, speciality, laboratory, performance, status
- Description of lab: city, # researchers,
- Inter-individual connections
- Inter-lab connections
Few remarks on the data

- Hierarchical: Labs/Researchers
- Graph
- Sparse
- Missing data
- Not bipartite graph
Small/big fish: Indegree centrality

Small/big pond

- Indegree centrality in inter-organizational networks
- Outdegree (indicating the potential resources to which its director declares having access)
- Size

A laboratory is a "big pond" if its values were above the median for at least two of these criteria.

Big Fish in a Big Pond: researcher’s indegree centrality must be higher than 5.2, that of the laboratory higher than 2.75; the laboratory’s outdegree must be higher than 2 and its size higher than 26 researchers.
Z_i = q: vertex i belongs to class q (Q classes). Ties are independent given the class memberships.

\[ X_{ij} | \{Z_i = q, Z_j = l\} \sim \mathcal{B}(\pi_{ql}) \]

Mixture model for graphs: \( X_{ij} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \alpha_q \alpha_l \mathcal{B}(\pi_{ql}) \)

Computations are performed by \texttt{wmixnet}

estimation of \( \alpha \)'s, \( \pi \)'s and latent group (probability of appartenance).
4 groups selected (group ≠ community).

Study on fully observed data ⇒ 95 researchers (from 76 labs).

Groups are not clearly linked to specialities.

<table>
<thead>
<tr>
<th></th>
<th>G 1</th>
<th>G 2</th>
<th>G 3</th>
<th>G 4</th>
<th>Sums</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Fish</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>33</td>
<td>45</td>
</tr>
<tr>
<td>Big Fish</td>
<td>16</td>
<td>12</td>
<td>17</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Sums</td>
<td>24</td>
<td>12</td>
<td>21</td>
<td>38</td>
<td>95</td>
</tr>
</tbody>
</table>
A multiplex stochastic block model for social networks. The figure on the left shows a researcher network with 4 groups, while the figure on the right displays a meta-researcher network (edges displayed if $\hat{\pi} \geq 0.1$). The matrix $\hat{\pi}$ is given by:

$$
\hat{\pi} = \begin{pmatrix}
0.08 & 0.08 & 0.01 & 0.00 \\
0.21 & 0.32 & 0.17 & 0.06 \\
0.01 & 0.08 & 0.36 & 0.05 \\
0.01 & 0.02 & 0.04 & 0.04 \\
\end{pmatrix}
$$
Stochastic Block Model with covariates

\[ Z_i = q : \text{vertex } i \text{ belongs to class } q \ (Q \text{ classes}) \]

\[ X_{ij} | \{Z_i = q, Z_j = l\} \sim \mathcal{B}(\pi_{ql}) \]

Covariates can be included, in that case:

\[ X_{ij} | \{Z_i = q, Z_j = l, V_{ij}\} \sim \mathcal{B}(g(\mu_{ql} + \beta_{ql}^T V_{ij})) \]

where \( g(x) = (1 + \exp(-x))^{-1} \) and \( \beta \) may depend on \( q \) and \( l \).

Mixture model for graphs: \( X_{ij} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \alpha_q \alpha_l \mathcal{B}(g(\mu_{ql} + \beta_{ql}^T V_{ij})) \)

In SBM, Ties are independent given the class memberships.
SBM with covariates on Researcher network

- with 5 covariates (describing edge or vertices: sender → receiver):
  1. Same speciality (0/1),
  2. Status of the sender (1/0: Director/not director),
  3. Status of the receiver (1/0: Director/not director),
  4. Lab relation (0/1): sender’s lab → receiver’s lab,
  5. Lab relation (0/1): receiver’s lab → sender’s lab.

- Lab network level is taken into account thanks to the covariates!

- Performance is not selected as a relevant covariate to explain edges.

- $\beta$ does not depend on groups.

$$X_{ij} | \{Z_i = q, Z_j = l, V_{ij}\} \sim \mathcal{B}(g(\mu_q l + \beta^T V_{ij}))$$
3 groups selected

\[ \hat{\beta} = \begin{pmatrix} 1.16 & 0.33 & 0.27 & 1.75 & 1.71 \end{pmatrix} \]

Lab effect is larger than Status effect

no directional effect
Groups obtained in SBM with covariates are quite different from groups in SBM without covariates.

<table>
<thead>
<tr>
<th></th>
<th>G 1</th>
<th>G 2</th>
<th>G 3</th>
<th>Sums</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Pond</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Fish</td>
<td>16</td>
<td>0</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Big Fish</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Big Pond</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Fish</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>Big Fish</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>Sums</td>
<td>55</td>
<td>10</td>
<td>30</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>G 1</td>
<td>G 2</td>
<td>G 3</td>
<td>Sums</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td><strong>Small Pond</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Fish</td>
<td>16</td>
<td>0</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Big Fish</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td><strong>Big Pond</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Fish</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>Big Fish</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td><strong>Sums</strong></td>
<td>55</td>
<td>10</td>
<td>30</td>
<td>95</td>
</tr>
</tbody>
</table>

“Average connections”

\[
\begin{pmatrix}
0.03 & 0.06 & 0.02 \\
0.04 & 0.53 & 0.06 \\
0.02 & 0.05 & 0.19 \\
\end{pmatrix}
\]

**Figure**: Meta-Researcher-network (edge displayed if $\pi \geq 0.05$)
SBM conclusions

- 3 researchers groups selected (sizes : 55 - 10 - 30).
- Relevant covariates to explain edges:
  - same speciality,
  - Status (R & S),
  - Lab links.
- Performance is NOT relevant.
- Posterior proba to be in a given group are mostly close to 99%.
Multiplex: model

∀ (i, j) ∈ \{1, \ldots n\}^2, i \neq j, ∀ (\delta_{ij}, \delta'_{ij}) \in \{0, 1\}^2,

\delta_{ij}: researcher's connections
\delta'_{ij}: lab's connections

\Pr(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij}) = \pi_{\delta_{ij}\delta'_{ij}} = \pi_{11}\pi_{01}\pi_{10}\pi_{00}\delta_{ij}\delta'_{ij}(1-\delta_{ij})(1-\delta'_{ij})(1-\delta_{ij})(1-\delta'_{ij})

where \pi_{00} + \pi_{11} + \pi_{01} + \pi_{10} = 1.

\Pr(X_{ij} = \delta_{ij}) = (\pi_{10} + \pi_{11})\delta_{ij}(\pi_{00} + \pi_{01})^{1-\delta_{ij}}.

\Pr(X_{ij} = \delta_{ij} | X'_{ij} = \delta'_{ij}) = \left(\frac{\pi_{1,\delta'_{ij}}}{\pi_{1,\delta'_{ij}} + \pi_{0,\delta'_{ij}}}\right)^{\delta_{ij}} \left(\frac{\pi_{0,\delta'_{ij}}}{\pi_{1,\delta'_{ij}} + \pi_{0,\delta'_{ij}}}\right)^{(1-\delta_{ij})}

Components of (X_{ij}, X'_{ij}) are independent if and only if

\log\left(\frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}}\right) = 0 \iff \pi_{00}\pi_{11} = \pi_{10}\pi_{01}.

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet  
Multiplex stochastic bloc model for social networks
Multiplex: estimation

Maximum likelihood estimates:

\[ \hat{\pi}_{00} = \frac{S_{00}}{n(n-1)}, \quad \hat{\pi}_{01} = \frac{S_{01}}{n(n-1)}, \quad \hat{\pi}_{10} = \frac{S_{10}}{n(n-1)}, \quad \hat{\pi}_{11} = \frac{S_{11}}{n(n-1)}. \]

\[
S_{00} = \sum_{i,j} (1 - X_{ij})(1 - X'_{ij}), \quad S_{01} = \sum_{i,j} (1 - X_{ij})X'_{ij}, \\
S_{10} = \sum_{i,j} X_{ij}(1 - X'_{ij}), \quad S_{11} = \sum_{i,j} X_{ij}X'_{ij}.
\]
Multiplex stochastic block model

\[ P(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij}) = \pi \delta_{ij} \delta'_{ij} = \pi_{11} \delta_{ij} \delta'_{ij} + \pi_{01} (1 - \delta_{ij}) \delta'_{ij} + \pi_{10} \delta_{ij} (1 - \delta'_{ij}) + \pi_{00} (1 - \delta_{ij}) (1 - \delta'_{ij}) \]

\[ P(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij} | Z_i = q, Z_j = l) = \pi_{q}^{q l} \delta_{ij} \delta'_{ij} \]

\[ = \left( \pi_{11}^{q l} \right) \delta_{ij} \delta'_{ij} \left( \pi_{01}^{q l} \right) (1 - \delta_{ij}) \delta'_{ij} \left( \pi_{10}^{q l} \right) \delta_{ij} (1 - \delta'_{ij}) \left( \pi_{00}^{q l} \right) (1 - \delta_{ij}) (1 - \delta'_{ij}) \]

\[ P(Z_i = q) = \alpha_q \]

where \( \sum_{q=1}^{Q} \alpha_q = 1 \) and \( \forall (q, l) \in \{1, \ldots Q\}^2, \pi_{00}^{q l} + \pi_{11}^{q l} + \pi_{01}^{q l} + \pi_{10}^{q l} = 1. \)
Multiplex stochastic block model

- Introduction of covariates:
  \[
  \logit \pi_{\delta_{ij},\delta'_{ij}}^{ql} = \mu_{\delta_{ij}^{1}...\delta_{ij}^{K}} + (\beta_{\delta_{ij}^{1}...\delta_{ij}^{K}})^\top y_{ij}
  \]
  but the number of parameters drastically increases

- For \( n \) large or large number of groups (Q), likelihood is not tractable: variational EM to maximize likelihood (\( \approx \) like SBM)

- Multivariate (\( K > 2 \)) Bernouilli.

- Number of groups chosen with penalized likelihood. Penalization:
  \[
  -\frac{1}{2} \left\{ Q^2 (2^K - 1) \log(Kn(n - 1)) + (Q - 1) \log n \right\} .
  \]
Application to social network: marginals

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet

Multiplex stochastic bloc model for social networks
Application to social network: Researchers

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet

Multiplex stochastic bloc model for social networks
Application to social network: Labs

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet
What about the groups?

- Performance1 < 28.56
- Performance1 < 18.17
- specialities = ef
- Laboratory.size < 16
- Laboratory.size >= 37

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet

Multiplex stochastic bloc model for social networks

32 links

477 users (including 325 that contribute for only one of the 32 articles)

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet

Multiplex stochastic bloc model for social networks
On going work: Collective experience in a rugby team

Is collective experience important for a rugby match

- number of common selection for each pairs of players
- National club for each players
- Opponents, results of the match
On going work: Collective experience in a rugby team

Holders in 2013 (with more than one selection)

Avner Bar-Hen, Pierre Barbillon, Sophie Donnet

Multiplex stochastic bloc model for social networks
Arising questions

- What is the sensitivity of the SBM grouping method if
  - a new edge appears in the graph?
  - a researcher’s lab changes?

- Are we able to detect the edge that should be set to maximize a criterion such that the mean performance of the researchers, the maximum performance...

- Scalability
Thank you for your attention (and your questions)
Since $X_{i,j}|\{Z_i = q, Z_j = l\} \sim \mathcal{B}(\pi_{ql})$ we have

$$X_{i,j} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \alpha_q \alpha_l \mathcal{B}(\pi_{ql})$$

Let $\theta = (\pi_{i,j}, \alpha_i)$. We are looking for

$$\hat{\theta} = \arg\max_{\theta} \log P(X, \theta)$$

but $P(X, \theta) = \sum_Z P(X, Z, \theta)$ is not tractable.

Classical decomposition (E-M trick)

$$\log P(X; \theta) = \log P(X, Z; \theta) - \log P(Z|X; \theta)$$

$$\mathbb{E}(\log P(X; \theta)|X) = \log P(X; \theta) = \mathbb{E}(\log P(X, Z; \theta)|X) - \mathbb{E}(\log P(Z|X; \theta)|X)$$
Variational EM

\[
\log P(X; \theta) = \mathbb{E}(\log P(X, Z; \theta)|X) - \mathbb{E}(\log P(Z|X; \theta)|X)
\]

- **E-step**: Calculation of \( P(Z|X; \hat{\theta}) \) (difficult: forward-backward recursion)

- **M-step**: \( \max_\theta \mathbb{E}(\log P(X, Z; \theta)|X) \) (similar to MLE)

Variational approximation: replace \( P(Z|X; \theta) \) with approximate distribution \( Q(Z) \) (\( Q(z) \) within a class of "good" distribution)

For any \( Q(z) \)

\[
\log P(X; \theta) \geq \log P(X; \theta) - KL(Q(Z), P(Z|X)) \quad (1)
\]
\[
= \mathbb{E}_Q (\log P(X, Z; \theta)) - \mathbb{E}_Q (\log Q(Z)) \quad (2)
\]

- **M-step**: \( \arg\max_\theta \mathbb{E}_Q^* (\log P(X, Z; \theta)) \)

- **E-step**: Replace calculation of \( P(Z|X; \hat{\theta}) \) with the search of

\[
Q^* = \arg\min_{Q(Z)} KL(Q(Z), P(Z|X))
\]
For any $Q(z)$

$$\log P(X; \theta) \geq \log P(X; \theta) - KL(Q(Z), P(Z|X))$$  \hspace{1cm} (3)

$$= E_Q(\log P(X, Z; \theta)) - E_Q(\log Q(Z))$$  \hspace{1cm} (4)

$Q(Z)$ within the set of factorisable distributions, ie

$Q(Z_i, Z_j) = Q(Z_i)Q(Z_j)$ (mean field approximation)

Fast to compute but not ML estimates...