Uncovering Latent Structure in Valued Graphs

M. Mariadassou, S. Robin

UMR AgroParisTech/INRA MIA 518, Paris

ECCS07, October 2007
Outline

1. Motivations

2. An Explicit Random Graph Model
   - Some Notations
   - Explicit Random Graph Model

3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative Algorithm
   - Model Selection Criterion

4. Simulation Study
   - Quality of the estimates
   - Number of Classes
Motivations for the study of networks

Networks... 

- Arise in many fields: 
  → Biology, Chemistry 
  → Physics, Internet.

- Represent an interaction pattern: 
  → $O(n^2)$ interactions 
  → between $n$ elements.

- Have a topology which: 
  → reflects the structure/function relationship

From Barabási website
Some Notations

- **Notations:**
  - $V$ a set of vertices in $\{1, \ldots, n\}$;
  - $E$ a set of edges in $\{1, \ldots, n\}^2$;
  - $X = (X_{ij})$ the adjacency matrix, with $X_{ij}$ the value of the edge between $i$ and $j$.

- **Random graph definition:**
  - To describe the network, we need the joint distribution of the $X_{ij}$.

- **Example:**

  - $V = \{1, 2, 3\}$
  - $E = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$
  - 

    $\begin{pmatrix}
    \cdot & 4 & 1 \\
    \cdot & \cdot & 2 \\
    \cdot & \cdot & \cdot
    \end{pmatrix}$
Some Notations

- **Notations:**
  - $V$ a set of vertices in $\{1, \ldots, n\}$;
  - $E$ a set of edges in $\{1, \ldots, n\}^2$;
  - $X = (X_{ij})$ the adjacency matrix, with $X_{ij}$ the value of the edge between $i$ and $j$.

- **Random graph definition:**
  - To describe the network, we need the joint distribution of the $X_{ij}$.

- **Example:**

  ![Graph Example](image)

  $V = \{1, 2, 3\}$

  $E = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$

  $\begin{pmatrix}
    \cdot & 4 & 1 \\
    \cdot & \cdot & 2 \\
    \cdot & \cdot & \cdot
  \end{pmatrix}$
Explicit Random Graph Model (vertices)

- **Vertices heterogeneity**
  - Hypothesis: the vertices are distributed among $Q$ classes with different connectivity;
  - $Z = (Z_i); \ Z_{iq} = 1\{i \in q\}$ are indep. hidden variables;
  - $\alpha = \{\alpha_q\}$, the *prior* proportions of groups;
  - $(Z_i) \sim M(1, \alpha)$.

- **Example:**
  - Example for 8 nodes and 3 classes with $\alpha = (0.25, 0.25, 0.5)$
Explicit Random Graph Model (vertices)

- **Vertices heterogeneity**
  - Hypothesis: the vertices are distributed among $Q$ classes with different connectivity;
  - $Z = (Z_i)_i; Z_{iq} = 1\{i \in q\}$ are indep. hidden variables;
  - $\alpha = \{\alpha_q\}$, the *prior* proportions of groups;
  - $(Z_i) \sim M(1, \alpha)$.

- **Example:**
  - Example for 8 nodes and 3 classes with $\alpha = (0.25, 0.25, 0.5)$
Explicit Random Graph Model (vertices)

- **Vertices heterogeneity**
  - Hypothesis: the vertices are distributed among $Q$ classes with different connectivity;
  - $Z = (Z_i)_i; Z_{iq} = 1\{i \in q\}$ are indep. hidden variables;
  - $\alpha = \{\alpha_q\}$, the prior proportions of groups;
  - $(Z_i) \sim M(1, \alpha)$.

- **Example:**
  - Example for 8 nodes and 3 classes with $\alpha = (0.25, 0.25, 0.5)$
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: $X_{ij}|\{i \in q, j \in \ell\} \sim f(., \theta_{q\ell})$;
  - $\theta = (\theta_{q\ell})$ is the connectivity parameter matrix;
  - ERMG: "Erdös-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;
Explicit Random Graph Model

- **X distribution**
  - Conditional distribution: \( X_{ij} \mid \{i \in q, j \in \ell \} \sim f(., \theta_{q\ell}); \)
  - \( \theta = (\theta_{q\ell}) \) is the connectivity parameter matrix;
  - ERMG: "Erdös-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;

\[
\begin{pmatrix}
0 & 0.9 & 0.25 \\
0.1 & 1 & 0.5 \\
0.1 & 0.5 & 1
\end{pmatrix}
\]
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: $X_{ij}|\{i \in q, j \in \ell\} \sim f(., \theta_{q\ell})$;
  - $\theta = (\theta_{q\ell})$ is the connectivity parameter matrix;
  - ERMG: "Erdős-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;

\[
\begin{pmatrix}
0 & 0.9 & 0.25 \\
.1 & 0 & .5 \\
. & . & 1
\end{pmatrix}
\]
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: $X_{ij}\mid\{i \in q, j \in \ell\} \sim f(., \theta_{q\ell})$;
  - $\theta = (\theta_{q\ell})$ is the connectivity parameter matrix;
  - ERMG: "Erdös-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;
  - Matrix for $\theta$:
    
    $\begin{pmatrix}
    0 & 0.9 & 0.25 \\
    .1 & 0.5 & .
    \\
    . & . & 1
    \end{pmatrix}$
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: $X_{ij}|\{i \in q, j \in \ell\} \sim f(\cdot, \theta_{q\ell})$;
  - $\theta = (\theta_{q\ell})$ is the connectivity parameter matrix;
  - ERMG: "Erdös-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;

![Graph Example Diagram]

- Connectivity parameter matrix:
  - $\begin{pmatrix} 0 & 0.9 & 0.25 \\ \cdot & 1 & 0.5 \\ \cdot & \cdot & 1 \end{pmatrix}$
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: $X_{ij}\{i \in q, j \in \ell\} \sim f(., \theta_{q\ell})$;
  - $\theta = (\theta_{q\ell})$ is the connectivity parameter matrix;
  - ERMG: "Erdös-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;
Explicit Random Graph Model (edges)

- **X distribution**
  - conditional distribution: \( X_{ij}|\{i \in q, j \in \ell \} \sim f(., \theta_{q\ell}) \);
  - \( \theta = (\theta_{q\ell}) \) is the connectivity parameter matrix;
  - ERMG: "Erdős-Rényi Mixture for Graphs".

- **Example:**
  - Example for 3 classes with Bernoulli-valued edges;
    - Example for 3 classes with Bernoulli-valued edges;
Random Edge Values

**Classical Distributions:**
- $f(., \theta_q \ell)$ can be any probability distribution;
- Bernoulli: presence/absence of an edge;
- Multinomial: nature of the connection (friend, lover, colleague);
- Poisson: in coauthorship networks, number of copublished papers;
- Gaussian: intensity of the connection (airport network);
- Bivariate Gaussian: directed networks where forward and backward edges are correlated;
- Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - \( f(., \theta_{q\ell}) \) can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - $f(., \theta_q \ell)$ can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - $f(., \theta_{q\ell})$ can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - \( f(., \theta_{q\ell}) \) can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - $f(., \theta_q \ell)$ can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - $f(., \theta_{q\ell})$ can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Random Edge Values

- **Classical Distributions:**
  - $f(., \theta_q \ell)$ can be any probability distribution;
  - Bernoulli: presence/absence of an edge;
  - Multinomial: nature of the connection (friend, lover, colleague);
  - Poisson: in coauthorship networks, number of copublished papers;
  - Gaussian: intensity of the connection (airport network);
  - Bivariate Gaussian: directed networks where forward and backward edges are correlated;
  - Etc.

Mixture Model to easily generate graphs
Log-Likelihood of the model

**First Idea:** Use maximum likelihood estimators

- **Complete data likelihood**

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

  with \( f_{\theta_{q\ell}}(X_{ij}) \) likelihood of edge value \( X_{ij} \) under \( i \sim q \) and \( j \sim \ell \).

- **Observed data likelihood**

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

  The observed data likelihood requires a sum over \( Q^n \) terms, and is thus **untractable**;

- EM-like strategies require the knowledge of \( \Pr(Z|X) \), also untractable (no conditional independence) and thus also fail.
Log-Likelihood of the model

**First Idea:** Use maximum likelihood estimators

- **Complete data likelihood**

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

with \(f_{\theta_{q\ell}}(X_{ij})\) likelihood of edge value \(X_{ij}\) under \(i \sim q\) and \(j \sim \ell\).

- **Observed data likelihood**

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

The observed data likelihood requires a sum over \(Q^n\) terms, and is thus untractable;

EM-like strategies require the knowledge of \(\Pr(Z|X)\), also untractable (no conditional independence) and thus also fail.
Log-Likelihood of the model

**First Idea:** Use maximum likelihood estimators

- **Complete data likelihood**

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

with \(f_{\theta q\ell}(X_{ij})\) likelihood of edge value \(X_{ij}\) under \(i \sim q\) and \(j \sim \ell\).

- **Observed data likelihood**

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

The observed data likelihood requires a sum over \(Q^n\) terms, and is thus untractable;

EM-like strategies require the knowledge of \(\Pr(Z|X)\), also untractable (no conditional independence) and thus also fail.
Main Idea: Replace *complicated* $\Pr(Z|X)$ by a *simple* $R_X[Z]$ such that $KL(R_X[Z], \Pr(Z|X))$ is minimal.

Optimize in $R_X$ the function $J(R_X)$ given by:

$$J(R_X[Z]) = L(X) - KL(R_X[Z], \Pr(Z|X))$$
$$= H(R_X[Z]) - \sum_Z R_X[Z]L(X, Z)$$

At best, $R_X = \Pr(Z|X)$ and

$$J(R_X[Z]) = L(X);$$

For simple $R_X$, $J(R_X[Z])$ is tractable.
Variational Inference: Pseudo Likelihood

Main Idea: Replace complicated $\Pr(Z|X)$ by a simple $R_X[Z]$ such that $KL(R_X[Z], \Pr(Z|X))$ is minimal.

- Optimize in $R_X$ the function $J(R_X)$ given by:

$$J(R_X[Z]) = L(X) - KL(R_X[Z], \Pr(Z|X)) = H(R_X[Z]) - \sum_Z R_X[Z] L(X, Z)$$

- At best, $R_X = \Pr(Z|X)$ and $J(R_X[Z]) = L(X)$;

- For simple $R_X$, $J(R_X[Z])$ is tractable.
Variational Inference: Pseudo Likelihood

**Main Idea:** Replace complicated $\Pr(Z|X)$ by a simple $R_X[Z]$ such that $KL(R_X[Z], \Pr(Z|X))$ is minimal.

- Optimize in $R_X$ the function $J(R_X)$ given by:

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

- At best, $R_X = \Pr(Z|X)$ and $J(R_X[Z]) = \mathcal{L}(X)$;

- For simple $R_X$, $J(R_X[Z])$ is tractable.
2 Step Algorithm

- **Step 1** Optimize $\mathcal{J}(R_X[Z])$ w.r.t. $R_X[Z]$:
  - Restriction to a "comfortable" class of functions;
  - $R_X[Z] = \prod_i h(Z_i; \tau_{i,X})$, with $h(.; \tau_{i,X})$ the multinomial distribution;
  - $\tau_{iq,X}$ is a variational parameter to be optimized using a fixed point algorithm:
    \[
    \tilde{\tau}_{iq,X} \propto \alpha_q \prod \prod \ell=1 f_{\theta_q\ell}(X_{ij}) \tilde{\tau}_{j\ell,X}
    \]

- **Step 2** Optimize $\mathcal{J}(R_X[Z])$ w.r.t. $(\alpha, \theta)$:
  - Constraint: $\sum_q \alpha_q = 1$
    \[
    \tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq,X}/n
    \tilde{\theta}_{q\ell} = \arg \max_{\theta} \sum_{ij} \tilde{\tau}_{iq,X} \tilde{\tau}_{j\ell,X} \log f_{\theta}(X_{ij})
    \]
  - Closed expression of $\tilde{\theta}_{q\ell}$ for classical distributions.
2 Step Algorithm

- **Step 1** Optimize $\mathcal{J}(R_X[Z])$ w.r.t. $R_X[Z]$:  
  - Restriction to a "comfortable" class of functions; 
  - $R_X[Z] = \prod_i h(Z_i; \tau_i,X)$, with $h(\cdot; \tau_i,X)$ the multinomial distribution; 
  - $\tau_{iq,X}$ is a variational parameter to be optimized using a fixed point algorithm:
    \[
    \tilde{\tau}_{iq,X} \propto \alpha_q \prod_{j \neq i} Q \prod_{\ell=1} f_{\theta_q \ell}(X_{ij}) \tilde{\tau}_{j \ell,X}
    \]

- **Step 2** Optimize $\mathcal{J}(R_X[Z])$ w.r.t. $(\alpha, \theta)$:  
  - Constraint: $\sum_q \alpha_q = 1$ 
    \[
    \tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq,X} / n \\
    \tilde{\theta}_{q\ell} = \arg\max_{\theta} \sum_{ij} \tilde{\tau}_{iq,X} \tilde{\tau}_{j\ell,X} \log f_\theta(X_{ij})
    \]
  - Closed expression of $\tilde{\theta}_{q\ell}$ for classical distributions.
Model Selection Criterion

- We derive a statistical 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:

  \[
  \mathcal{L}(X|Z, Q) \rightarrow \text{pen}_{X|Z} = \frac{Q(Q + 1)}{2} \log \frac{n(n - 1)}{2}
  \]

  \[
  \mathcal{L}(Z|Q) \rightarrow \text{pen}_Z = (Q - 1) \log(n)
  \]

  \[
  ICL(Q) = \max_{\theta} \mathcal{L}(X, \tilde{Z}|\theta, m_Q) - \frac{1}{2} \left( \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - (Q - 1) \log(n) \right)
  \]
We derive a statistical 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:

\[
\mathcal{L}(X|Z, Q) \rightarrow \text{pen}_{X|Z} = \frac{Q(Q + 1)}{2} \log \frac{n(n - 1)}{2}
\]

\[
\mathcal{L}(Z|Q) \rightarrow \text{pen}_{Z} = (Q - 1) \log(n)
\]

\[
ICL(Q) = \max_{\theta} \mathcal{L}(X, \tilde{Z}|\theta, m_Q) - \frac{1}{2} \left( \frac{Q(Q + 1)}{2} \log \frac{n(n - 1)}{2} - (Q - 1) \log(n) \right)
\]
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);
  • $\gamma <> 1$: classes are different;
  • $\lambda$: mean value of an edge;
→ 100 repeats for each setup.
Simulation Study

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);
  
  ● $\gamma <> 1$: classes are different;
  
  ● $\lambda$: mean value of an edge;

→ 100 repeats for each setup.
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);
  • $\gamma <> 1$: classes are different;
  • $\lambda$: mean value of an edge;

→ 100 repeats for each setup.
Simulation Study

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);
  - $\gamma <> 1$: classes are different;
  - $\lambda$: mean value of an edge;

→ 100 repeats for each setup.
Simulation Study

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);
- $\gamma <> 1$: classes are different;
- $\lambda$: mean value of an edge;

→ 100 repeats for each setup.
Results

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

- **Root Mean Square Error (RMSE)**

  \[ \text{RMSE} = \sqrt{\text{Bias}^2 + \text{Variance}} \]

  **RMSE for the } \alpha_q \text{**
  
  \[-\text{x-axis: } \alpha_1, \alpha_2, \alpha_3 \]

  **RMSE for the } \lambda_{ql} \text{**
  
  \[-\text{x-axis: } \lambda_{11}, \lambda_{22}, \lambda_{33}, \lambda_{12}, \lambda_{13}, \lambda_{23} \]

\[(n, \lambda, \gamma, a) \text{ from left (hard) to right (easy):}\]

\[(100, 2, 0.9, 0.2), (100, 2, 0.5, 0.5), (500, 5, 0.1, 1)\]
Simulation Setup and Results

→ Undirected graph with $Q^* = 3$ classes;
→ Poisson-valued edges;
→ $n = 50, 100, 500, 1000$ vertices;
→ $\alpha_q = (57.1\%, 28, 6\%, 14, 3\%)$ (or $a = 0.5$);
→ $\lambda = 2, \gamma = 0.5$;
→ Retrieve $Q$ that maximizes ICL;
→ 100 repeats for each value of $n$;

<table>
<thead>
<tr>
<th>$n$</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>82</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

Frequency (in %) at which $Q$ is selected for various $n$. 

Mariadassou (AgroParisTech) 
Uncovering Structure in Valued Graphs 
ECCS07 14 / 17
Simulation Setup and Results

→ Undirected graph with $Q^* = 3$ classes;

→ Poisson-valued edges;

→ $n = 50, 100, 500, 1000$ vertices;

→ $\alpha_q = (57.1\%, 28, 6\%, 14, 3\%)$ (or $a = 0.5$);

→ $\lambda = 2, \gamma = 0.5$;

→ Retrieve $Q$ that maximizes ICL;

→ 100 repeats for each value of $n$;

<table>
<thead>
<tr>
<th>$n$</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>82</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

Frequency (in %) at which $Q$ is selected for various $n$. 
Summary

Flexibility of ERMG
- A simple way to simulate networks;
- Many distributions to model different networks;
- Probabilistic model which captures features of real-networks (data not shown).

Estimation and Model selection
- Variational approaches to compute approximate MLE when dependencies are complex,
- A statistical criterion to choose the number of classes (ICL).
**E. Coli reaction network** [http://www.biocyc.org/](http://www.biocyc.org/)

- Dot-plot representation (605 nodes and 1,782 vertices)
  - 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.
Discussion

*E. Coli* reaction network [http://www.biocyc.org/](http://www.biocyc.org/)

→ 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></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>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
<td>.01</td>
<td>ε</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Adjacency matrix (sample)