Uncovering Structure in Biological Networks

J-J. Daudin[1], V. Lacroix[2], M. Mariadassou[1], V. Miele[3], F. Picard[3], S. Robin[1], M-F. Sagot[2].

[1]UMR INAPG/ENGREF/INRA MIA 518, Paris

[2]Projet HELIX, INRIA Rhône-Alpes

[3]UMR CNRS-8071/INRA-1152, Statistique et Génome, Évry

RIAMS 2006, December 2006

Outline

1. Motivations
2. An Explicit Random Graph Model
   - Some Notations
   - ER and ERMG Graph Model
3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative algorithm
   - Model Selection Criterion
4. Application

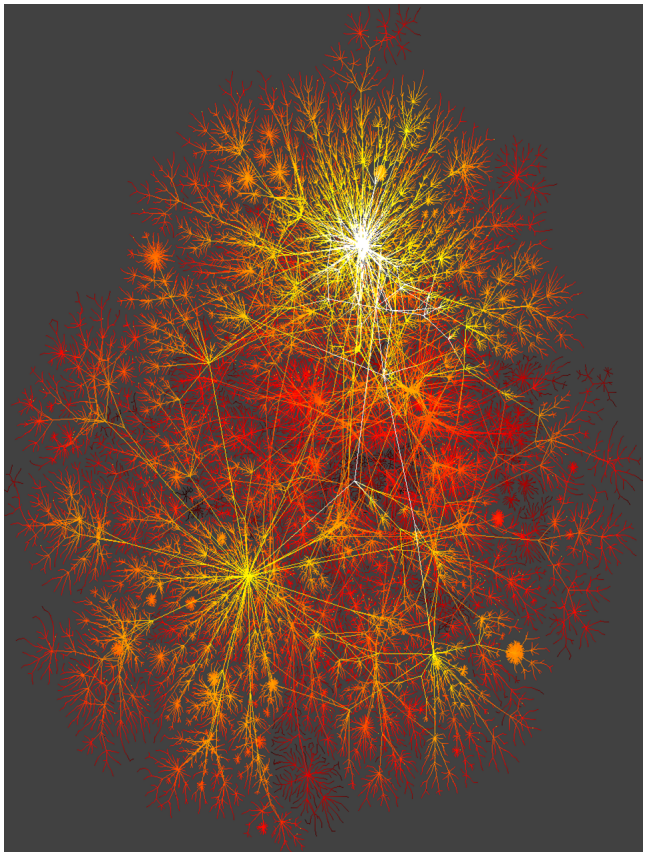
Outline

1. Motivations
2. An Explicit Random Graph Model
   - Some Notations
   - ER and ERMG Graph Model
3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative algorithm
   - Model Selection Criterion
4. Application

Motivations for the study of networks

Networks...

- Arise in many fields:
  - → Biology, Chemistry
  - → Physics, Internet.
- Represent an interaction pattern:
  - → $\mathcal{O}(n^2)$ interactions
  - → between n elements.
- Have a topology which:
  - → reflects the structure/function relationship



From Barabási website

Outline

1. Motivations
2. An Explicit Random Graph Model
   - Some Notations
   - ER and ERMG Graph Model
3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative algorithm
   - Model Selection Criterion
4. Application

Some Notations

- **Notations:**
  - → V a set of vertices in $\{1, \ldots, n\}$
  - → E a set of edges in $\{1, \ldots, n\}^2$
  - → $\mathbf{X} = (X_{ij})$ the adjacency matrix, with $X_{ij} = \mathbb{1}(i \sim j)$.

- **Possible extensions**
  - → Directed graphs: $X_{ij} \neq X_{ji}$
  - → Valued graphs: $X_{ij} \in \mathbb{N}, \mathbb{R}$

- **Random graph definition:**
  - → The joint distribution of the X_{ij} describes the topology of the network.

ER (Erdös-Rényi) Graph Model

- **The Model**
  - → The oldest and best-known graph model,
  - → The (X_{ij}) are independent, with distribution $\mathcal{B}(p)$.
- **Some Properties and Problems**
  - → Degree K_i of a node i has a Poisson distribution,

$$K_i = \sum_{j \neq i} X_{ij} \sim \mathcal{P}(\lambda)$$

  - → Clustering coefficient c is low: $c = p$

$$c = \Pr\{X_{jk} = 1 | X_{ij} = X_{ik} = 1\} = \Pr\{\nabla | \mathbf{V}\}$$

  - → Highly inaccurate to describe real networks.

ERMG: A New Probabilistic Model

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

- **$\mathbf{X}$ distribution**
  - → conditional distribution : $X_{ij}|\{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql})$,
  - → $\boldsymbol{\pi} = (\pi_{ql})$ is the connectivity matrix,
  - → ERMG : "Erdös-Rényi Mixture for Graphs".

ERMG is a model to easily generate graphs

ERMG: A New Probabilistic Model

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

- $\mathbf{X}$ **distribution**
  - → conditional distribution : $X_{ij}|\{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql})$,
  - → $\boldsymbol{\pi} = (\pi_{ql})$ is the connectivity matrix,
  - → ERMG : "Erdös-Rényi Mixture for Graphs".

ERMG is a model to easily generate graphs

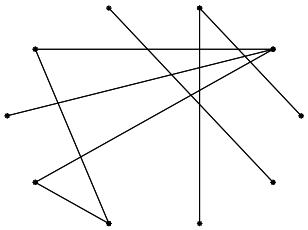
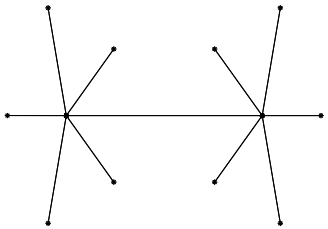
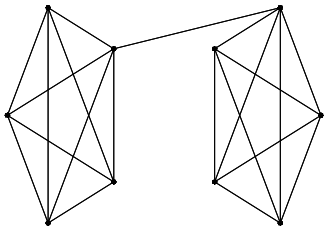
Some Properties of ERMG

- **Degree distribution**
  - $\rightarrow$ $K_i|\{Z_{iq} = 1\} \sim \mathcal{P}(\lambda_q)$, $\lambda_q = (n-1)\bar{\pi}_q$, $\bar{\pi}_q = \sum_l \alpha_l \pi_{ql}$,
  - $\rightarrow$ $K_i \sim \sum_q \alpha_q \mathcal{P}(\lambda_q)$,
  - $\rightarrow$ Mixture distribution of K_i is a sub-product of ERMG,

- **Clustering coefficient:** ERMG and the probabilistic definition give:

$$c = \sum_{q,l,m} \alpha_q \alpha_l \alpha_m \pi_{ql} \pi_{qm} \pi_{lm} \Big/ \sum_{q,l,m} \alpha_q \alpha_l \alpha_m \pi_{ql} \pi_{qm} \ .$$

ERMG couterpart to some topologies

Description	Network	Q	π	Clustering coef.
Random		1	p	p
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	0
Clusters (affiliation networks)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\frac{1+3\varepsilon^2}{(1+\varepsilon)^2}$

Outline

1. Motivations
2. An Explicit Random Graph Model
   - Some Notations
   - ER and ERMG Graph Model
3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative algorithm
   - Model Selection Criterion
4. Application

Log-Likelihood of the model

First Idea: Use maximum likelihood estimators

- **Complete data likelihood**

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

where $b(\pi_{ql}, X_{ij}) = \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{(1-X_{ij})}$

- **Observed data likelihood**

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

- The observed data likelihood requires a sum over Q^n terms, and is thus untractable
- EM-like strategies require the knowledge of $\Pr(\mathbf{Z}|\mathbf{X})$, also untractable (no conditional independence) and thus also fail.

Log-Likelihood of the model

First Idea: Use maximum likelihood estimators

- **Complete data likelihood**

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

where $b(\pi_{ql}, X_{ij}) = \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{(1-X_{ij})}$

- **Observed data likelihood**

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

- The observed data likelihood requires a sum over Q^n terms, and is thus untractable
- EM-like strategies require the knowledge of $\Pr(\mathbf{Z}|\mathbf{X})$, also untractable (no conditional independence) and thus also fail.

Log-Likelihood of the model

First Idea: Use maximum likelihood estimators

- **Complete data likelihood**

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

where $b(\pi_{ql}, X_{ij}) = \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{(1-X_{ij})}$

- **Observed data likelihood**

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

- The observed data likelihood requires a sum over Q^n terms, and is thus untractable
- EM-like strategies require the knowledge of $\Pr(\mathbf{Z}|\mathbf{X})$, also untractable (no conditional independence) and thus also fail.

Variational Inference: Pseudo Likelihood

Main Idea: Replace complicated $\Pr(\mathbf{Z}|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$ such that $KL(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal.

- Optimize in $\mathcal{R}_{\mathbf{X}}$ the function $\mathcal{J}(\mathcal{R}_{\mathbf{X}})$ given by :

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

- For simple $\mathcal{R}_{\mathbf{X}}$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ is tractable,

- At best, $\mathcal{R}_{\mathbf{X}} = \Pr(\mathbf{Z}|\mathbf{X})$ and $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathcal{L}(\mathbf{X})$.

Variational Inference: Pseudo Likelihood

Main Idea: Replace complicated $\Pr(\mathbf{Z}|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$ such that $KL(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal.

- Optimize in $\mathcal{R}_{\mathbf{X}}$ the function $\mathcal{J}(\mathcal{R}_{\mathbf{X}})$ given by :

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

- For simple $\mathcal{R}_{\mathbf{X}}$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ is tractable,
- At best, $\mathcal{R}_{\mathbf{X}} = \Pr(\mathbf{Z}|\mathbf{X})$ and $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathcal{L}(\mathbf{X})$.

Variational Inference: Pseudo Likelihood

Main Idea: Replace complicated $\Pr(\mathbf{Z}|\mathbf{X})$ by a simple $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$ such that $KL(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal.

- Optimize in $\mathcal{R}_{\mathbf{X}}$ the function $\mathcal{J}(\mathcal{R}_{\mathbf{X}})$ given by :

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

- For simple $\mathcal{R}_{\mathbf{X}}$, $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ is tractable,

- At best, $\mathcal{R}_{\mathbf{X}} = \Pr(\mathbf{Z}|\mathbf{X})$ and $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathcal{L}(\mathbf{X})$.

2 Steps Iterative Algorithm

- Step 1 **Optimize $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ w.r.t. $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$:**
  - → Restriction to a "comfortable" class of functions,
  - → $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}] = \prod_i h(\mathbf{Z}_i; \boldsymbol{\tau}_i)$, with $h(.; \boldsymbol{\tau}_i)$ the multinomial distribution,
  - → τ_{iq} is a variational parameter to be optimized using a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{l=1}^{Q} b(\pi_{ql}, X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ w.r.t. (α, π):**
  - → Constraint: $\sum_q \alpha_q = 1$

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\pi}_{ql} &= \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl}X_{ij} / \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \end{aligned}$$

2 Steps Iterative Algorithm

- Step 1 **Optimize** $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ **w.r.t.** $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$:
  - → Restriction to a "comfortable" class of functions,
  - → $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}] = \prod_i h(\mathbf{Z}_i; \boldsymbol{\tau}_i)$, with $h(.; \boldsymbol{\tau}_i)$ the multinomial distribution,
  - → τ_{iq} is a variational parameter to be optimized using a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{l=1}^{Q} b(\pi_{ql}, X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize** $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ **w.r.t.** $(\boldsymbol{\alpha}, \boldsymbol{\pi})$:
  - → Constraint: $\sum_q \alpha_q = 1$

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\pi}_{ql} &= \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl}X_{ij} / \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \end{aligned}$$

Model Selection Criterion

- We derive a statistical 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:

$$\begin{aligned} \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) &\rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} = \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} \\ \mathcal{L}(\mathbf{Z}|Q) &\rightarrow \text{pen}_{\mathbf{Z}} = (Q-1)\log(n) \end{aligned}$$

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

Model Selection Criterion

- We derive a statistical 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:

$$\begin{aligned} \mathcal{L}(\mathbf{X}|\mathbf{Z}, Q) &\rightarrow \text{pen}_{\mathbf{X}|\mathbf{Z}} = \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} \\ \mathcal{L}(\mathbf{Z}|Q) &\rightarrow \text{pen}_{\mathbf{Z}} = (Q-1)\log(n) \end{aligned}$$

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

Outline

1. Motivations
2. An Explicit Random Graph Model
   - Some Notations
   - ER and ERMG Graph Model
3. Parametric Estimation
   - Log-likelihoods and Variational Inference
   - Iterative algorithm
   - Model Selection Criterion
4. Application

Data and short results

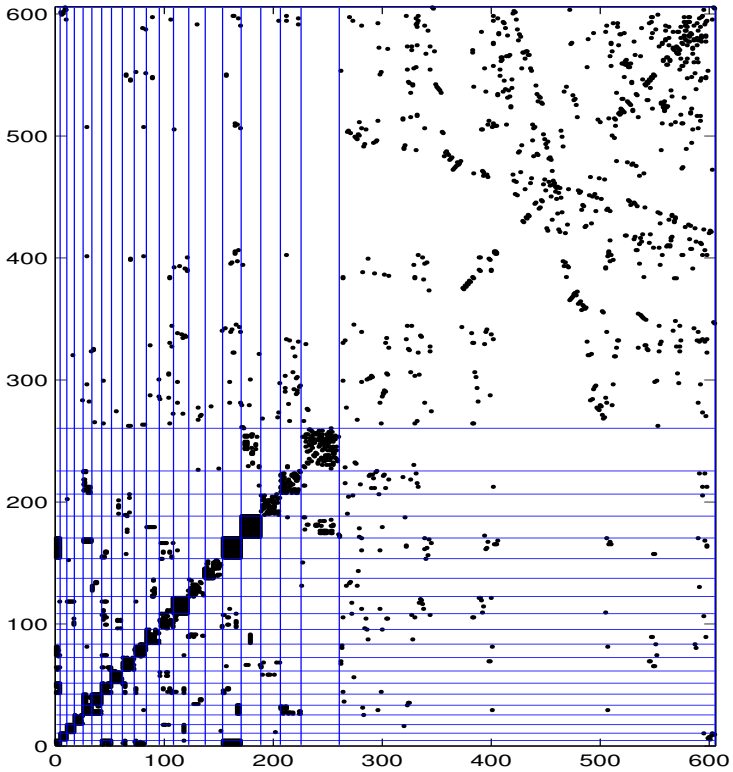
- **Reaction Network of E.Coli**:
  - → data from `http://www.biocyc.org/`,
  - → $n = 605$ vertices (reactions) and $1\,782$ edges.
  - → 2 reactions i and j are connected if the product of i is the substrate of j (cofactors excluded),
  - → V. Lacroix and M.-F. Sagot (INRIA - Hélix).

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

- **ERMG results**:
  - → ICL gives $\hat{Q} = 21$ classes,
  - → Most classes correspond to pseudo-cliques,
  - → Clustering coefficient and degree distribution asses a good fit of the model to the data.

Biological interpretation of the groups I

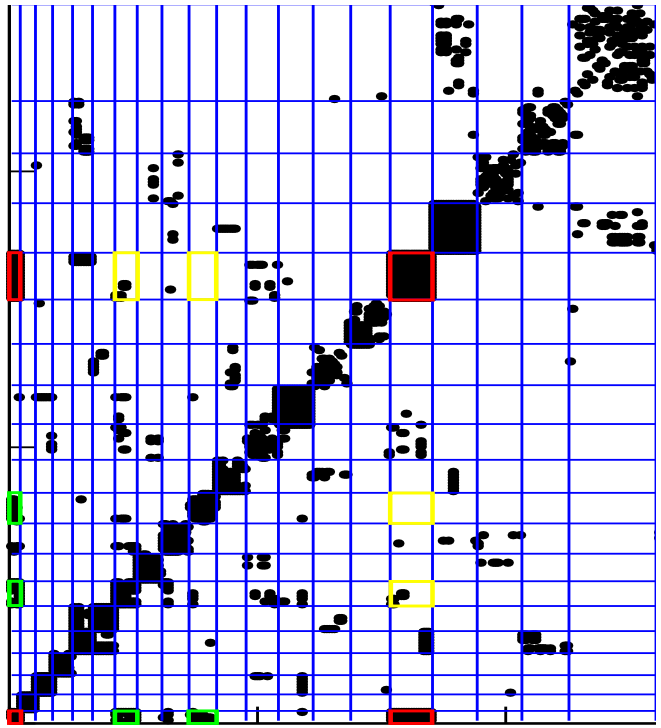
- Dot-plot representation
  - → adjacency matrix (sorted)
- Biological interpretation:
  - → Groups 1 to 20 gather reactions involving all the same compound either as a substrate or as a product,
  - → A compound (chorismate, pyruvate, ATP,*etc*) can be associated to each group.
- The structure of the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

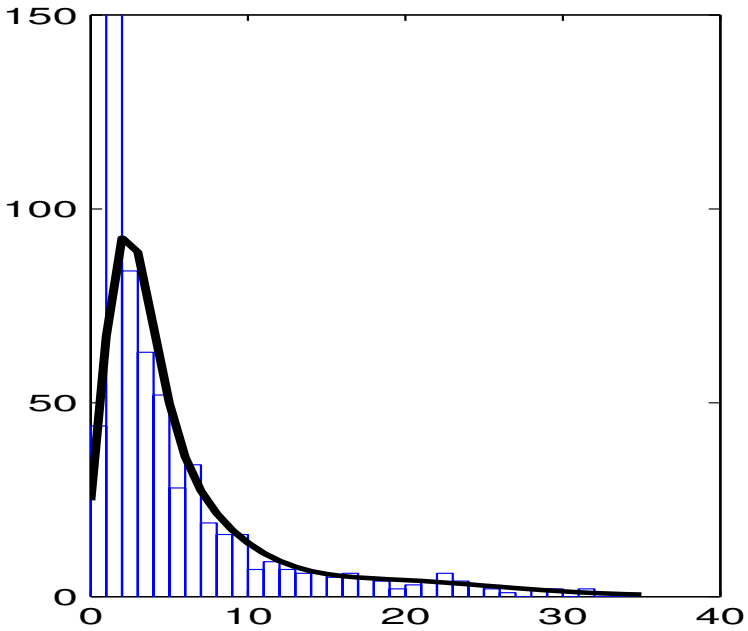
q, l	1	7	10	16
1	1.0			
7	.11	.65		
10	.43		.67	
16	1.0	.01	ϵ	1.0

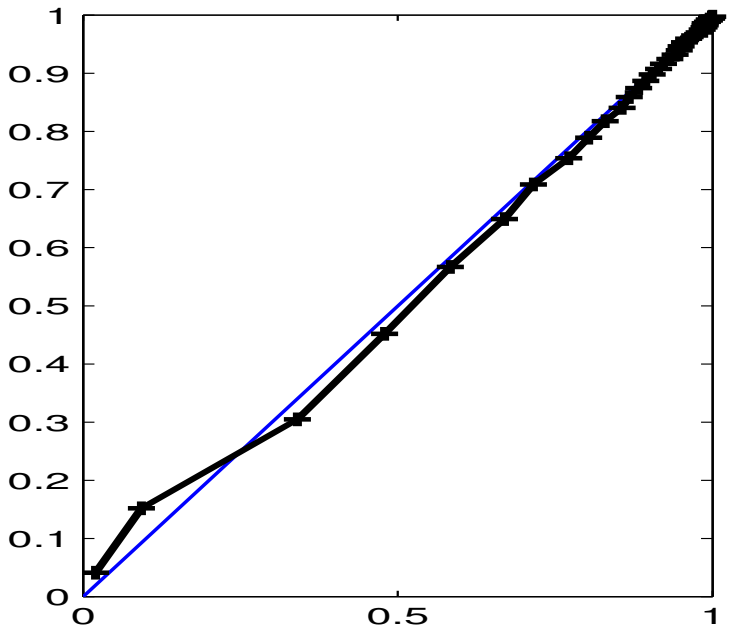


Adjacency matrix (sample)

Goodness of Fit of the ERMG Model

Degree distribution (histogram and PP-plot)





Clustering coefficient

Empirical	ERMG ($Q = 21$)	ER ($Q = 1$)
0.626	0.544	0.0098

Summary

Flexibility of ERMG

- Probabilistic model which captures features of real-networks,
- Models various network topologies,
- A promising alternative to existing methods.

Estimation and Model selection

- Variational approaches to compute approximate MLE when dependencies are complex,
- A statistical criterion to choose the number of classes (ICL).

Extensions

- Directed graphs and valued graphs
- Network motifs (*cf.* Sophie Schbath's talk)