Uncovering structure in biological networks

J-J. Daudin[†], V. Lacroix[‡], F. Picard[⋆], S. Robin[†], M-F. Sagot[‡].

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

[‡] UMR 5558 Biométrie et Biologie Évolutive, Lyon,

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

Statistics for Systems Biology (SSB) group

INRA-MIG	E. Roquain, S. Schbath,
Stat. et Génome-Évry	E. Birmelé, C. Matias, V. Miele.

The network revolution

- **Many scientific fields**:

  $\longrightarrow$ biology,

  $\longrightarrow$ sociology, physics, "internet".

- **Nature of the data under study**:

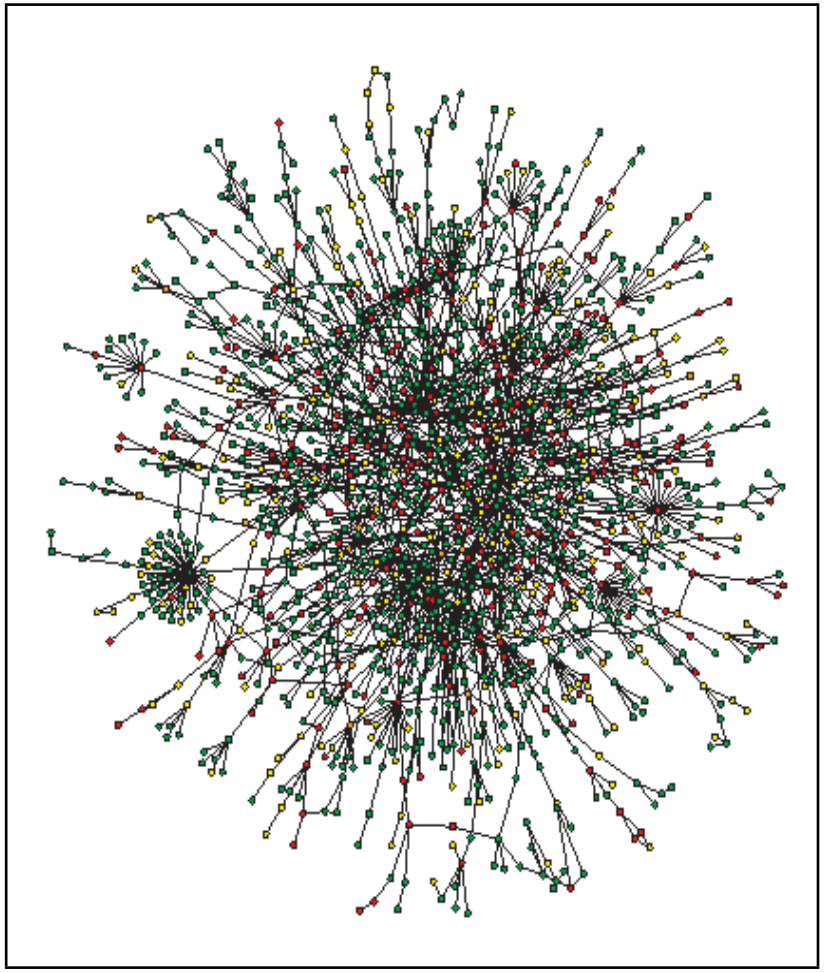
  $\longrightarrow$ interactions between n elements,

  $\longrightarrow$ $\mathcal{O}(n^2)$ possible interactions.

- **Topology of the network:**

  $\longrightarrow$ describes the way genes/proteins interact,

  $\longrightarrow$ structure/function relationship.



From Barabasi et al. (2004)

Mathematical tool: random graphs

- **Notations** :

  $\longrightarrow$ V a set of vertices in $\{1, \dots, n\}$,

  $\longrightarrow$ E a set of edges in $\{1, \dots, n\}^2$,

  $\longrightarrow$ $\mathbf{X} = (X_{ij})$ the adjacency matrix such that $\{X_{ij} = 1\} = \mathbb{I}\{i \leftrightarrow j\}$.

- **Possible graphs**:

  $\longrightarrow$ directed: $X_{ij} \neq X_{ji}$,

  $\longrightarrow$ valuated: $X_{ij} \in \mathbb{R}$.

- **Random graph definition** :

  $\longrightarrow$ the distribution of $\mathbf{X}$ describes the topology of the network.

- **Erdös Rényi (ER) model (1959)** :

  $\longrightarrow$ (X_{ij}) independent, with Bernoulli distribution $\mathcal{B}(p)$.

The ER model and the degree distribution of real networks

- Degree distribution p_k:

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

- Real networks:

  $\longrightarrow$ heterogeneous connectivity,

  $\longrightarrow$ Scale-free networks: $K_i \sim k^{-\gamma}$.

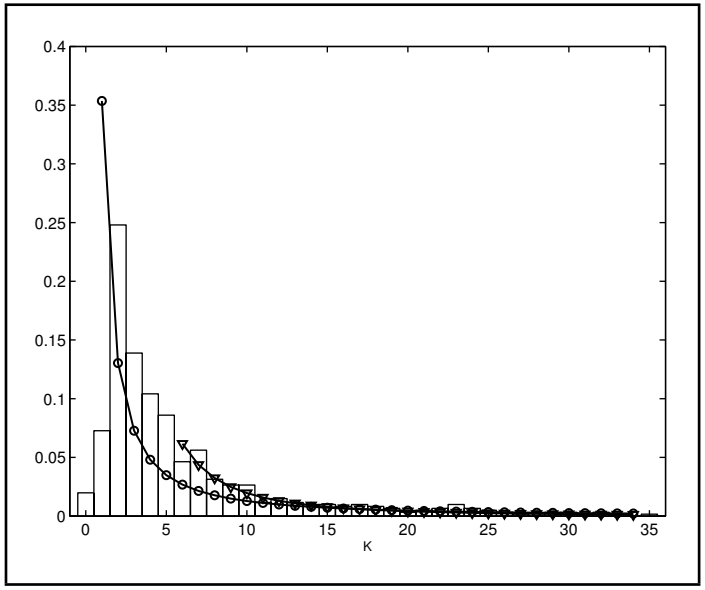
  $\longrightarrow$ Mecanistic interpretation.

  $\longrightarrow$ No consensus for the form of p_k.

- Current strategies :

  $\longrightarrow$ description of networks using p_k,

  $\longrightarrow$ theoretical results when p_k is fixed

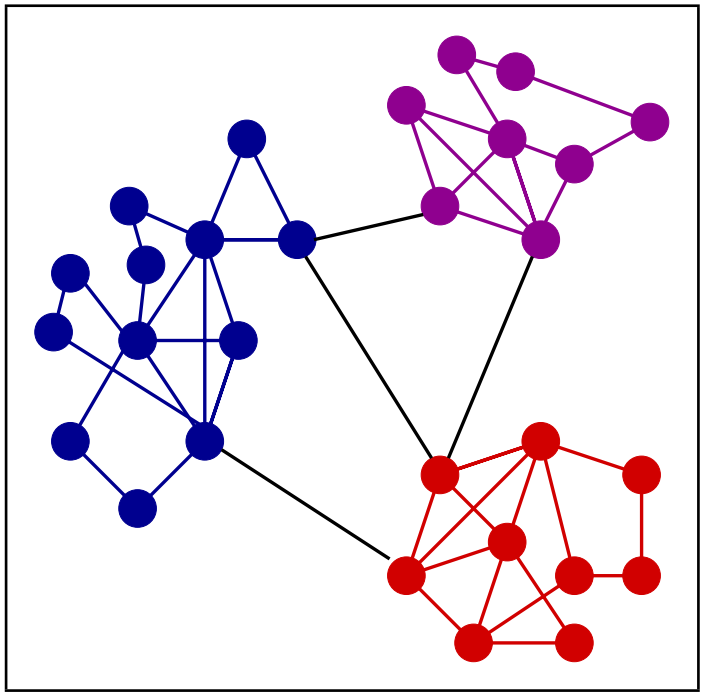


Network of metabolic reactions (E. Coli).

p_k does not give the distribution of X_{ij}.

The ER model and clustering

- The clustering coefficient c:
  - $\longrightarrow \Pr\{X_{jk} = 1 | X_{ij} = X_{ik} = 1\}$,
  - $\longrightarrow \Pr\{\nabla | \mathbf{V}\}$,
  - $\longrightarrow c = p$ in ER.
  - $\longrightarrow$ real networks : high clustering coef.
  - $\longrightarrow$ Interpretation ?
- Community structure/modularity:
  - $\longrightarrow$ heterogeneity intra/inter-clusters,
  - $\longrightarrow$ modularity of biological networks,
  - $\longrightarrow$ current strategies are algorithmic,
  - $\longrightarrow$ choosing the number of modules ?



Community structure

ERMG: a new probabilistic model

- **Modelling connection heterogeneity**

  $\longrightarrow$ hypothesis: there exists a hidden structure into Q classes of connectivity,

  $\longrightarrow$ $\mathbf{Z} = (\mathbf{Z}_i)_i$, $Z_{iq} = \mathbb{I}\{i \in q\}$ are indep. hidden variables,

  $\longrightarrow$ $\boldsymbol{\alpha} = \{\alpha_q\}$, the $prior$ proportions of groups,

  $\longrightarrow$ $(\mathbf{Z}_i) \sim \mathcal{M}(1, \boldsymbol{\alpha})$.

- $\mathbf{X}$ **distribution**

  $\longrightarrow$ conditional distribution : $X_{ij}|\{Z_{iq}Z_{j\ell} = 1\} \sim \mathcal{B}(\pi_{q\ell})$,

  $\longrightarrow$ $\boldsymbol{\pi} = (\pi_{q\ell})$ is the connectivity matrix.

  $\longrightarrow$ Marginal distribution : $X_{ij} \sim \sum_{q\ell} \alpha_q \alpha_\ell \mathcal{B}(\pi_{q\ell})$,

  $\longrightarrow$ ERMG : "Erdös-Rényi Mixture for Graphs".

Some properties of ERMG

- **Degree distribution**

  $\rightarrow K_i|\{Z_{iq}=1\} \sim \mathcal{P}(\lambda_q)$, $\lambda_q = (n-1)\bar{\pi}_q$, $\bar{\pi}_q = \sum_\ell \alpha_\ell \pi_{q\ell}$,

  $\rightarrow K_i \sim \sum_q \alpha_q \mathcal{P}(\lambda_q)$.

  $\rightarrow$ The mixture distribution of K_i is a sub-product of ERMG.

  $\rightarrow$ It models the observed heterogeneity among degrees with an intuitive interpretation.

- **Clustering coefficient** : ERMG allows us to derive a probabilistic definition:

$$c = \sum_{q,\ell,m} \alpha_q \alpha_\ell \alpha_m \pi_{ql} \pi_{qm} \pi_{\ell m} \Big/ \sum_{q,\ell,m} \alpha_q \alpha_\ell \alpha_m \pi_{q\ell} \pi_{qm} \ .$$

Parameters estimation

- **Log-likelihood(s) of the model**:
  - $\longrightarrow$ Observed data : $\mathcal{L}(\mathbf{X}) = \log\left(\sum_{\mathbf{Z}} \exp \mathcal{L}(\mathbf{X}, \mathbf{Z})\right)$.
  - $\longrightarrow$ Complete data : $\mathcal{Q}(\mathbf{X}) = \mathbb{E}\left[\mathcal{L}(\mathbf{X}, \mathbf{Z})|\mathbf{X}\right]$.
  - $\longrightarrow$ EM-like strategies require the knowledge of $\Pr(\mathbf{Z}|\mathbf{X})$.
  - $\longrightarrow$ In our case, this distribution is not tractable (no conditional independence).
- **Variational methods**:
  - $\longrightarrow$ $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$ chosen such that $KL(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal.
  - $\longrightarrow$ Optimizing $\mathcal{J}(\mathcal{R}_{\mathbf{X}})$ w.r.t. $\mathcal{R}_{\mathbf{X}}$ gives an approximation of $\mathcal{L}(\mathbf{X})$ such that:

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

  - $\longrightarrow$ If $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}] = \Pr(\mathbf{Z}|\mathbf{X})$ then $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathcal{L}(\mathbf{X})$.
  - $\longrightarrow$ Moreover $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathcal{H}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]) - \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{X}}[\mathbf{Z}]\mathcal{L}(\mathbf{X}, \mathbf{Z})$ (tractable)

An iterative algorithm

(h) **Optimizing $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ w.r.t. $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$:**

$\longrightarrow$ Restriction to a "confortable" class of distributions,

$\longrightarrow$ $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}] = \prod_i h(\mathbf{Z}_i; \boldsymbol{\tau}_i)$, with $h(\bullet; \boldsymbol{\tau}_i)$ the multinomial distribution.

$\longrightarrow$ τ_{iq} is the variational parameter to optimize using a fixed-point algorithm:

$$\tilde{\tau}_{iq} = \Pr\{Z_{iq} = 1 | \mathbf{X}, \tilde{\mathbf{Z}}^i\}.$$

$\longrightarrow$ $\tilde{\tau}_i$ is an approximation of the conditional expectation: $\tilde{\tau}_i = \mathbb{E}_{\mathcal{R}_{\mathbf{X}}}[\mathbf{Z}_i]$.

(h+1) **Optimizing $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ w.r.t. $(\boldsymbol{\alpha}, \boldsymbol{\pi})$:**

$\longrightarrow$ Constraint: $\sum_q \alpha_q = 1$,

$\longrightarrow$ $\tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq}/n$,

$\longrightarrow$ $\tilde{\pi}_{q\ell} = \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{j\ell}X_{ij} / \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{j\ell}$.

Model selection criterion

▶ We derive a statistical criterion to select the number of classes, using the integrated likelihood of the complete data:

$$\mathcal{L}(\mathbf{X}, \mathbf{Z}|m_Q) = \int_{\mathbf{\Theta}} \mathcal{L}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, m_Q) g(\boldsymbol{\theta}|m_Q) \mathrm{d}\boldsymbol{\theta}.$$

▶ This likelihood can be split: $\mathcal{L}(\mathbf{X}, \mathbf{Z}|m_Q) = \mathcal{L}(\mathbf{X}|\mathbf{Z}, m_Q) + \mathcal{L}(\mathbf{Z}|m_Q)$.

▶ These terms can be penalized separately :

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

$$ICL(m_Q) = \max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{X}, \tilde{\mathbf{Z}}|\boldsymbol{\theta}, m_Q) - \tfrac{Q(Q+1)}{4} \log \tfrac{n(n-1)}{2} - \tfrac{Q-1}{2} \log(n).$$

Application

- **Reaction Network of E.Coli** :

  $\longrightarrow$ data from `http://www.biocyc.org/`,

  $\longrightarrow$ $n = 605$ vertices (reactions) and $1\ 782$ edges.

  $\longrightarrow$ 2 reactions i and j are connected if the product of i is the substrate of j (cofactors excluded),

  $\longrightarrow$ V. Lacroix and M.-F. Sagot (INRIA - Hélix).

- **ERMG results**:

  $\longrightarrow$ ICL gives $\hat{Q} = 21$ classes.

  $\longrightarrow$ Most classes correspond to pseudo-cliques.

  $\longrightarrow$ Interpretation of the connectivity structure of classes ?

  $\longrightarrow$ Degree distribution ?

  $\longrightarrow$ Clustering coefficient?

Biological interpretation of the groups

- **Dot-plot representation**

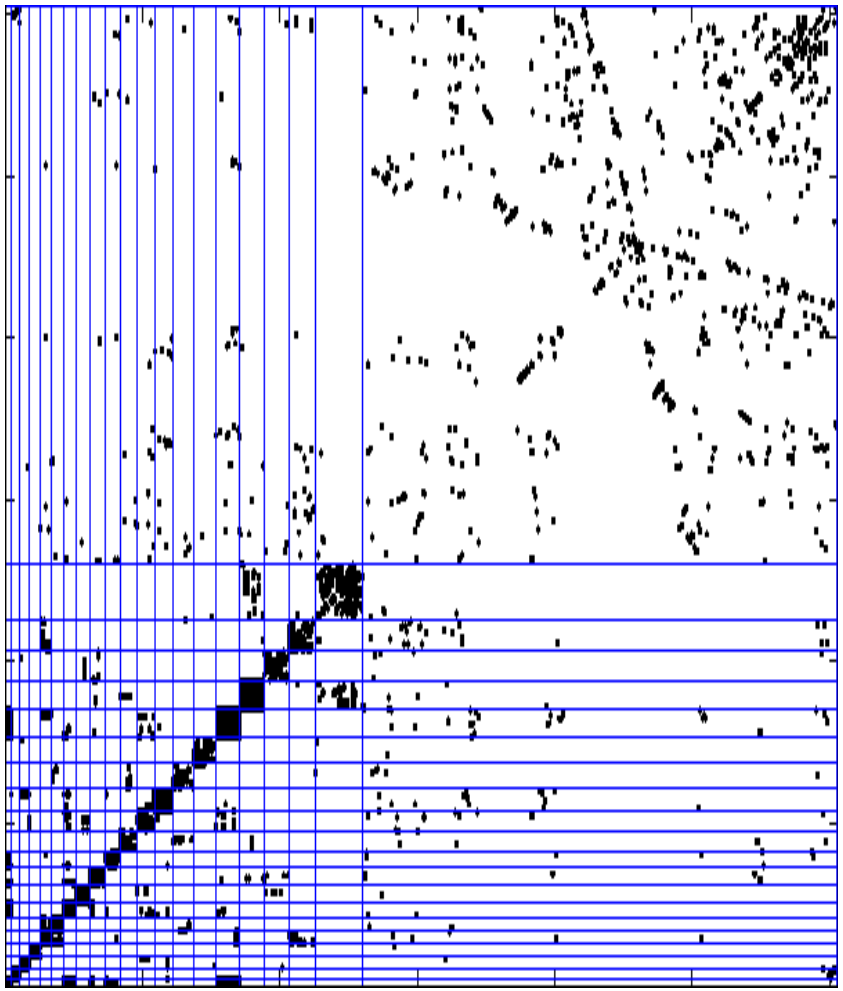
  $\longrightarrow$ adjacency matrix (sorted)

- **Biological interpretation**:

  $\longrightarrow$ Groups 1 to 20 gather reactions involving all the same compound either as a substrate or as a product.

  $\longrightarrow$ A compound (chorismate, pyruvate, ATP, *etc*) can be associated to each group.

- **The structure of the metabolic network is governed by the compounds**



Detailed example with pyruvate

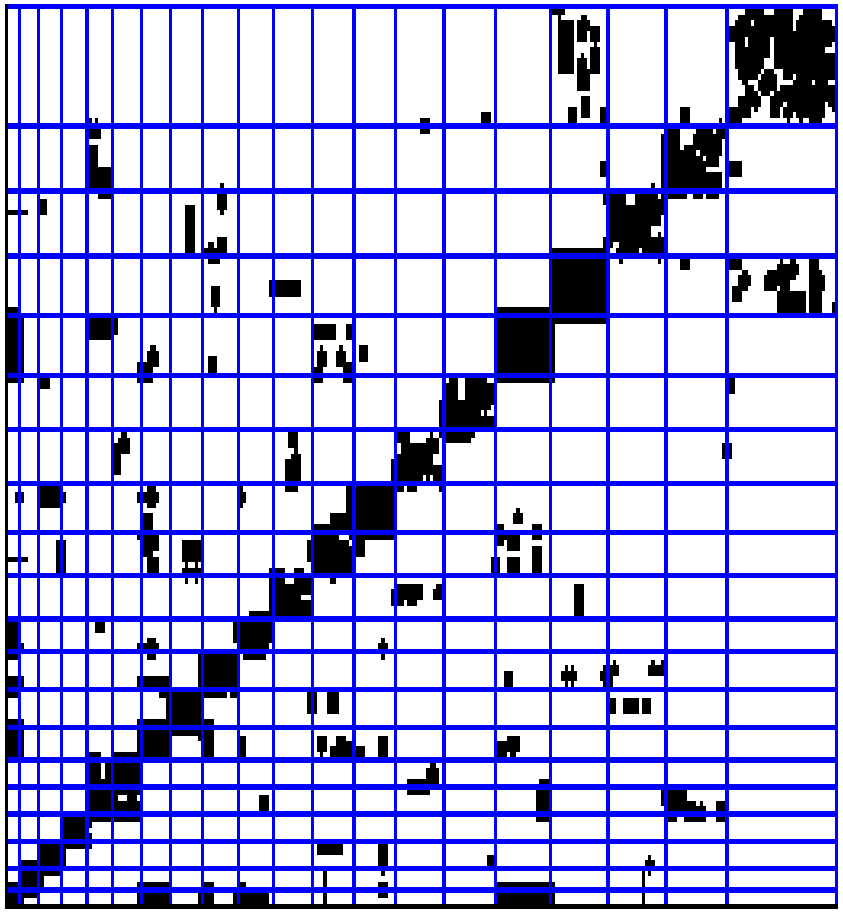
- **Biological interpretation**:

  $\longrightarrow$ classes 1 and 16 consitute a clique which corresponds to a single compound (pyruvate).

  $\longrightarrow$ They are split into 2 sub-cliques because of their connection with classes 7 (CO2) and 10 (AcetylCoA)

- **Connectivity matrix (sample)**:

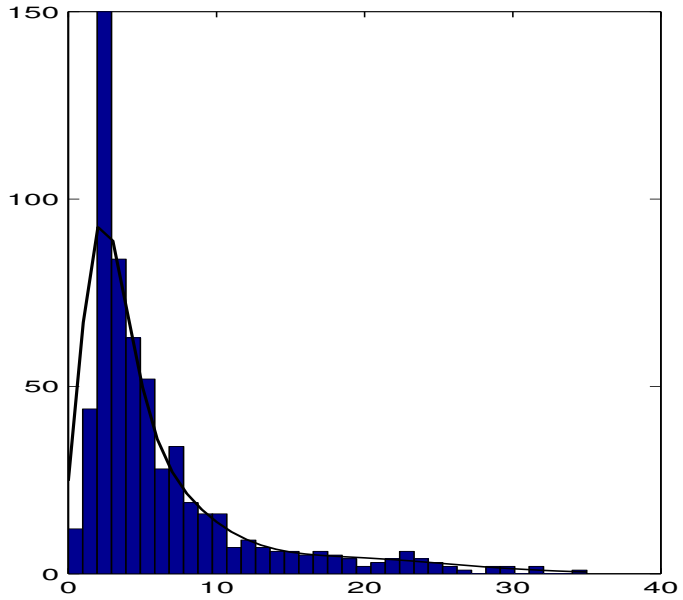
q, ℓ	1	7	10	16
1	**1.0**			
7	.11	.65		
10	.43		.67	
16	**1.0**	.01		**1.0**

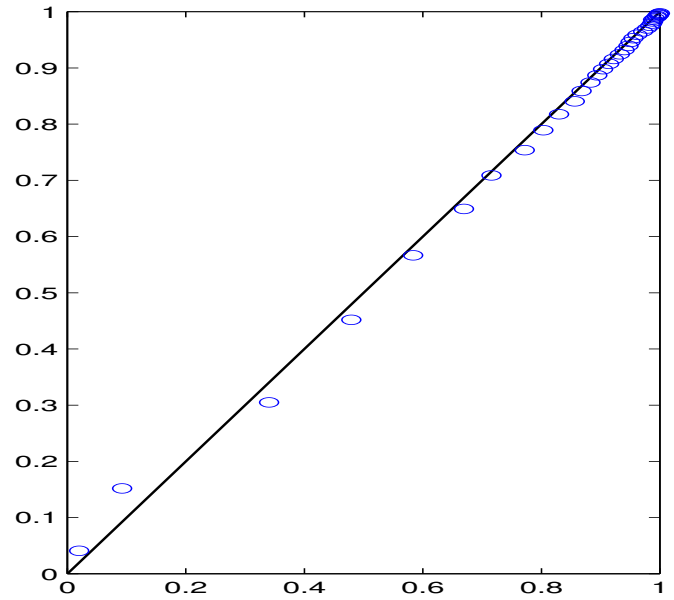


Adjacency matrix (sample)

Is ERMG more realistic than other models ?

Degree distribution (histogram and PP-plot)





Clustering coefficient

Empirical	ERMG ($Q = 6$)	ERMG ($Q = 21$)	ER ($Q = 1$)
0.626	0.436	0.544	0.0098

Conclusions

- **Flexibility of ERMG**:

  $\longrightarrow$ ERMG is a probabilistic model which captures features of real-networks,

  $\longrightarrow$ it can be used to model various network topologies,

  $\longrightarrow$ it constitutes a promising alternative to existing methods.

- **Estimation and Model selection:**

  $\longrightarrow$ variational approaches allow us to compute approximate MLE estimators when the dependency structure can not be simplified.

  $\longrightarrow$ We developed a statistical criterion to choose the number of classes (ICL).

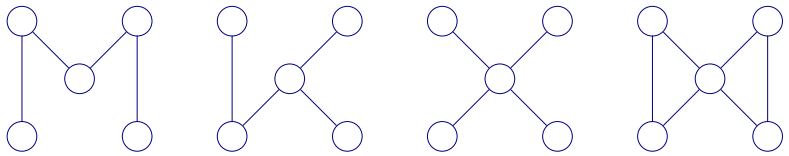
- **Extensions:**

  $\longrightarrow$ directed graphs and regulation networks,

  $\longrightarrow$ valuated graphs.

Perspectives: network motifs

- Network motifs provide insights regarding the local organization of a network.
- Examples of motifs



- Denoting $N(\mathbf{m})$, the count of motif $\mathbf{m}$,

  $\longrightarrow$ Is $N_{obs}(\mathbf{m})$ exceptionnal ?

- **Need of a probabilistic model under $\mathcal{H}_0$**:

  $\longrightarrow$ ERMG can be used for this purpose.