Uncovering Latent Structure in Valued Graphs

M. Mariadassou[1], S. Robin[1], C. Vacher[2]

[1]UMR AgroParisTech/INRA MIA 518, Paris, France

[2]INRA, UMR1202 BioGeCo, Bordeaux, France

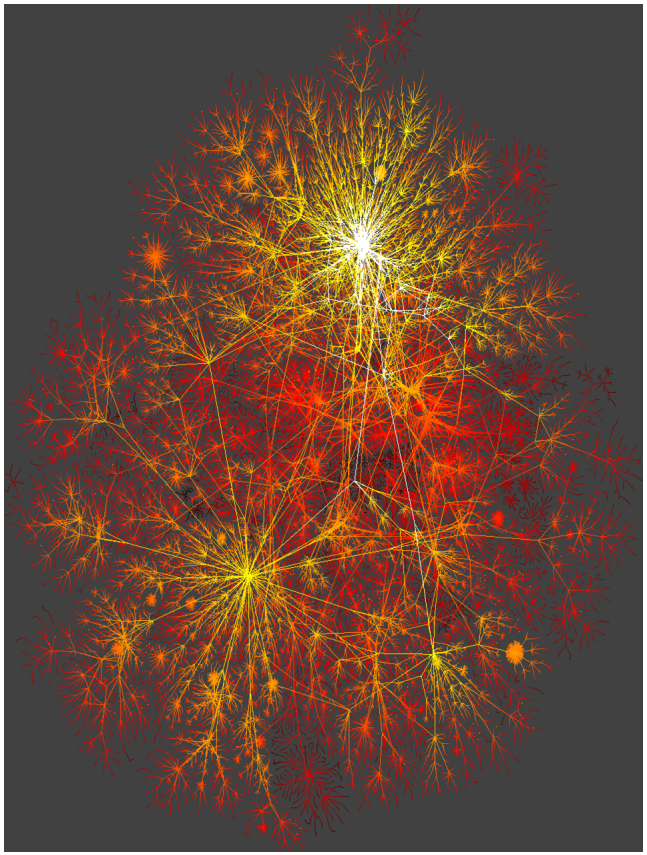
NIPS Graphs Workshop, Whistler, December 2008

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Motivations for the Study of Networks

Networks. . .

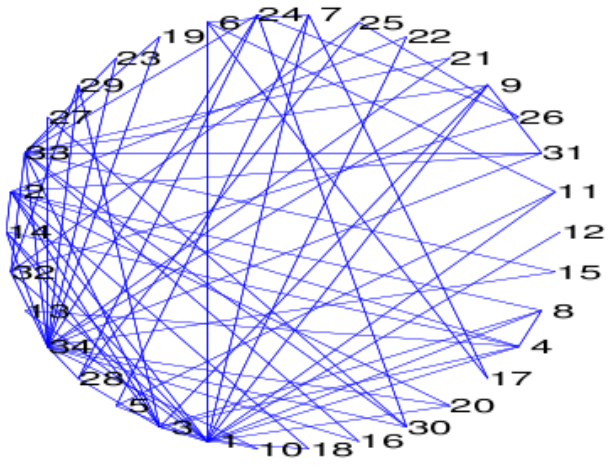
- Do exist:
  - → Electricity, transport,
  - → www networks,
- Represent or model interactions:
  - → Social relations,
  - → Metabolic pathways,
  - → Chemical reactions,
- Are hard to interpret as such:
  - → Many nodes/edges,
  - → Structure is hidden.

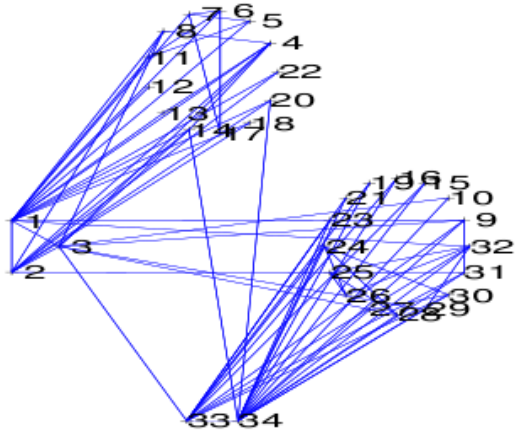


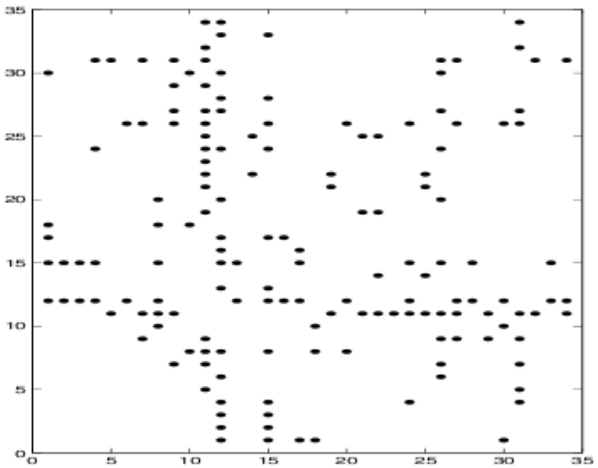
From Barabási website

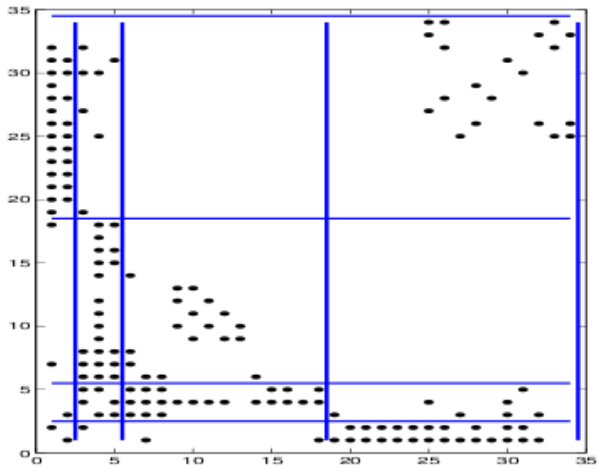
AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Goal: Simple Representation of the Graph









Some Notations

Nodes

- n nodes labeled in $\{1, \ldots, n\}$;
- Q classes with different connectivity;
- $Z = Z_{iq}$ discrete latent variable, $Z_{iq} = 1$ if node i belongs to class q;
- $(Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_Q)$.

Edges

- X_{ij} = value of the edge from node i to node j: $i \overset{X_{ij}}{\rightarrow} j$
- Conditionally to Z, the X_{ij}s are independent with:

$$X_{ij} | i \in q, j \in l \sim f_{\theta_{ql}}$$

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Some Notations

Nodes

- n nodes labeled in $\{1, \ldots, n\}$;
- Q classes with different connectivity;
- $Z = Z_{iq}$ discrete latent variable, $Z_{iq} = 1$ if node i belongs to class q;
- $(Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_Q)$.

Edges

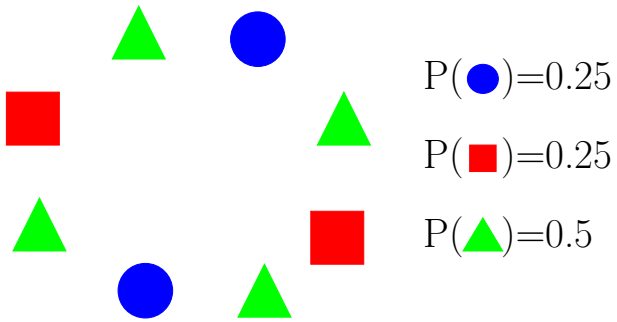
- X_{ij} = value of the edge from node i to node j: $i \overset{X_{ij}}{\rightarrow} j$
- Conditionally to Z, the X_{ij}s are independent with:

$$X_{ij} | i \in q, j \in l \sim f_{\theta_{ql}}$$

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Example

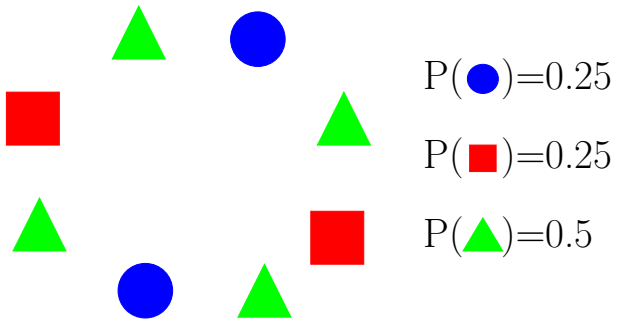
- **Nodes:**



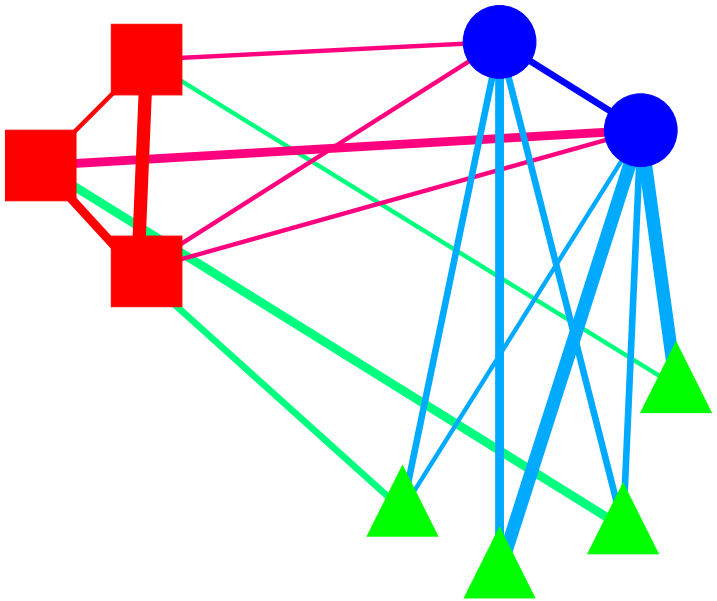
- **Edge (Poisson)**

Example

- **Nodes:**



- **Edge (Poisson)**



$$\begin{array}{c} \\ \blacktriangle \\ \bullet \\ \blacksquare \end{array} \begin{array}{c} \begin{array}{ccc} \blacktriangle & \bullet & \blacksquare \end{array} \\ \begin{pmatrix} 0 & 0.9 & 0.25 \\ . & 1 & 0.5 \\ . & . & 1 \end{pmatrix} \end{array}$$

Flexibility of MixNet

- **Classical Distributions:**

→ f_θ can be any probability distribution;

→ Bernoulli (interaction graph): presence/absence of an edge;

→ Multinomial (labelled graph): nature of the connection (friend, lover, colleague);

→ Poisson (count): in coauthorship networks, number of copublished papers;

→ Gaussian (intensity): intensity of the connection (airport network);

MixNet easily generates graphs.

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Pseudo-Likelihood

Maximum Likelihood Approach

- 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 also fail.

Variational Inference: Pseudo Likelihood

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

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

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

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:

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

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

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\theta}_{ql} &= arg\max_{\theta} \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \log f_{\theta}(X_{ij}) \end{aligned}$$

  - → Simple expression of $\tilde{\theta}_{ql}$ for classical distributions.

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

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:

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

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

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\theta}_{ql} &= arg\max_{\theta} \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \log f_{\theta}(X_{ij}) \end{aligned}$$

  - → Simple expression of $\tilde{\theta}_{ql}$ for classical distributions.

Model Selection Criterion

- 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:

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

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

Model Selection Criterion

- 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:

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

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

Number of Classes

→ Undirected graph with $Q^\star = 3$ classes and Poisson edges;

→ $n = 50,\ 100,\ 500,\ 1000$ vertices;

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

→ Connectivity matrix of the form $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

	Q		
n	2	3	4
50	82	**17**	1
100	7	**90**	3
500	0	**100**	0
1000	0	**100**	0

Table: Frequency of selected Q for various n.

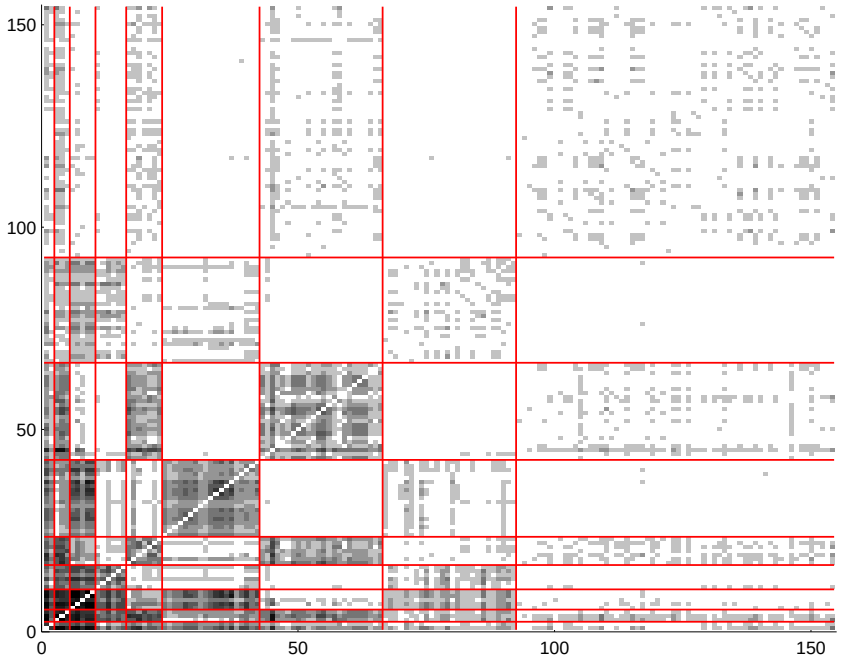
Fungi Trees Interactions

- **Dataset** Interactions between 154 fungi and 51 trees European species. Fungus f is connected to tree t if it has been collected on it.
- **Projected Graphs** For each species, we define the projected graph:
  for trees $X_{tt'}$ = Number of common fungi,
  for fungi $X_{ff'}$ = Number of common trees,
- **Poisson Model** For both species, we assume that the intensities have Poisson distributions: $X \simeq \mathcal{P}(\lambda_{ql})$.
- **Number of classes**. The ICL criterion selects:
  - 7 classes for trees;
  - 9 classes for fungi.

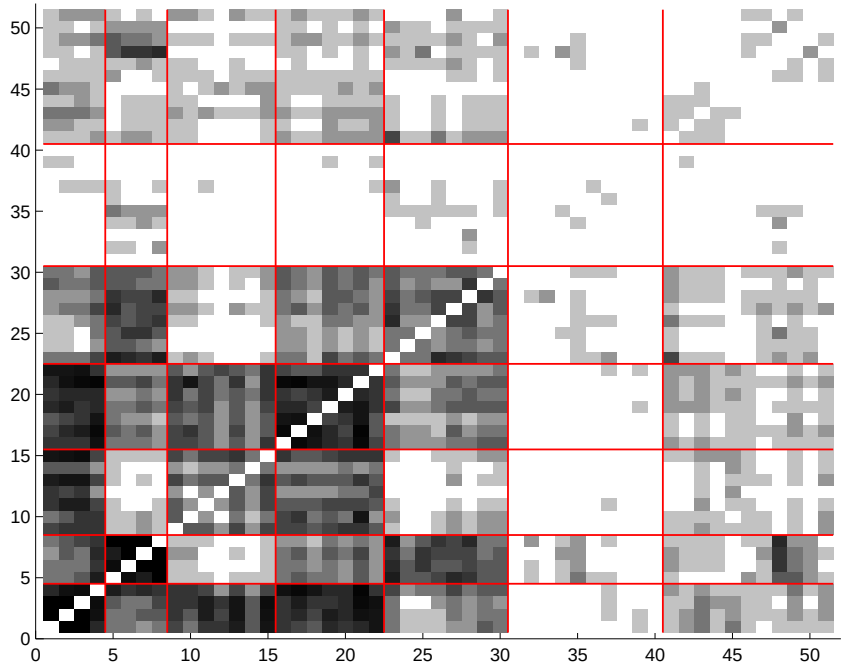
AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Results of Clustering

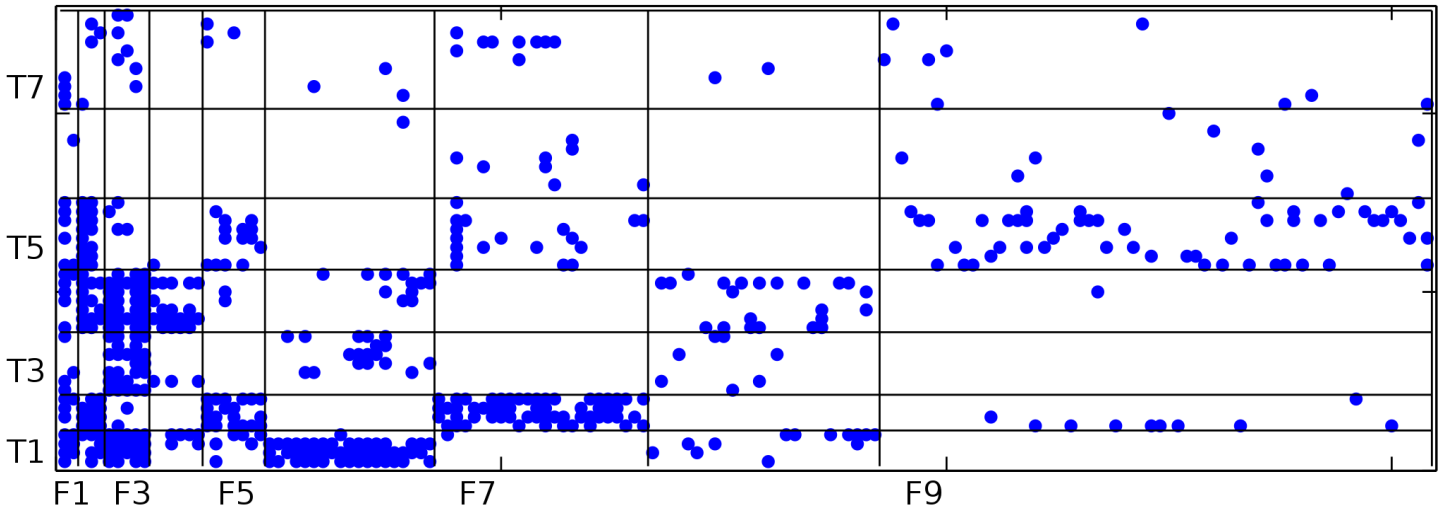
Fungi Network



Trees Network



Clustered Interaction Network



- Tree classes are associated with phylogenetic history;
- Fungus classes are associated with nutritional strategy;
- Consistent with (known) tree speciation prior to fungus speciation;
- Identify generalist and specialist fungi classes.

Summary

- Flexible probabilistic model to detect structure in complex valued graphs;
- Pseudo-likelihood estimators computed obtained through variational EM;
- A statistical criterion to choose the number of classes;
- Package available at `http://pbil.univ-lyon1.fr/software/MixNet`.

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

E. Coli Reaction Network

- **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).

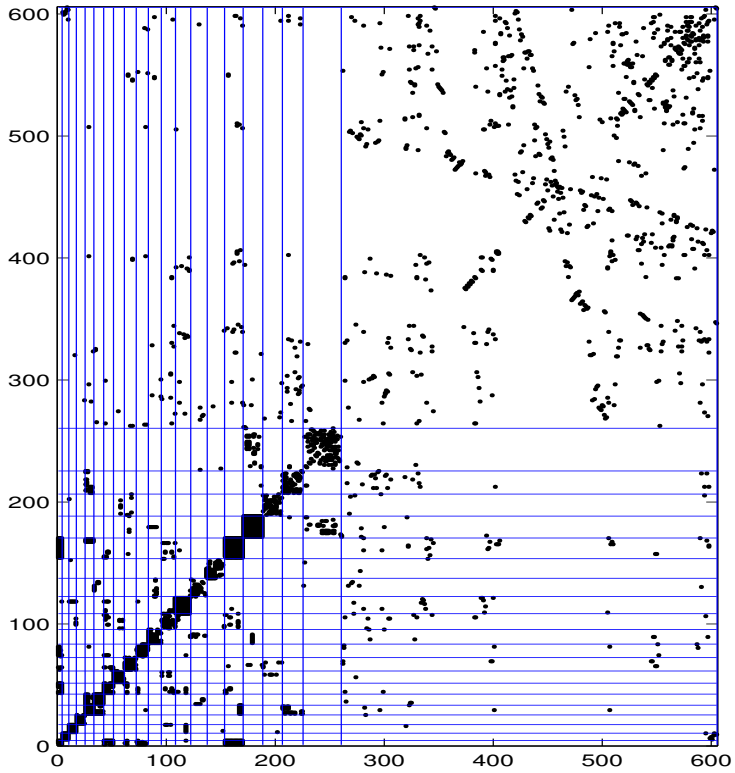
- **Question**:
  - → Interpretation of the connectivity structure of classes?

- **MixNet results**:
  - → ICL gives $\hat{Q} = 21$ classes,
  - → Most classes correspond to pseudo-cliques,

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Biological interpretation of the groups I

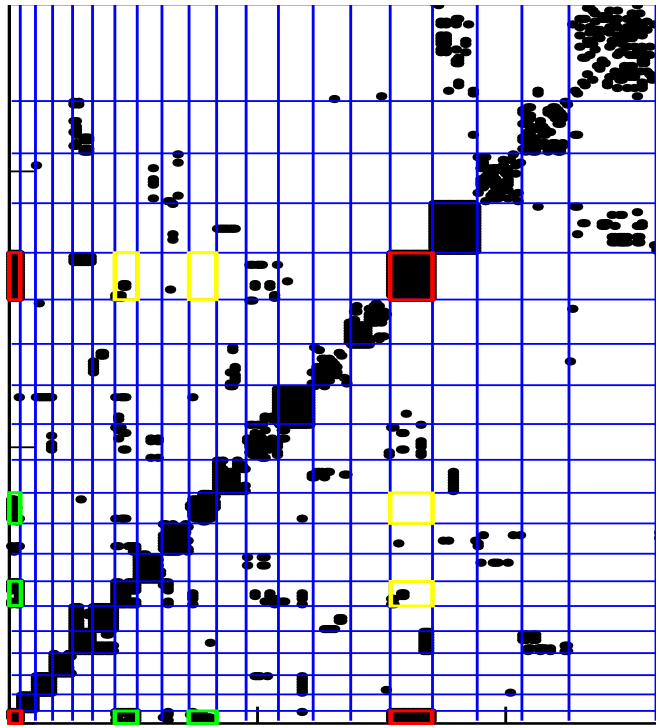
- Dot-plot representation
  - → adjacency matrix (sorted)
- Biological interpretation:
  - → Groups 1 to 20 gather reactions involving all the same compound either as a substrate or as a product,
  - → A compound (chorismate, pyruvate, ATP,*etc*) can be associated to each group.
- The structure of the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