Uncovering latent structure in valued graphs: A variational approach

S. Robin

Joint work with J.-J. Daudin, M. Mariadassou, F. Picard, C. Vacher

UMR AgroParisTech / INRA, Paris, Mathématique et Informatique Appliquées :
`www.agroparistech.fr/mia/`

Statistics for Biological Sequences (SSB) group:
`genome.jouy.inra.fr/ssb/`

Research report: `genome.jouy.inra.fr/ssb/preprint/`

- `SSB-RR-4.ermg.pdf` + Stat. Comput., 18(2):173-83, Jun 2008.
- `SSB-RR-10.valued-graphs.pdf`

1 - Looking for structure in networks

Networks . . .

- Arise in many fields:

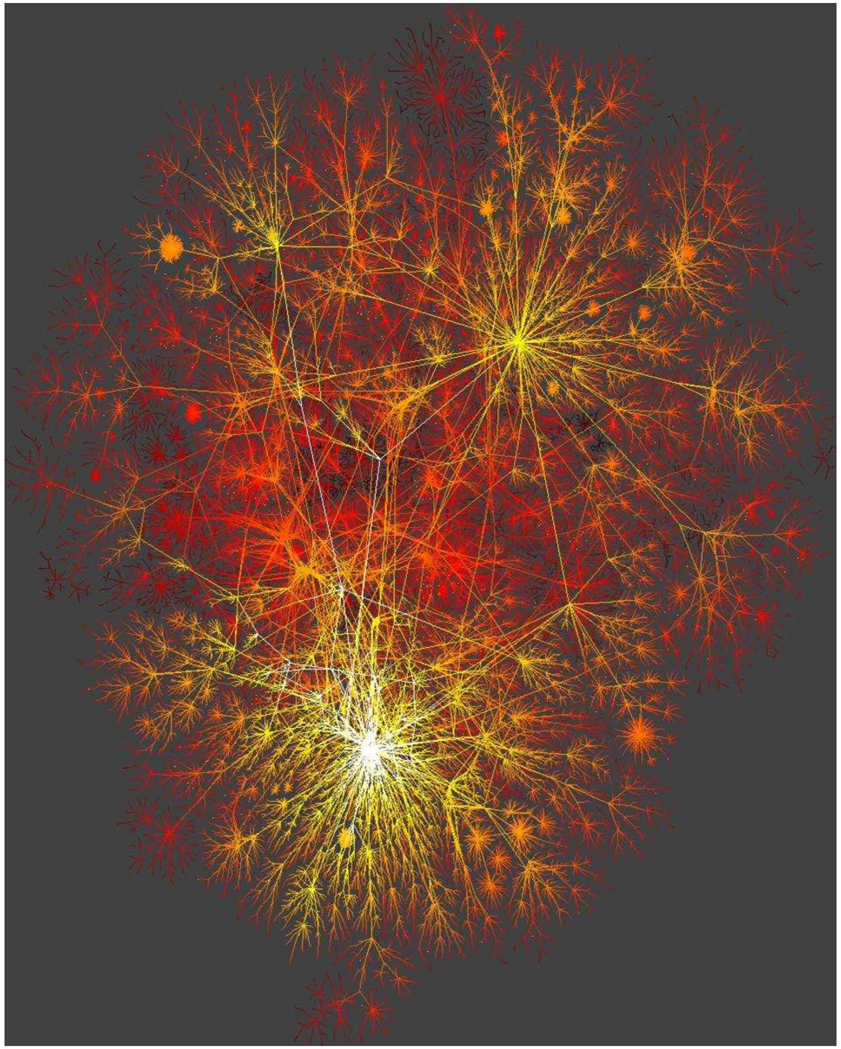
  → Biology, Chemistry
  → Physics, Internet.

- Represent an interaction pattern:

  → $\mathcal{O}(n^2)$ interactions
  → between n elements.

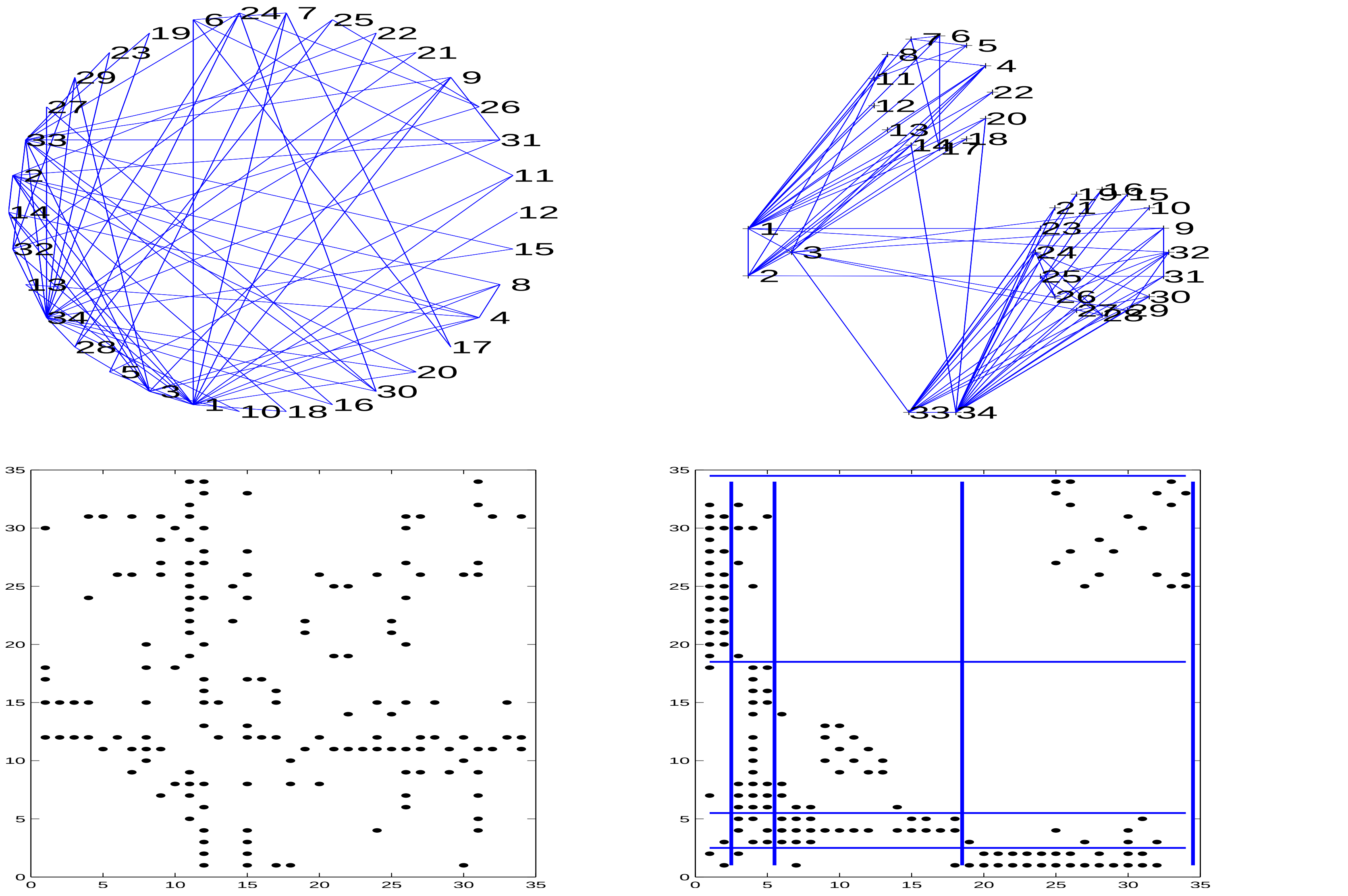
- Have a topology which:

  → reflects the structure / function relationship



From Barabási website

Uncovering structure in networks: A simple example



1.1 - Heterogeneity in random graphs

Nodes may have different connectivity behaviour.

Looking for connected sub-groups:

- Detection of cliques or groups of highly connected nodes: *Gethor & Diehl, 04*
- Edge betweenness: *Girvan & Newman, 02*
- Spectral clustering: *Von Luxburg & al., 07*

Model based:

- Underlying topology: *Hoff & al., 02* (Latent space)
- Mixture model *Nowicki & Snijders, 01* (Block structure), *Daudin & al., 08* (Mixture for graphs)
- General model for heterogeneous networks: *Bollobás al., 07* (Topological properties: Giant component, diameter, degree distribution = compound Poisson, *etc.*).
- General review on random graph models: *Pattison & Robbins, 07*

1.2 - Inhomogeneous random graphs

General definition for binary graphs. (*Bollobás al., 07*)

- n nodes ($i = 1 \dots n$)
- $n(n-1)/2$ possible edges: $X_{ij} = \mathbb{I}\{i \sim j\}$
- Each i is characterised by a *latent variable* Z_i sampled in some space $\mathcal{Z}$ with distribution α:

$$\{Z_i\}_i \text{ i.i.d.}, \qquad Z_i \sim \alpha$$

- Edge (i,j) is present with probability $\pi(Z_i, Z_j)$, where π is a *kernel function*:

$$\{X_{ij}\}_{i,j} \text{ independent given } \{Z_i\}_i, \qquad X_{ij} \sim \mathcal{B}[\pi(Z_i, Z_j)].$$

Latent space: $\mathcal{Z} = \mathbb{R}^k$, $\qquad \pi(z, z') = \dfrac{\exp(a - |z - z'|)}{1 + \exp(a - |z - z'|)}$.

Mixture model: $\mathcal{Z} = \{1, \dots, Q\}$, $\qquad \pi(z, z') = \pi_{q\ell}$ for $z = q, z' = \ell$.

2 - Mixture model for valued graphs

Our approach

- is model based:

Mixture model

- deals with valued graphs:

$$X_{ij} \in \{0,1\}, \mathbb{N}, \mathbb{R}, \mathbb{R}^d, etc.$$

- and makes frequentist inference using a variational method:

Approximate maximum likelihood.

2.1 - Model

- n nodes $(i = 1 \dots n)$;
- each node i belong to class q with probability α_q:

$$\{Z_i\}_i \text{ i.i.d.}, \qquad Z_i \sim \mathcal{M}(1; \boldsymbol{\alpha})$$

  where $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_Q)$;

- The values of the edges $\{X_{ij}\}_{i,j}$ are conditionally independent given the Z_i's:

$$(X_{ij} \mid Z_i = q, Z_j = \ell) \sim f_{q\ell}(\cdot).$$

  where $f_{q\ell}(\cdot)$ is some parametric distribution $f_{q\ell}(x) = f(x; \theta_{q\ell})$.

We denote: $\mathbf{Z} = \{Z_i\}_i$, $\mathbf{X} = \{X_{ij}\}_{i,j}$, $\boldsymbol{\theta} = \{\theta_{q\ell}\}_{q,\ell}$, $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$.

2.2 - Some distributions $f_{q\ell}$

Bernoulli $\mathcal{B}(\pi_{ql})$. Binary oriented or non-oriented *interaction graphs*:
Relation network, protein-protein interaction, gene regulation.

Multinomial $\mathcal{M}(\boldsymbol{\pi}_{ql})$. *Labelled edges*:
Social networks ('friend', 'lover', colleague'), Directed graphs with correlated edges (' ', '$\longrightarrow$', '$\longleftarrow$', '$\longleftrightarrow$').

Poisson $\mathcal{P}(\lambda_{ql})$. The edge value is a *count*:
Number of co-publications of two authors, Number of times two species were observed in the same place, Number of alleles shared by two species.

Gaussian $\mathcal{N}(\mu_{q\ell}, \sigma^2)$. *Traffic intensity*:
Airport network, Electric network.

Linear regression. If *covariates* $\mathbf{y}_{ij}$ are available for each couple of nodes:

$$X_{ij} = \mathbf{y}_{ij}\boldsymbol{\beta}_{q\ell} + E_{ij}, \qquad \{E_{ij}\}_{i,j} \text{ independent, } E_{ij} \sim \mathcal{N}(0, \sigma^2).$$

3 - Variational inference

3.1 - Maximum Likelihood Inference

Likelihoods. The log-likelihood of the complete dataset $(\mathbf{X}, \mathbf{Z})$ is

$$\begin{aligned}\log \mathbb{P}(\mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \log \mathbb{P}(\mathbf{Z}; \boldsymbol{\alpha}) + \log \mathbb{P}(\mathbf{X}|\mathbf{Z}; \boldsymbol{\theta}) \\ &= \sum_i \sum_q Z_{iq} \log \alpha_q + \sum_{i \neq j} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log f_{q\ell}(X_{ij}).\end{aligned}$$

The log-likelihood of the observed dataset $(\mathbf{X})$ is

$$\log \mathbb{P}(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} \log \mathbb{P}(\mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta})$$

and cannot be evaluated since $\mathbf{Z}$ may take Q^n different values.

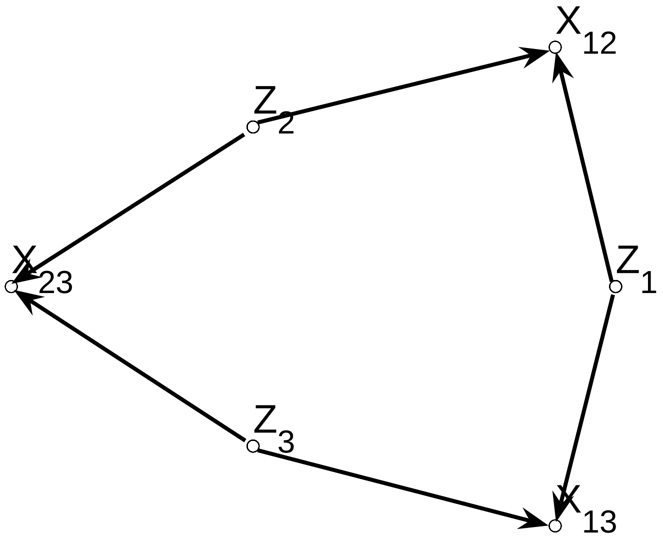
Most popular solution: E-M algorithm.

E-M algorithm. To achieve the E-step, we need to calculate the conditional distribution of the unobserved data given the observed ones: $\log \mathbb{P}(\mathbf{Z}|\mathbf{X})$.

Due to intricate dependencies this distribution is *intractable*:

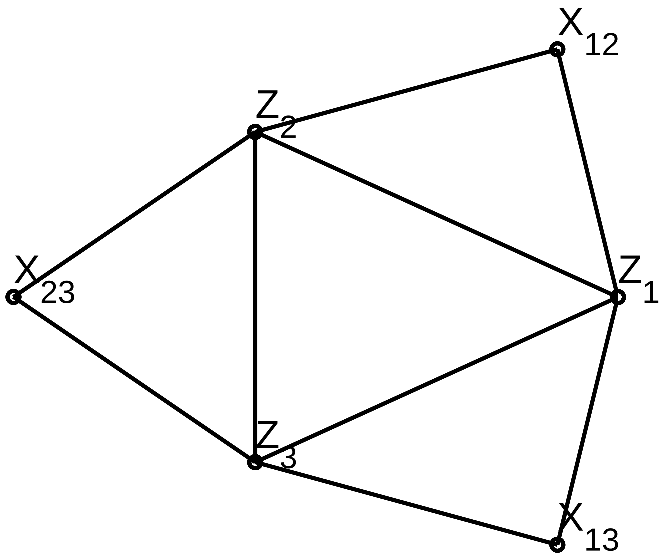
Dependency graph (oriented)

Edge X_{ij} only depends on its two parents Z_1 and Z_2



Moral graph (parents are married)

Conditional on the edges, labels Z_i's all depend on each others



$\Rightarrow$ All edges are actually *'neighbours'* (unlike in Bayesian networks).

3.2 - Variational strategy

Variational trick: Maximise a *lower bound* of the incomplete likelihood

$$\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta}) - KL[R_{\mathbf{X}}(\cdot), \mathbb{P}(\cdot|\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta})]$$

where

- KL denotes the Kullback-Leibler divergence
- $R_{\mathbf{X}}$ is some distribution for $\mathbf{Z}$.

Thanks to the definition of KL, we get for any $R_{\mathbf{X}}$ (*Jaakkola, 00*)

$$\begin{aligned}\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \log \mathbb{P}(\mathbf{X}) - \sum_{\mathbf{Z}} \log[R_{\mathbf{X}}(\mathbf{Z})] R_{\mathbf{X}}(\mathbf{Z}) + \sum_{\mathbf{Z}} \log[P(\mathbf{Z}|\mathbf{X})] R_{\mathbf{X}}(\mathbf{Z}) \\ &= \mathcal{H}(R_{\mathbf{X}}) + \sum_{\mathbf{Z}} R_{\mathbf{X}}(\mathbf{Z}) \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\alpha}, \boldsymbol{\theta})\end{aligned}$$

where $\mathcal{H}(R_{\mathbf{X}})$ stands for the entropy of distribution $R_{\mathbf{X}}$.

Choice of $R_{\mathbf{X}}$. $R_{\mathbf{X}}$ approximates the conditional distribution $\mathbb{P}(\mathbf{Z}|\mathbf{X})$. We want it to be

- tractable (e.g. factorised):

$$R_{\mathbf{X}}(\mathbf{Z}) = \prod_i h(\mathbf{Z}_i, \boldsymbol{\tau}_i)$$

  where $h(\cdot, \boldsymbol{\tau})$ denotes the multinomial distribution;

- as close to $\mathbb{P}(\mathbf{Z}|\mathbf{X})$ as possible:

$$\widehat{\boldsymbol{\tau}} = \arg\min KL[R_{\mathbf{X}}(\cdot), \mathbb{P}(\cdot|\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\theta})].$$

We get

$$\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_{i \neq j} \sum_{q,\ell} \tau_{iq}\tau_{j\ell} \log f_{q\ell}(X_{ij}).$$

The τ_i's are interpreted as *approximate posterior probabilities* $\mathbb{P}\{Z_i = q|\mathbf{X}\}$;

3.3 - Estimation algorithm

The optimisation of $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is achieved via two alternative steps.

M-step: Maximises $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\alpha}, \boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$ given $\boldsymbol{\tau}$. We get

$$\widehat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}, \qquad \widehat{\theta}_{q\ell} = \arg\max_{\theta_{q\ell}} \sum_{i \neq j} \tau_{iq} \tau_{j\ell} \log f(X_{ij}; \theta_{q\ell}).$$

Pseudo E-step: Finds the optimal $\boldsymbol{\tau}$ given $(\boldsymbol{\alpha}, \boldsymbol{\theta})$. We end up with a *fix point relation*.

- Oriented graphs:

$$\log \widehat{\tau}_{iq} = \text{cst} + \log \alpha_q + \sum_{j \neq i} \sum_{\ell} \widehat{\tau}_{j\ell} \left[\log f(X_{ij}; \theta_{q\ell}) \log f(X_{ji}; \theta_{\ell q}) \right].$$

- Non-oriented graphs:

$$\log \widehat{\tau}_{iq} = \text{cst} + \log \alpha_q + \sum_{j \neq i} \sum_{\ell} \widehat{\tau}_{j\ell} \log f(X_{ij}; \theta_{q\ell}).$$

3.4 - Model selection

Penalised likelihood. Standard criteria, such as BIC or AIC are based on the log-likelihood of observed data $\log \mathbb{P}(\mathbf{X})$, so they can not be used here.

Integrated Classification Likelihood (ICL). The ICL criterion (*Biernacki & al., 00*) is an approximation of the complete-data integrated log-likelihood:

$$\log \mathbb{P}(\mathbf{X}, \mathbf{Z} | m_Q) = \int \log \mathbb{P}(\mathbf{X}, \mathbf{Z} | \boldsymbol{\gamma}, m_Q) g(\boldsymbol{\gamma} | m_Q) d\boldsymbol{\gamma},$$

where $\log \mathbb{P}(\mathbf{X}, \mathbf{Z} | \gamma, m_Q)$ is the log-likelihood of model m_Q with Q classes.

We get

$$ICL(m_Q) = \max_{\boldsymbol{\gamma}} \log \mathbb{P}(\mathbf{X}, \widehat{\mathbf{Z}} | \gamma, m_Q) - \frac{1}{2} \left\{ P_Q \log[n(n-1)] - (Q-1)\log(n) \right\}.$$

where P_Q denotes the number of parameters in $\boldsymbol{\theta}$ and $\widehat{\mathbf{Z}}$ can be replaced by $\widehat{\boldsymbol{\tau}}$ or by the Maximum A posteriori (MAP) prediction of $\mathbf{Z}$.

4 - Applications

4.1 - Metabolic network of *E. coli*

Dataset.

- The network is made of 605 reaction (nodes) and 1782 edges (*V Lacroix & M.-F. Sagot, INRIA*).
- Reactions i and j are connected if the compound of i is the substrate of j.
- Because most reactions are reversible, the network is not oriented.
- The only information about edges is terms of presence/absence.

Results

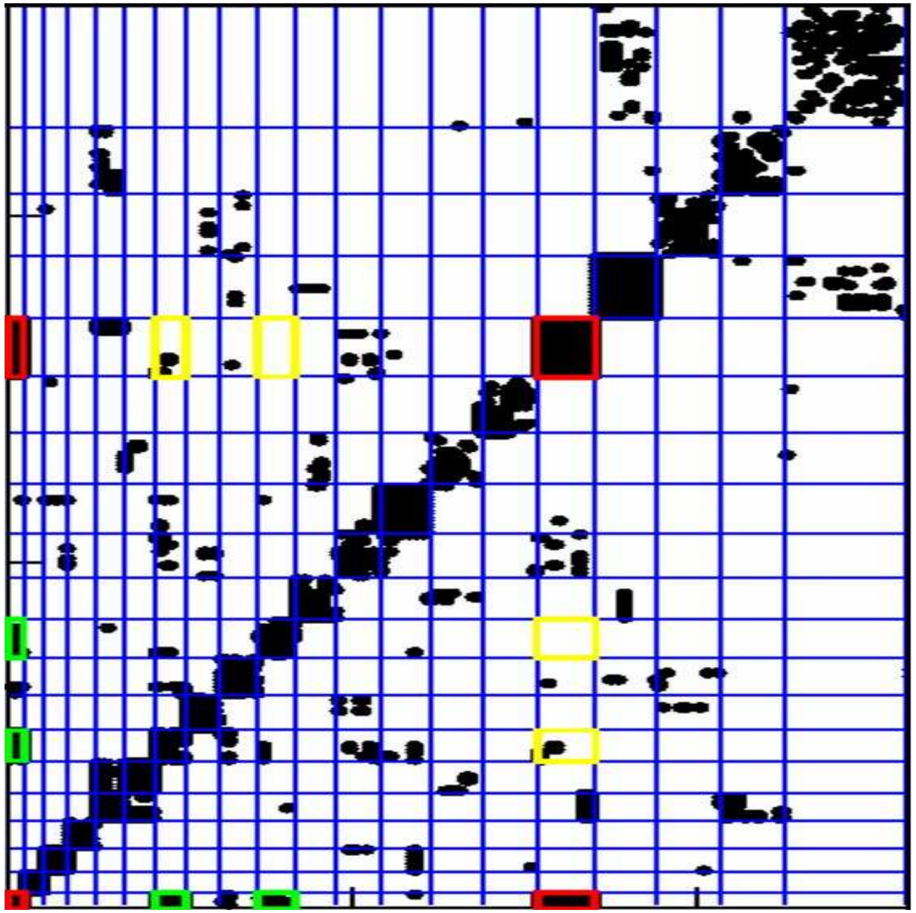
- The ICL criterion applied to a mixture with Bernoulli edge values select $\widehat{Q} = 21$ classes.
- Groups 1 to 20 gather reactions involving all the *same compound* either as a substrate or as a product.
- A compound (chorismate, pyruvate, ATP,*etc*) can be associated to each group.

Dot-plot representation.

- 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):

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

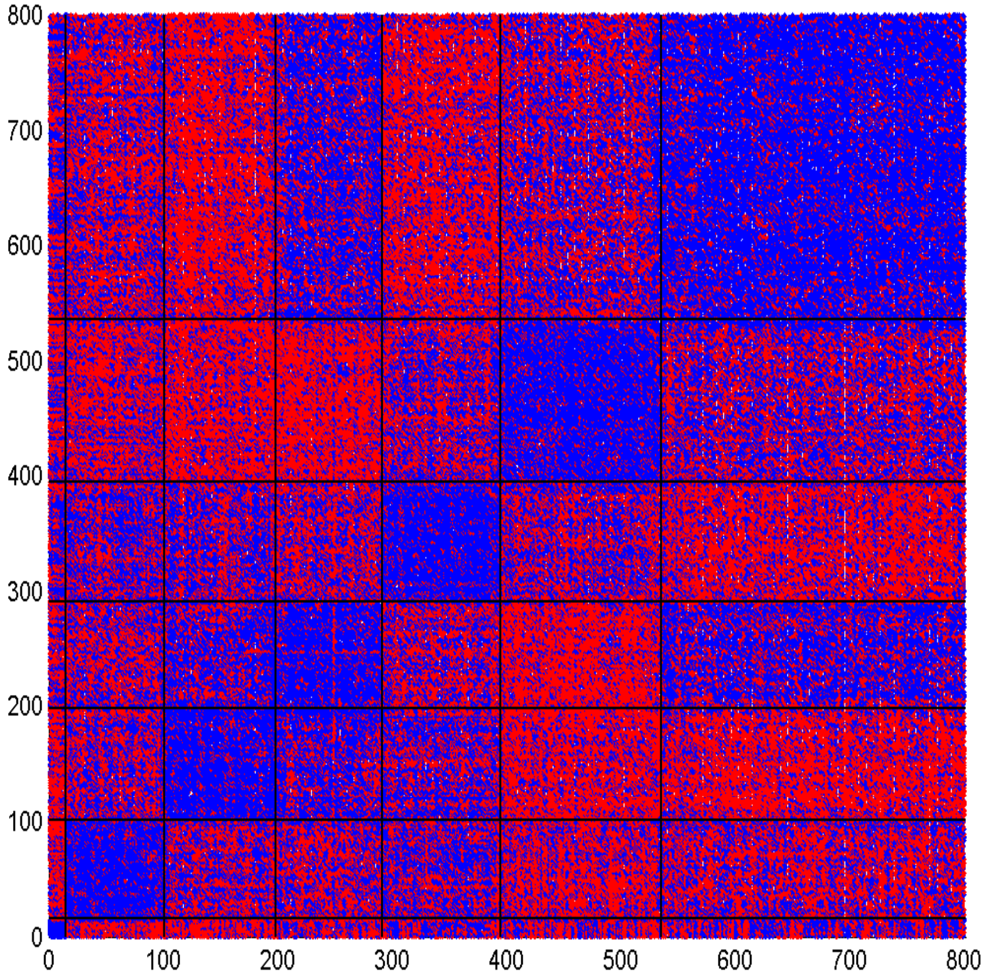
Adjacency matrix
(zoom on the *first 20 classes*)



4.2 - Gene regulations in *A. Thaliana*

Dataset. *Partial correlations* between the expression levels of 800 genes in various conditions (*Opgen-Rhein & Strimmer, 06*).

Dot-plot. Dot size = absolute correlation, Color = sign ($-$, $+$).



Results.

- Using a Gaussian model, we get $\widehat{Q} = 7$ classes.
- Groups are made of positively correlated genes.
- Between group correlations are weaker than within-group correlation and have different signs (see classes 3/4 with class 7).
- Total computational time for $Q = 1..15$ classes on a standard PC: 1h.

4.3 - Fungus - Tree 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 (*Data from C. Vacher, INRA*).

Projected graphs. For each species we define the projected graph:

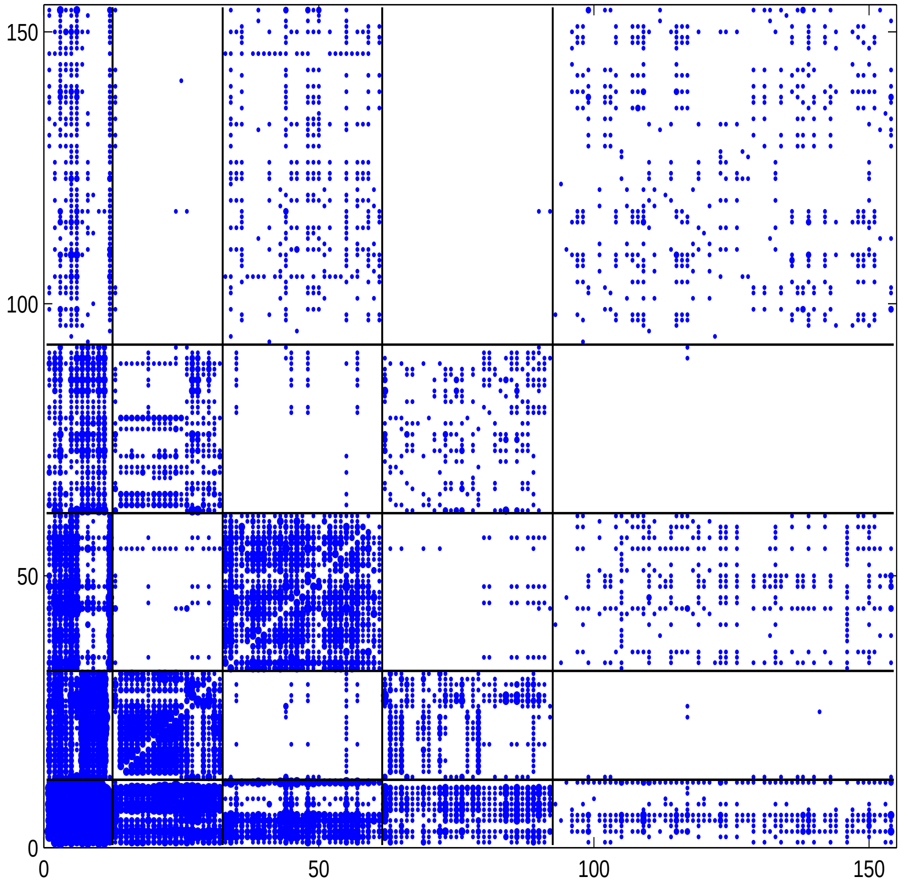
$$\begin{array}{lrcl} \text{for trees} & X_{tt'} & = & \text{Number of common fungi,} \\ \text{for fungi} & X_{ff'} & = & \text{Number of common trees.} \end{array}$$

Poisson model. For both species, we assume that the intensities have Poisson distributions: $X \sim \mathcal{P}(\lambda_{q\ell})$.

Number of classes. The ICL criterion selects

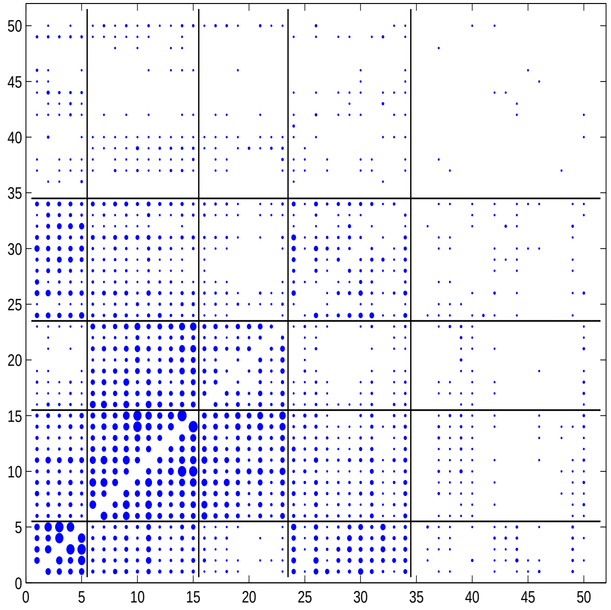
- 5 classes for trees
- and 6 classes for fungi.

Fungus network



- A group of generalist fungi is detected.
- Others are more specific.

Tree network



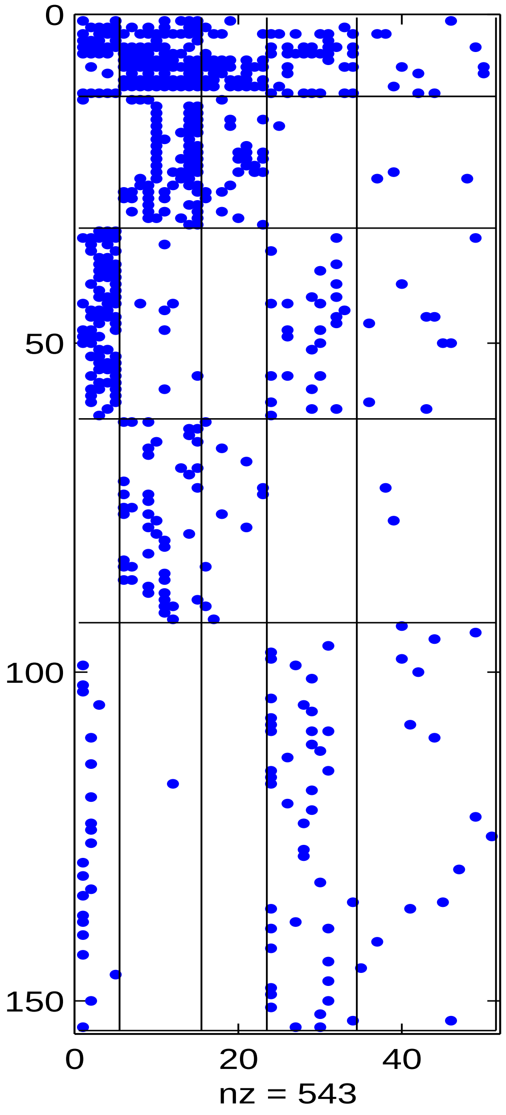
- Trees are mainly clustered according to the number of fungi they host.
- Tree groups are less contrasted.

Crossed clusterings

The comparison of the two clusterings exhibits *specific correspondences* between groups of fungi (rows) and trees (columns).

Work in progress. Compare these groups according to their phyla, the time of their introduction in Europe, *etc.*.

Biclustering. A direct clustering could be performed on the interaction matrix Fungi × Tree. The method proposed by *Govaert & Nadif (05)* also relies on a variational approach.



5 - Discussion & Work in progress

Inference for heterogeneous valued graphs

- Mixture models constitutes a natural way to describe heterogeneity in a network.
- The variational approach is a general and efficient alternative to MCMC algorithms.

Applications of the mixture model

- 'Realistic' heterogeneous networks can be simulated according to mixture models with given parameters.
- Once fitted to a given network, the mixture model allows to detect unexpetedly frequent motifs in biological (binary) networks (see *5.1*).

Extension

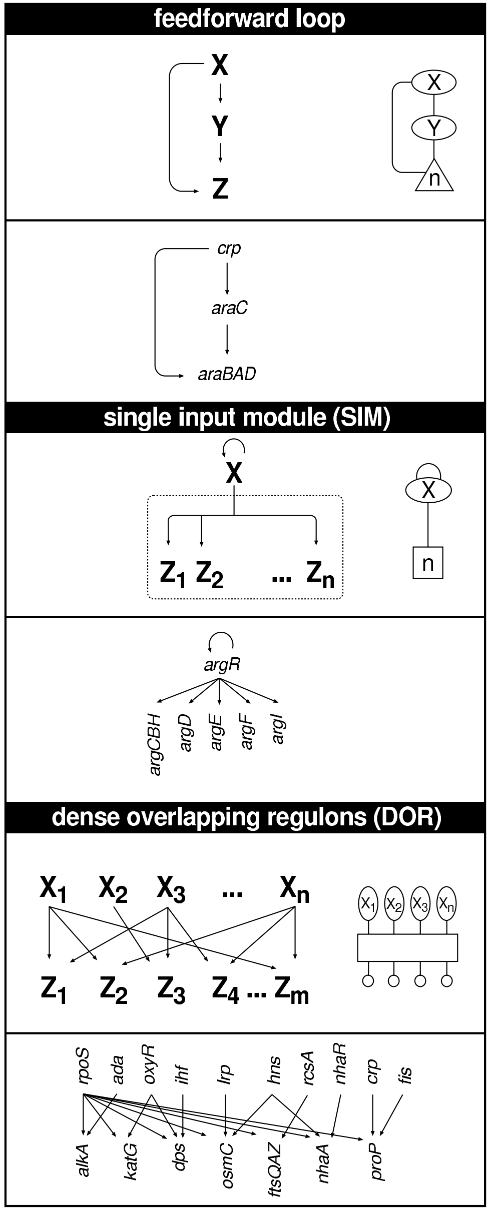
- The variational approach does not provide any measure of the precision of the estimates.
  → A variational Bayes approach would provide the (approximate) posterior distribution of the parameters (see *5.2*).

5.1 - Mixture model as a null model for heterogeneous networks

Looking for over-represented motifs in *E. coli* transcriptional network.

Strategy proposed by *Shen-Orr & al, 02*.

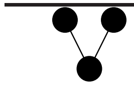
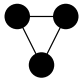
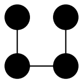
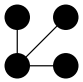
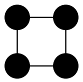
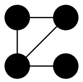
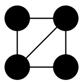
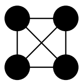
1. Count the number of occurrences $N_{\text{obs}}(\mathbf{m})$;
2. Resample a large number of random networks similar to *E.coli*'s one (using the edge swapping algorithm);
3. Estimate $\mathbb{E}N(\mathbf{m})$ and $\mathbb{V}N(\mathbf{m})$;
4. Derive a p-value implicitly based on a Gaussian approximation.



Direct computation using heterogenous models

Exact moments. For several heterogeneous models (mixture, EDD), we can get the exact formula for the mean $\mathbb{E}N$ and variance $\mathbb{V}N$ of the count (*Picard & al., 07*).

Distribution. Based on theoretical results (Erdös) and an analogy with sequence motifs, we fit a *compound Poisson* distribution to derive a p-value.

Motif	$N_{\text{obs}}(\mathbf{m})$	λ	$\frac{1}{(1-a)}$	p-value
	14 113	25.5	514.9	$3.36\ 10^{-1}$
	75	10.4	6.2	$2.87\ 10^{-1}$
	98 697	11.9	7 543.2	$3.46\ 10^{-1}$
	112 490	11.4	7 812.0	$1.85\ 10^{-1}$
	1 058	5.9	82.9	*9.34* 10^{-3}
	3 535	6.4	428.7	$2.22\ 10^{-1}$
	79	2.9	11.5	*2.56* 10^{-2}
	0	0.1	1.1	1.00

Results for *E. coli*'s network. 2 motifs appear to be unexpectedly frequent.

According to the permutation-based strategy, all of them are significantly over-represented!

5.2 - Variational Bayes approach

Beal & Ghahramani (2003) propose a

- variational
- Bayes
- E-M algorithm

to deal with for incomplete data models in the exponential family context.

1 - Variational approximation. Denoting $\boldsymbol{\theta}$ the set of parameters, for any distribution Q, we have

$$\log P(\mathbf{X}) \geq \int Q(\mathbf{Z}, \boldsymbol{\theta}) \log \frac{P(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{Q(\mathbf{Z}, \boldsymbol{\theta})} \mathrm{d}\mathbf{Z} \mathrm{d}\boldsymbol{\theta} =: \mathcal{F}(\mathbf{X}, Q).$$

2 - Optimal approximate distribution. If we choose $Q = Q_{\boldsymbol{\theta}} Q_{\mathbf{Z}}$, the optimal $Q_{\mathbf{Z}}$ and $Q_{\boldsymbol{\theta}}$ must satisfy

$$\begin{aligned} Q_{\mathbf{Z}}(\mathbf{Z}) &\propto \exp \int Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log P(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \\ Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) &\propto \exp \int Q_{\mathbf{Z}}(\mathbf{Z}) \log P(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \mathrm{d}\mathbf{Z}. \end{aligned}$$

This can be viewed as a *mean field* approximation.

3 - Exponential family. Suppose the complete likelihood belongs to the exponential family is and that parameter prior is conjugate

$$P(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) = f(\mathbf{X}, \mathbf{Z}) g(\boldsymbol{\theta}) \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})' \mathbf{u}(\mathbf{X}, \mathbf{Z})\},$$

$$P(\boldsymbol{\theta} | \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) g(\boldsymbol{\theta})^{\eta} \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})' \boldsymbol{\nu}\}.$$

Variational Bayes E-M algorithm

The optimal approximate conditional distribution $Q_{\boldsymbol{\theta}}$ and $Q_{\mathbf{Z}}$ must satisfy

$$
\begin{array}{rclcrcl}
Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) & \propto & g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})'\tilde{\boldsymbol{\nu}}\}, & \quad & \tilde{\eta} & = & \eta + 1, \\
\overline{\mathbf{u}}(\mathbf{X}) & = & \int Q_{\mathbf{Z}}(\boldsymbol{\theta})\mathbf{u}(\mathbf{X}, \mathbf{Z})\mathrm{d}\mathbf{Z}; & & \tilde{\boldsymbol{\nu}} & = & \boldsymbol{\nu} + \overline{\mathbf{u}}(\mathbf{X}, \mathbf{Z}), \\
Q_{\mathbf{Z}}(\mathbf{Z}) & \propto & f(\mathbf{X}, \mathbf{Z}) \exp\left\{\overline{\boldsymbol{\phi}}'\mathbf{u}(\mathbf{X}, \mathbf{Z})\right\}, & & \overline{\boldsymbol{\phi}} & = & \int Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\boldsymbol{\phi}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.
\end{array}
$$

Iterative algorithm. The variational Bayes E-M algorithm consists in alternative updates of $Q_{\boldsymbol{\theta}}$ ('E-step') and $Q_{\mathbf{Z}}$ ('M-step'):

$$
\begin{array}{rrcl}
\textbf{E-step:} & Q_{\boldsymbol{\theta}}^{t+1}(\boldsymbol{\theta}) & = & h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}^t) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\{[\boldsymbol{\phi}(\boldsymbol{\theta})]'\tilde{\boldsymbol{\nu}}^t\}; \\
\textbf{M-step:} & Q_{\mathbf{Z}}^{t+1}(\mathbf{Z}) & \propto & f(\mathbf{X}, \mathbf{Z}) \exp\left\{\left[\overline{\boldsymbol{\phi}}^{t+1}\right]' \mathbf{u}(\mathbf{X}, \mathbf{Z})\right\}.
\end{array}
$$

Application to mixture in networks?

Interest.

- Get *'confidence intervals'* for the parameter;
- Still avoids costly MCMC algorithms.

Problems.

- The approximate distribution $Q_{\mathbf{Z}}$ still needs to be restricted (e.g. $Q_{\mathbf{Z}} = \prod_i Q_{\mathbf{Z}_i}$);
- Initialisation (same as E-M);
- Uniqueness of the fix point?
- The *intrinsic identifiability problem* of mixture models...