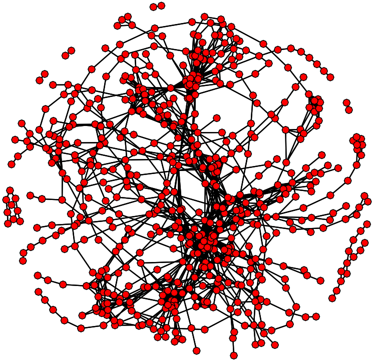
Random graph models for the clustering of nodes in networks and visualization

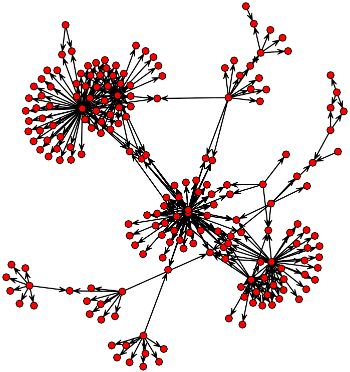
Pierre Latouche

Université Paris 1 Panthéon-Sorbonne
Laboratoire SAMM

NETBIO, AgroParisTech, 19/09/2014

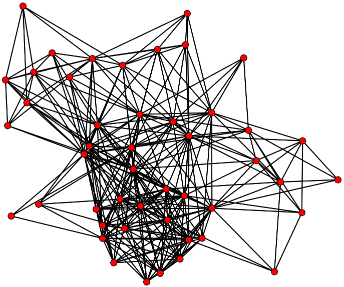


The metabolic network of bacteria *Escherichia coli* (Lacroix et al., 2006).



Subset of the yeast transcriptional regulatory network (Milo et al., 2002).

In Biology



Identification and classification of hubs in brain networks (O. Sporns et al., 2007).

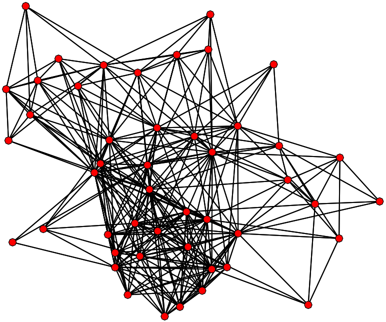
Goal

- **Two types of approaches** :
  - Unconvering clusters of vertices
  - Visualizing the network

Goal

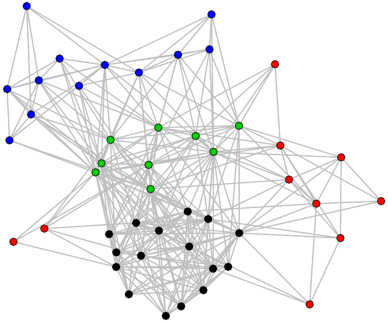
Extracting knowledge, summarizing the data

Graph clustering



Identification and classification of hubs in brain networks (O. Sporns et al., 2007).

Graph clustering



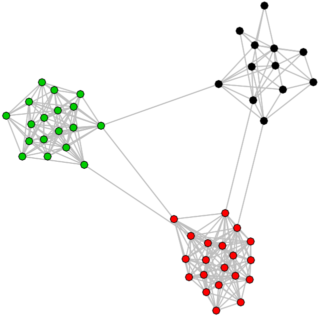
Identification and classification of hubs in brain networks (O. Sporns et al., 2007).

Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure

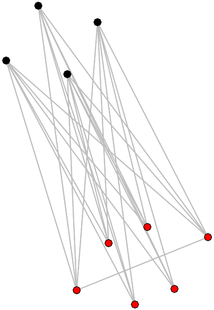
Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure



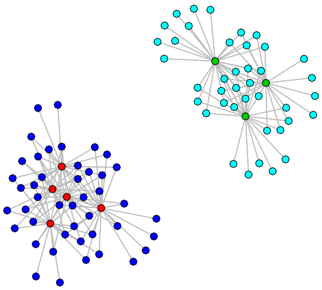
Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure

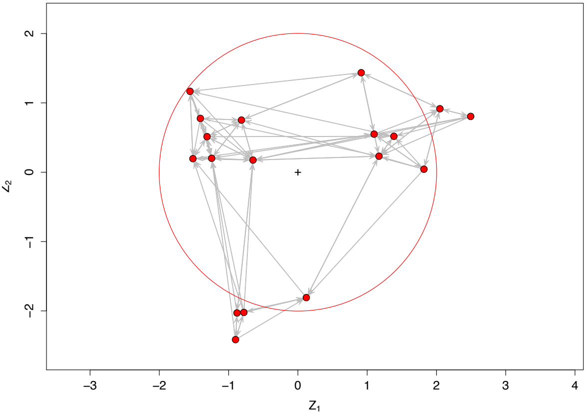


Graph clustering

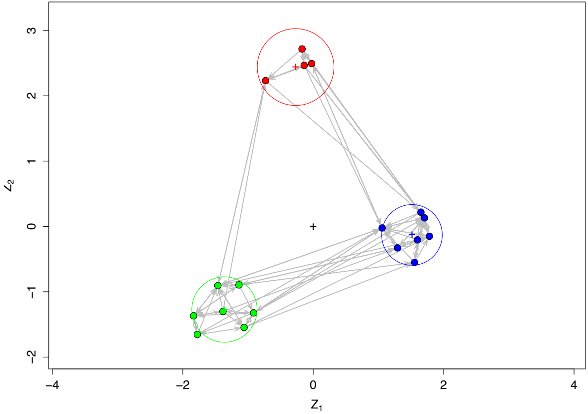
- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure



Visualization



Latent position model (LPCM) (Hoff et al. 2002).



Latent position cluster model (LPCM) (Handcock et al. 2007).

Random graph models

Erdös-Rényi model (1959)
Latent position model (Hoff et al. 2002)
Latent position cluster model (Handcock et al. 2007)
Stochastic block model (SBM) (Nowicki and Snijders 2001)
Mixed membership SBM (Airoldi et al. 2008)
Overlapping SBM (Latouche et al. 2011)
...
$\hookrightarrow$ W-graph model ???

Contents

Contents

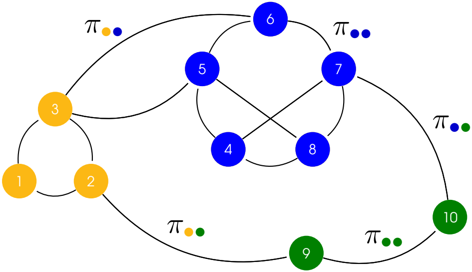
Stochastic block model (SBM)

- Nowicki and Snijders (2001)
  - Earlier work : Govaert et al. (1977)
- $\mathbf{Z}_i$ independent hidden variables :
  - $\mathbf{Z}_i \sim \mathcal{M}\Big(1,\ \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)\Big)$
  - $Z_{ik} = 1$: vertex i belongs to class k
- $\mathbf{X} \,|\, \mathbf{Z}$ edges drawn independently :

$$X_{ij} | \{i \in k, j \in l\} \sim \mathcal{B}(\pi_{kl})$$

- A mixture model for graphs :

$$X_{ij} \sim \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k \alpha_l \mathcal{B}(\pi_{kl})$$



$\pi_{\bullet\bullet}$
6
$\pi_{\bullet\bullet}$
5
7
3
$\pi_{\bullet\bullet}$
4
8
1
2
10
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
9

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$

Problem

$p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable (no conditional independence)

Variational EM

Daudin et al. (2008)

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) \right\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$

Problem

$p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable (no conditional independence)

Variational EM

Daudin et al. (2008)

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) \right\}$
    $\hookrightarrow K^N$ terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$

Problem

$p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable (no conditional independence)

Variational EM

Daudin et al. (2008)

Model selection

Criteria

Since $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable, we *cannot* rely on:

- $AIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - M$
- $BIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - \frac{M}{2} \log \frac{N(N-1)}{2}$

ICL

Biernacki et al. (2000) $\hookrightarrow$ Daudin et al. (2008)

Variational Bayes EM $\hookrightarrow ILvb$

Latouche et al. (2012)

Others

McDaid et al. (2012), ...

Model selection

Criteria

Since $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable, we *cannot* rely on:

- $AIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - M$
- $BIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - \frac{M}{2} \log \frac{N(N-1)}{2}$

ICL

Biernacki et al. (2000) $\hookrightarrow$ Daudin et al. (2008)

Variational Bayes EM $\hookrightarrow$ $ILvb$

Latouche et al. (2012)

Others

McDaid et al. (2012), ...

Bayesian framework

- **Conjugate prior distributions** :
  - $p\Big(\boldsymbol{\alpha} \,|\, \mathbf{n}^0 = \{n_1^0, \dots, n_K^0\}\Big) = \mathrm{Dir}(\boldsymbol{\alpha};\, \mathbf{n}^0)$
  - $p\Big(\boldsymbol{\Pi} \,|\, \boldsymbol{\eta}^0 = (\eta_{kl}^0), \boldsymbol{\zeta}^0 = (\zeta_{kl}^0)\Big) = \prod_{k \le l} \mathrm{Beta}(\pi_{kl};\, \eta_{kl}^0, \zeta_{kl}^0)$
- **Non informative Jeffreys prior** :
  - $n_k^0 = 1/2$
  - $\eta_{kl}^0 = \zeta_{kl}^0 = 1/2$

- $p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi} \,|\, \mathbf{X})$ not tractable

Decomposition

$$\log p(\mathbf{X}) = \mathcal{L}\left(q\right) + \mathrm{KL}\left(q(\cdot) \,||\, p(\cdot \,|\, \mathbf{X})\right)$$

where

$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi})} \right\} d\,\boldsymbol{\alpha}\, d\,\boldsymbol{\Pi}$$

Factorization

$$q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = q(\boldsymbol{\alpha})q(\boldsymbol{\Pi})q(\mathbf{Z}) = q(\boldsymbol{\alpha})q(\boldsymbol{\Pi}) \prod_{i=1}^{N} q(\mathbf{Z}_i)$$

Variational Bayes EM
Latouche et al. (2009)

E-step

- $q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i;\ 1, \boldsymbol{\tau_i} = \{\tau_{i1}, \ldots, \tau_{iK}\})$

M-step

- $q(\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha;\ \mathbf{n})$
- $q(\mathbf{\Pi}) = \prod_{k \leq l}^{K} \mathrm{Beta}(\pi_{kl};\ \eta_{kl}, \zeta_{kl})$

A new model selection criterion : ILvb
Latouche et al. (2012)

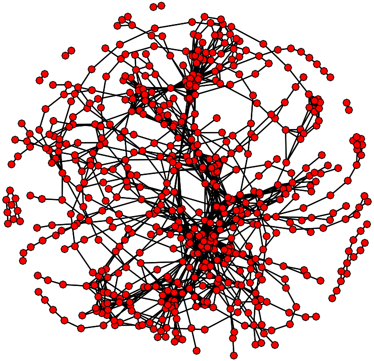
- $\log p(\mathbf{X} \,|K) = \mathcal{L}\,(q) + \mathrm{KL}(...)$
- After convergence, use $\mathcal{L}\,(q)$ as an approximation of $\log p(\mathbf{X} \,|K)$

ILvb

$$IL_{vb} = \log \left\{ \frac{\Gamma(\sum_{k=1}^{K} n_k^0) \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\sum_{k=1}^{K} n_k) \prod_{k=1}^{K} \Gamma(n_k^0)} \right\} + \sum_{k \leq l}^{K} \log \left\{ \frac{\Gamma(\eta_{kl}^0 + \zeta_{kl}^0)\Gamma(\eta_{kl})\Gamma(\zeta_{kl})}{\Gamma(\eta_{kl} + \zeta_{kl})\Gamma(\eta_{kl}^0)\Gamma(\zeta_{kl}^0)} \right\} - \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}$$

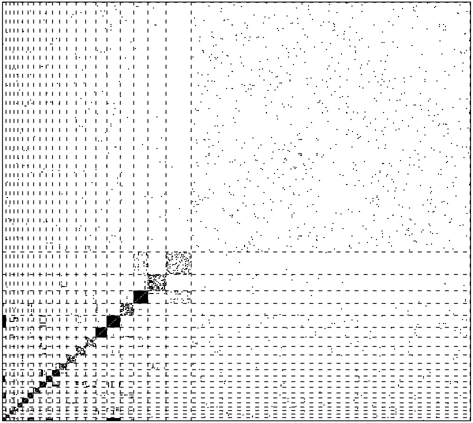
- Lacroix et al. (2006)
- Lab : Biométrie et Biologie Évolutive (Lyon 1)
- Represents pathways of biochemical reactions
- 605 vertices, 1782 edges

The metabolic network of ecoli



The metabolic network of bacteria *Escherichia coli* (Lacroix et al., 2006).

Results (1)

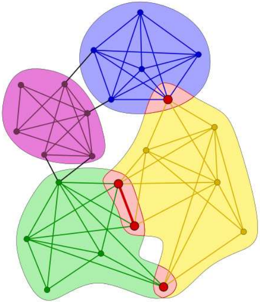


Dot plot representation of the metabolic network after classification of the vertices into $K_{VB} = 22$ classes.

Results (2)

- Among the classes, eight are cliques
- Six have within probability connectivity greater than 0.5
- Cliques and pseudo-cliques gather reactions involving a same compound
  - Responsible for cliques : chorismate, pyruvate, L-aspartate, L-glutamate, D-glyceraldehyde-3-phosphate and ATP
- Classes 1 and 17 both associated to pyruvate

Contents



Palla et al. (2006)

Problem

The stochastic block model (SBM) and most existing methods assume that each vertex belongs to a single class

Stochastic Block Model (SBM)

- Nowicki and Snijders (2001)
- $\mathbf{Z}_i$ independent hidden variables :

$$\mathbf{Z}_i \sim \mathcal{M}\Big(1,\ \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)\Big)$$

Overlapping Stochastic Block model (OSBM)

- Latouche et al. (2011)
- Z_{ik} independent hidden variables :

$$\mathbf{Z}_i \sim \prod_{k=1}^{K} \mathcal{B}(Z_{ik};\ \alpha_k) = \prod_{k=1}^{K} \alpha_k^{Z_{ik}} (1 - \alpha_k)^{1 - Z_{ik}}$$

Overlapping Stochastic Block model (OSBM)

- Latouche et al. (2011)
- $\mathbf{X} \,|\, \mathbf{Z}$ edges drawn independently :

$$X_{ij} \,|\, \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}\big(X_{ij};\, \mathbf{\Pi}_{\mathbf{Z}_i,\mathbf{Z}_j})\big)$$

- $\mathbf{\Pi}_{\mathbf{Z}_i,\mathbf{Z}_j} = g\big(a_{\mathbf{Z}_i,\mathbf{Z}_j}\big)$
- $a_{\mathbf{Z}_i,\mathbf{Z}_j} = \underbrace{\mathbf{Z}_i^\intercal \mathbf{W} \mathbf{Z}_j}_{i \leftrightarrow j} + \underbrace{\mathbf{Z}_i^\intercal \mathbf{U}}_{i \rightarrow ?} + \underbrace{\mathbf{V}^\intercal \mathbf{Z}_j}_{? \rightarrow j} + \underbrace{W^*}_{\text{bias}}$
- $g(t) = 1/\left(1 + \exp(-t)\right)$ is the logistic function

OSBM

- $\tilde{\mathbf{Z}}_i = (\mathbf{Z}_i, 1)^\intercal$
- $\tilde{\mathbf{W}} = \begin{pmatrix} \mathbf{W} & \mathbf{U} \\ \mathbf{V}^\intercal & W^* \end{pmatrix}$
- $a_{\mathbf{Z}_i,\mathbf{Z}_j} = \tilde{\mathbf{Z}}_i^\intercal \, \tilde{\mathbf{W}} \, \tilde{\mathbf{Z}}_j$
- Parameter set : $\left\{\boldsymbol{\alpha}, \tilde{\mathbf{W}}\right\}$

Bayesian framework

- **Conjugate prior distributions** :
  - $p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathrm{Beta}(\alpha_k;\, \eta_k^0, \zeta_k^0)$
  - $p(\tilde{\mathbf{W}}^{\mathrm{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\mathrm{vec}};\, \tilde{\mathbf{W}}_0^{\mathrm{vec}}, \mathbf{S}_0)$
- The vec operator : if

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

then

$$\mathbf{A}^{\mathrm{vec}} = \begin{pmatrix} A_{11} \\ A_{21} \\ A_{12} \\ A_{22} \end{pmatrix}$$

Bayesian framework

- $\mathbf{x}^\intercal \, \mathbf{A} \, \mathbf{y} = (\mathbf{y} \otimes \mathbf{x})^\intercal \, \mathbf{A}^{\mathrm{vec}}$
- In practice : set $\tilde{\mathbf{W}}_0^{\mathrm{vec}} = \mathbf{0}$ and $\mathbf{S}_0 = \frac{\mathbf{I}}{\beta}$

Problem

$p(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}} \mid \mathbf{X})$ not tractable

q Transformation

Decomposition

$$\log p(\mathbf{X}) = \mathcal{L}(r) + \mathrm{KL}(r||p)$$

where

$$\mathcal{L}(r) = \sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \log \Big(\frac{p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}})}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})}\Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}$$

Lower bound

$$\log p(\mathbf{X}) \geq \mathcal{L}(r)$$

Problem

$\mathcal{L}(r)$ has a too complex form $\hookrightarrow$ no variational Bayes EM algorithm ??

Local bound

- Use the bound of Jaakkola and Jordan (2000) for Bayesian logistic regression

$$\log p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) \geq \log h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}), \forall\, \boldsymbol{\xi} \in \mathbb{R}^{N \times N}$$

where

$$\log h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) = \sum_{i \neq j}^{N} \Big\{ (X_{ij} - \frac{1}{2}) a_{\mathbf{Z}_i, \mathbf{Z}_j} - \frac{\xi_{ij}}{2} + \log g(\xi_{ij}) - \lambda(\xi_{ij})(a^2_{\mathbf{Z}_i, \mathbf{Z}_j} - \xi^2_{ij}) \Big\}$$

and

$$\lambda(\xi) = \frac{1}{4\xi} \tanh(\frac{\xi}{2}) = \frac{1}{2\xi} \{ g(\xi) - \frac{1}{2} \}$$

Lower Bound

$$\log p(\mathbf{X}) = \log \left\{ \sum_{\mathbf{Z}} \int p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}} \right\}$$

$$\geq \mathcal{L}(\boldsymbol{\xi})$$

where

$$\mathcal{L}(\boldsymbol{\xi}) = \log \left\{ \sum_{\mathbf{Z}} \int h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}} \right\}$$

Decomposition

$$\mathcal{L}(\boldsymbol{\xi}) = \mathcal{L}(r;\ \boldsymbol{\xi}) + \mathrm{KL}(r||p)$$

where

$$\mathcal{L}(r;\ \boldsymbol{\xi})$$
$$= \sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \log \Big(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}})}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})}\Big) d\alpha d\, \tilde{\mathbf{W}}$$

Lower bound

$$\log p(\mathbf{X}) \geq \mathcal{L}(\boldsymbol{\xi}) \geq \mathcal{L}(r;\ \boldsymbol{\xi})$$

Inference

Local optimization

- $\boldsymbol{\xi} = \text{argmax}_{\boldsymbol{\xi}} \mathcal{L}(r;\, \boldsymbol{\xi})$

E-step

- $r(Z_{ik}) = \mathcal{B}(Z_{ik};\, \tau_{ik})$

M-step

- $r(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \text{Beta}(\alpha_k;\, \eta_k^N, \zeta_k^N)$
- $r(\tilde{\mathbf{W}}^{vec}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\, \tilde{\mathbf{W}}_N^{\text{vec}}, \mathbf{S}_N)$

- After convergence, use $\mathcal{L}(\hat{r};\, \hat{\boldsymbol{\xi}})$ as an approximation of $\log p(\mathbf{X}\,|K)$

ILosbm

$$IL_{osbm} = \mathcal{L}(\hat{r};\, \hat{\boldsymbol{\xi}})$$

L_2 regularization

$$p(\tilde{\mathbf{W}}^{\text{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\, \mathbf{0}, \frac{\mathbf{I}}{\beta})$$

- β too small $\hookrightarrow$ overfit
- β too large $\hookrightarrow$ IL_{osbm} maximized for very large values of K

Question

Can we estimate β from the data ?

Bayesian framework

- **Conjugate prior distributions :**
  - $p(\tilde{\mathbf{W}}^{\text{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\ \mathbf{0}, \frac{\mathbf{I}}{\beta})$
  - $p(\beta) = \text{Gamma}(\beta;\ a_0, b_0)$

- Use a variational Bayes EM algorithm to maximize:

$$\mathcal{L}(r;\, \boldsymbol{\xi}) =$$

$$\sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta) \log \Big(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) p(\beta)}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta)} \Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}\, d$$

- $r(\beta) = \mathrm{Gamma}(\beta;\, a_N, b_N)$, where

$$a_N = a_0 + \frac{(K+1)^2}{2}$$

and

$$b_N = b_0 + \frac{1}{2}\mathrm{Tr}\big(S_N + (\tilde{\mathbf{W}}_N^{\mathrm{vec}})^\intercal\, \tilde{\mathbf{W}}_N^{\mathrm{vec}}\big)$$

Criterion

$$IL_{osbm} = \mathcal{L}(\hat{r};\, \hat{\boldsymbol{\xi}})$$

- Use a variational Bayes EM algorithm to maximize:

$$\mathcal{L}(r;\, \boldsymbol{\xi}) =$$

$$\sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta) \log \Big(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) p(\beta)}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta)} \Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}\, d$$

- $r(\beta) = \mathrm{Gamma}(\beta;\, a_N, b_N)$, where

$$a_N = a_0 + \frac{(K+1)^2}{2}$$

and

$$b_N = b_0 + \frac{1}{2} \mathrm{Tr}\big(S_N + (\tilde{\mathbf{W}}_N^{\mathrm{vec}})^{\intercal}\, \tilde{\mathbf{W}}_N^{\mathrm{vec}}\,\big)$$

Criterion

$$IL_{osbm} = \mathcal{L}(\hat{r};\, \hat{\boldsymbol{\xi}})$$

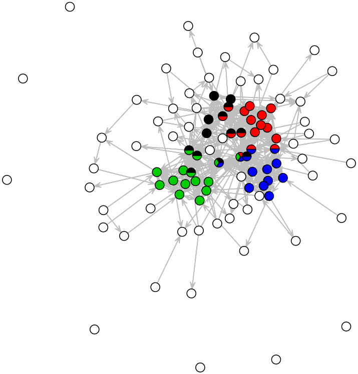
Experiments on simulated data

- Community structures (affiliation) :

$$\mathbf{W} = \begin{pmatrix} \boldsymbol{\lambda} & -\epsilon & \dots & -\epsilon \\ -\epsilon & \boldsymbol{\lambda} & & \vdots \\ \vdots & & \ddots & -\epsilon \\ -\epsilon & \dots & -\epsilon & \boldsymbol{\lambda} \end{pmatrix}$$

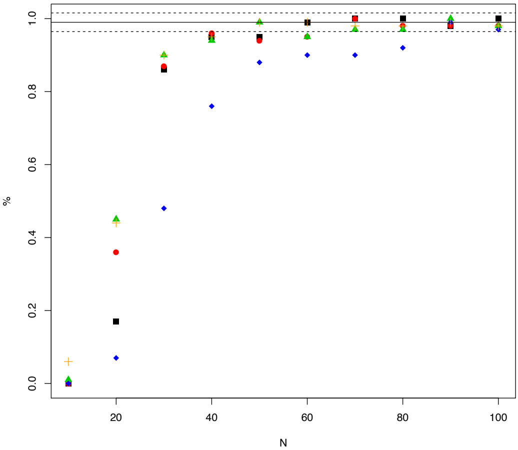
- $\mathbf{U} = \mathbf{V} = (\epsilon, \dots, \epsilon)$

Community structures



Example of an overlapping stochastic block model (OSBM) network with community structures.

- $\lambda = 1.5$, $\epsilon = 1$, $W^* = -2$, $K = 3$, $\alpha_k = 1/K$, $N \in \{10, 20, \ldots, 100\}$ simulated 100 networks



Proportions of the simulations where 99% credibility intervals obtained with the VBEM algorithm contain the true value of the parameters.

Experiments on simulated data

- $N = 100$
- $\lambda \in \{6, 4, 3.5\}$
- $\epsilon = 1$
- $W^* = -5.5$
- $\alpha_k \propto a^k$
  - $a = 1$: balanced proportions
  - $a = 0.7$: unbalanced proportions
- $K_{True} \in \{2, \dots, 7\}$
- 100 simulations

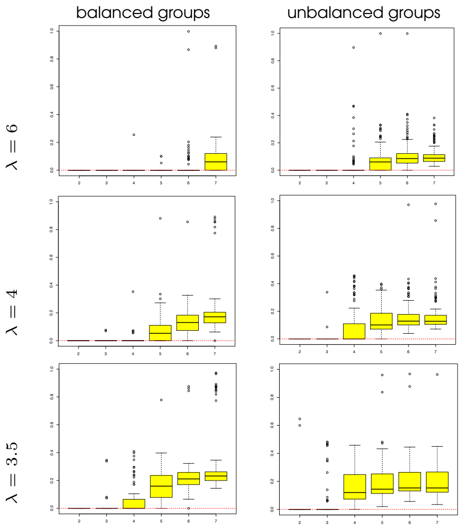
How to evaluate the clustering ?

- Compute $\mathbf{P} = \mathbf{Z}\,\mathbf{Z}^\intercal$ and $\hat{\mathbf{P}} = \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\intercal$:
  - invariant to column permutations of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$
  - number of shared clusters between each pair of vertices
- Compute

$$\sqrt{\frac{1}{N(N-1)}\sum_{i\neq j}|(\mathbf{Z}\,\mathbf{Z}^\intercal)_{ij} - (\hat{\mathbf{Z}}\hat{\mathbf{Z}}^\intercal)_{ij}|}$$

Results

		balanced groups							unbalanced groups						
		2	3	4	5	6	7	8	2	3	4	5	6	7	8
$\lambda = 6$	2	**100**	0	0	0	0	0	0	**100**	0	0	0	0	0	0
	3	0	**100**	0	0	0	0		0	**100**	0	0	0	0	0
	4	0	0	**99**	0	1	0	0	0	6	**85**	5	3	1	0
	5	0	0	2	**98**	0	0	0	0	3	34	**50**	8	4	1
	6	0	0	0	8	**85**	6	1	0	0	29	49	**15**	6	1
	7	0	0	0	1	24	**56**	19	0	0	30	50	13	**6**	1
$\lambda = 4$	2	**100**	0	0	0	0	0	0	**100**	0	0	0	0	0	0
	3	0	**100**	0	0	0	0	0	0	**99**	1	0	0	0	0
	4	0	0	**99**	1	0	0	0	0	14	**68**	9	7	2	0
	5	0	0	4	**79**	14	1	2	0	18	50	**22**	4	6	0
	6	0	0	1	22	**49**	22	6	0	20	46	16	**13**	4	1
	7	0	0	0	16	47	**24**	13	0	22	56	14	5	**3**	0
$\lambda = 3.5$	2	**100**	0	0	0	0	0	0	**98**	2	0	0	0	0	0
	3	0	**98**	2	0	0	0	0	1	**91**	7	0	1	0	0
	4	0	0	**87**	9	3	1	0	1	43	**32**	16	4	1	3
	5	0	0	15	**44**	26	12	3	2	34	44	**9**	8	3	0
	6	0	1	11	28	**22**	25	13	0	47	32	15	**5**	1	0
	7	0	0	6	34	28	**17**	15	2	30	46	14	5	**3**	0

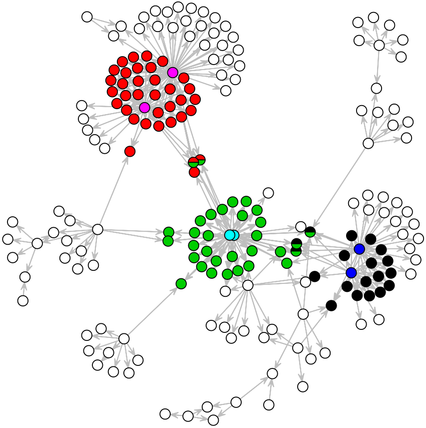

balanced groups
unbalanced groups
λ = 6
λ = 4
λ = 3.5

Experiments on yeast transcription network

cluster	size	genes
1	2	STE12 TEC1
2	35	DDR48 YLR042C YPS1 YPL114W YNL159C YNL051W YLR414C YJL142C YJL017W YHR156C **TKL2** YGR149W YEL033W YDL222C YBR070C WSC2 TSL1 TOS11 YHL021C **SSA4** SRL1 SRD1 SPO12 SFP1 RTS2 RTA1 PST1 PRM5 PGU1 MPT5 MID2 HTB2 GAT4 DHH1 BNI5
3	2	MSN4 MSN2
4	35	UBI4 SSA3 HSP26 HSP12 HSP104 **CTT1** TPS1 DOG2 GRE3 **SSA4** YNL077W YGR086C TTR1 SPS100 SOD2 RAS2 PNC1 **PGM2** MTC2 MDJ1 HXK1 **HSP78** HSP42 HOR2 GRX1 **TKL2** GLO1 GLK1 GDH3 CPH1 ARA1 ALD3 ALD2 YKL151C DDR2
5	2	YAP1 SKN7
6	19	YMR318C **CTT1** TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C **HSP78** CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 **PGM2**

Table : Classification of the genes into $K = 6$ clusters. Genes in bold belong to multiple clusters.

Experiments on yeast transcription network



Contents

W-graph model $\hookrightarrow$ *graphon* (Borgs et al. 2007)

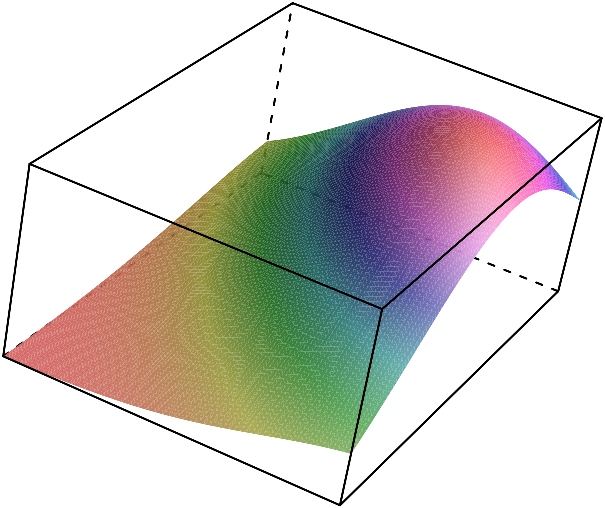
Graphon function

- $W : [0,1]^2 \rightarrow [0,1]$
- $W(u,v)$: probability that nodes (i,j) with coordinates u and v connect

Sampling

- Sample $U_i \sim \mathcal{U}(0,1), \ \forall i$
- Sample edges $X_{ij} | U_i, U_j \sim \mathcal{B}\left(W(U_i, U_j)\right)$

W-graph model



Example of a graphon function.

Estimation ?

Problem Given a network, how to estimate the graphon function ?

Approach

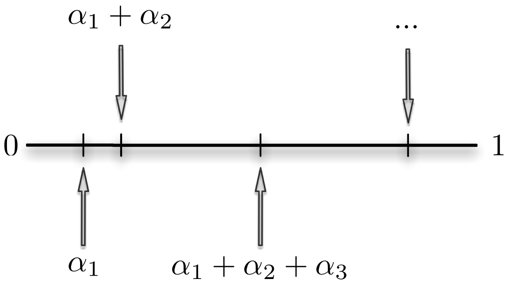
Use SBM + inference strategies

SBM and W-graph models

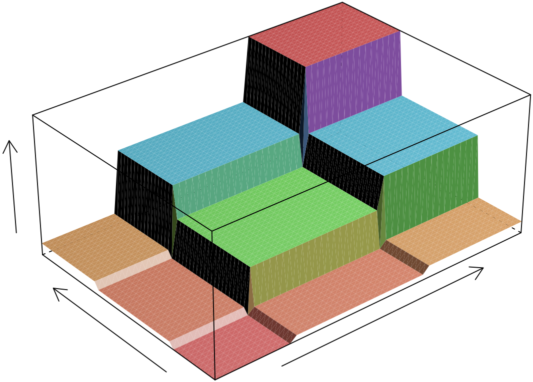
- SBM : special case of a W-graph model
- Recall : SBM : K + $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ + $\boldsymbol{\Pi} = \left(\pi_{kl}\right)_{kl}$

Connection

- Define $\sigma_k = \sum_{l=1}^{k} \alpha_l, \ \forall k$
- $C_{\boldsymbol{\alpha}}(u) = 1 + \sum_{k=1}^{K} \mathbb{I}\{\sigma_k \leq u\}$
- $W(u, v) = \pi_{C_{\boldsymbol{\alpha}}(u), C_{\boldsymbol{\alpha}}(v)}$



SBM and W-graph models



Graphon function of a SBM model with $K = 3$ classes.

Estimation

Bayesian approach

Estimate the posterior distribution of $W(u, v)$

- $\hat{K}$ with $ILvb$ (Latouche et al. 2009)
- $\boldsymbol{\alpha} \,|\, \mathbf{X}$ and $\boldsymbol{\Pi} \,|\, \mathbf{X}$ with VBEM (Latouche et al. 2012)

Posterior distribution of W

$$
\begin{aligned}
\tilde{p}(w|\,\mathbf{X},K) &= \tilde{p}(\pi_{C_{\boldsymbol{\alpha}}(u),C_{\boldsymbol{\alpha}}(v)}|\,\mathbf{X},K) \\
&= \sum_{k\le l}^{K} \tilde{p}(\pi_{k,l}|\,\mathbf{X},K)\tilde{Pr}(C(u)=k,C(v)=l|\,\mathbf{X},K) \\
&= \sum_{k\le l}^{K} \mathrm{Beta}(w;\eta_{kl},\zeta_{kl})\tilde{\mathrm{Pr}}(C(u)=k,C(v)=l|\,\mathbf{X},K)
\end{aligned}
$$

$$
\tilde{\mathrm{Pr}}(C(u)=k,C(v)=l|\,\mathbf{X},K) = \tilde{\mathrm{Pr}}(\sigma_{k-1}<u<\sigma_k,\sigma_{l-1}<v<\sigma_l|\,\mathbf{X},K)
$$

Proposition

For given $(u,v)\in[0,1]^2$, $u\le v$, using a SBM with K groups, the variational Bayes approximate pdf of $W(u,v)$ is $\tilde{p}(w|\,\mathbf{X},K) =$

$$
\sum_{k\le\ell} \mathrm{Beta}\Big(w;\eta_{k\ell},\zeta_{k\ell}\Big)\Big[F_{k-1,l-1}(u,v;\mathbf{a}) - F_{k,l-1}(u,v;\mathbf{a})
$$

Posterior distribution of W

Proposition

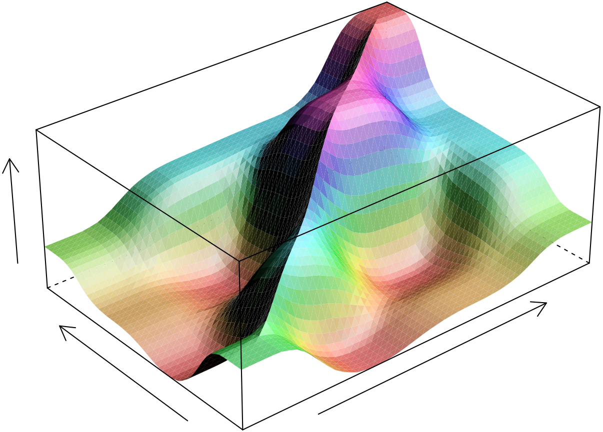
For given $(u, v) \in [0, 1]^2$, $u \leq v$, using a SBM with K groups, the variational Bayes approximate pdf of $W(u, v)$ is $\tilde{p}(w|\,\mathbf{X}, K) =$

$$\sum_{k \leq \ell} \mathrm{Beta}\Big(w; \eta_{k\ell}, \zeta_{k\ell}\Big) \Big[F_{k-1,l-1}(u, v; \mathbf{a}) - F_{k,l-1}(u, v; \mathbf{a}) - F_{k-1,l}(u, v; \mathbf{a}) + F_{k,l}(u, v; \mathbf{a})\Big]$$

where

- $F_{k,l}(u, v; \mathbf{a})$ is the joint cdf of (σ_k, σ_l) when $\boldsymbol{\alpha}$ has Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$

Macaque



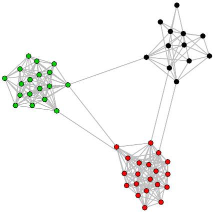
Estimation of the graphon function of the macaque cortex network.

References

- K. Nowicki and T.A.B. Snijders (2001), Estimation and prediction for stochastic blockstructures. 96, 1077-1087
- E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing (2008), Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9, 1981-2014
- J-J. Daudin, F. Picard et S. Robin (2008), A mixture model for random graphs. Statistics and Computing, 18, 2, 151-171
- P. Latouche, E. Birmelé, C. Ambroise (2011), Overlapping stochastic block models with application to the French political blogosphere network. Annals of Applied Statistics, 5, 1, 309-336
- P. Latouche, E. Birmelé, C. Ambroise (2012), Variational Bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12, 1, 93-115

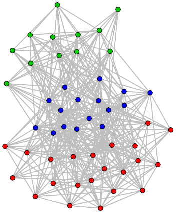
- **Two topological structures** :
  - Affiliation :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon \\ \epsilon & \lambda & & \vdots \\ \vdots & & \ddots & \epsilon \\ \epsilon & \dots & \epsilon & \lambda \end{pmatrix}$$



  - Affiliation and a class of hubs :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon & \lambda \\ \epsilon & \lambda & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \lambda & \dots & \dots & \dots & \lambda \end{pmatrix}$$



(a) $Q_{True} \backslash Q_{VBMOD}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	0	**100**	0	0
6	0	0	0	0	**97**	3
7	0	0	0	2	14	**84**

(b) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	0	**99**	1	0
6	0	0	4	23	**73**	0
7	0	2	14	44	27	**13**

(c) $Q_{True} \backslash Q_{VBMOD}$

	2	3	4	5	6	7
3	95	**0**	3	0	0	2
4	1	95	**4**	0	0	0
5	0	0	94	**6**	0	0
6	0	0	1	83	**16**	0
7	0	0	2	15	78	**5**

(d) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	2	**98**	0	0
6	0	0	1	29	**70**	0
7	0	0	3	34	45	**18**

The French blogosphere network

	UMP	UDF	liberal	PS	analysts	others
cluster 1	30 + 3	0 + 1	0	0	0 + 1	0
cluster 2	2 + 3	29 + 1	0	0	1 + 3	0
cluster 3	0	0	24	0	1 + 1	0
cluster 4	0	0 + 2	0	40	0 + 4	1
outliers	5	1	1	17	5	30

Clustering of the blogs into $Q = 4$ clusters using OSBM. 196 vertices, 2864 edges.

The French blogosphere network