# Clustering the nodes of a graph

J.-B. Leger and J.-J. Daudin

February 9, 2012

- Why ?
- Exemples
- Basic definitions
    - Communities and SES
    - Similarity transformation
- Methods based on a algorithm
    - Markov Cluster algorithm (MCL)
    - Pons-Latapy distance
    - Spectral Clustering
    - Edge-Betweeness
    - Hierarchical agglomerative clustering algorithm
- Methods based on a optimization criterion
    - Modularity criterion
    - Cut cost
- Methods using a probabilistic model
    - SBM
        - Karate Club
        - PPI human cells
    - Continuous Stochastic bloc model (CSBM)
- Summary
- Benchmark

# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

Methods based on a optimization criterion

Methods using a probabilistic model

Summary

Benchmark

# PPI of *Saccharomyces C.*

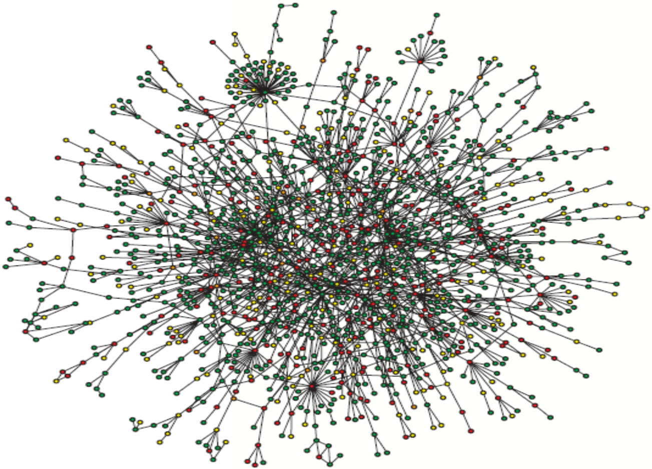


Figure 2 | **Yeast protein interaction network.** A map of protein–protein interactions[18] in *Saccharomyces cerevisiae*, which is based on early yeast two-hybrid measurements[23], illustrates that a few highly connected nodes (which are also known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from REF. 18 © Macmillan Magazines Ltd.

# Questions

Find some structure

- identify ”independent modules”
- classify the nodes into few classes of nodes with similar connections, i.e. connected to the same nodes.

Same questions for for social and ecological networks.

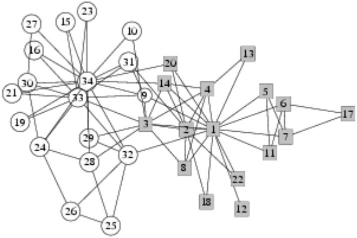
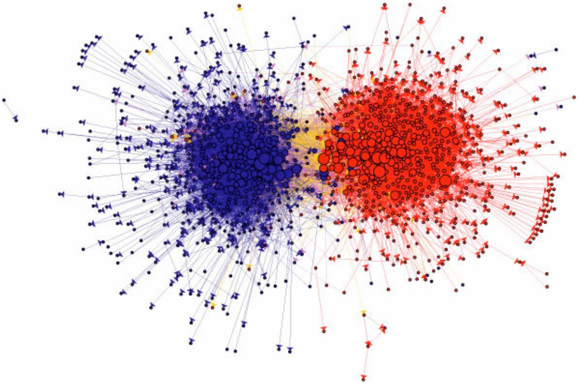


FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

# Links among Web pages between political blogs prior to the 2004 U.S. Presidential election reveals two natural and well-separated clusters. [1]



[1] Image from http://wwwpersonal. umich.edu/ ladamic/img/politicalblogs.jpg

# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

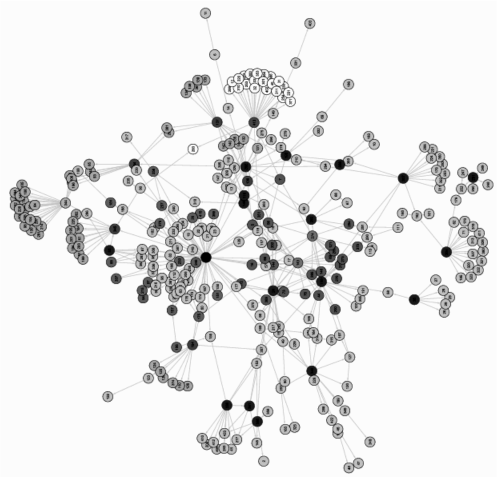
Methods based on a optimization criterion

Methods using a probabilistic model

Summary

Benchmark

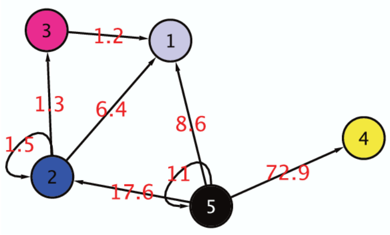
# Transcriptional regulatory network of E. Coli

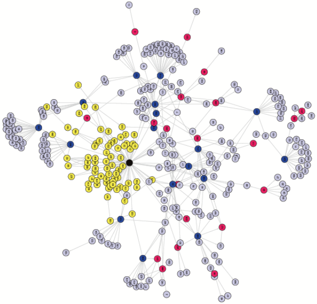


- nodes are operons
- edges between 2 operons if one regules the other
- known properties: sparseness, no feed-back circuits, hierarchical organization.

Data from Shen-Or et al. Nature genetics, 2002

# Mixnet results for TRN of E. Coli

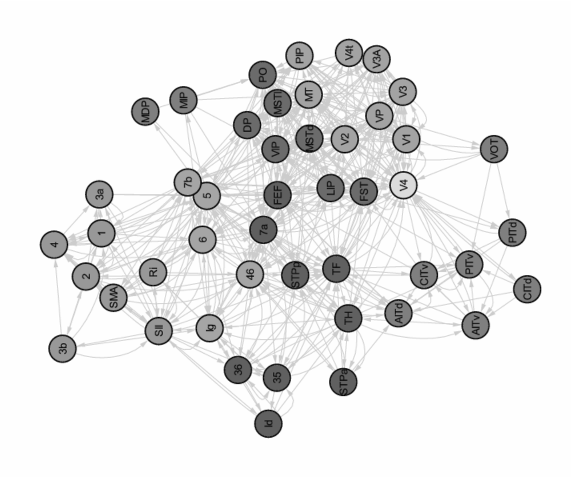




|  | MixNet Classes |  |  |  |  |
|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 |
| 1 | . | . | . | . | . |
| 2 | 6.40 | 1.50 | 1.34 | . | . |
| 3 | 1.21 | . | . | . | . |
| 4 | . | . | . | . | . |
| 5 | 8.64 | 17.65 | . | 72.87 | 11.01 |
| alpha | 65.49 | 5.18 | 7.92 | 21.10 | 0.30 |

Meta Hierarchical structure, Meta Single Input Modules and Feed Forward Loops.

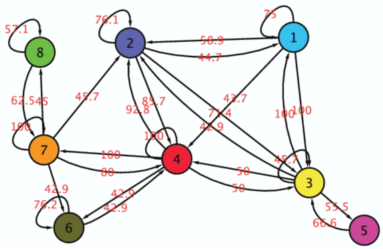
# Macaque Cortex Network



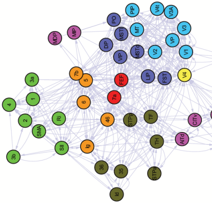
- nodes are cortical regions
- edges between 2 regions if one is connected to the other
- known properties: highly connected network, central and ”provincial hubs”.

Data from Sporns et al. PLoS one, 2007

# Mixnet results for Cortex network

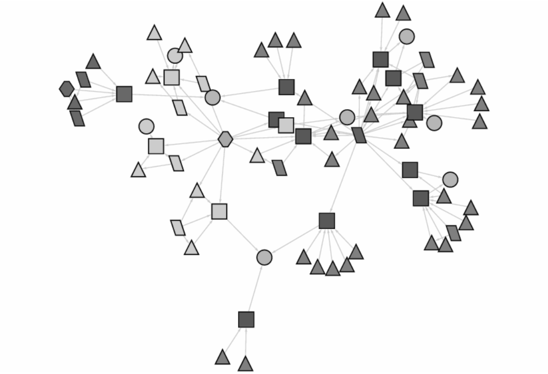


|  | MixNet Classes |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 75.0 | 58.9 | 100.0 | 43.7 | 2.8 | 3.6 | 10.0 | . |
| 2 | 44.7 | 76.1 | 71.4 | 85.7 | 3.2 | 12.2 | 25.7 | . |
| 3 | 100.0 | 42.9 | 45.7 | 50.0 | 55.5 | 28.6 | 20.0 | . |
| 4 | 6.2 | 92.8 | 50.0 | 100.0 | 11.1 | 42.9 | 100.0 | . |
| 5 | 4.2 | 6.4 | 66.6 | 27.8 | 23.6 | 4.8 | 4.4 | . |
| 6 | 8.9 | 12.2 | 28.6 | 42.9 | 12.7 | 76.2 | 31.4 | 1.8 |
| 7 | 15.0 | 45.7 | . | 80.0 | 6.7 | 42.9 | 100.0 | 45.0 |
| 8 | . | . | . | 18.7 | . | 7.1 | 62.5 | 57.1 |
| alpha | 17.0 | 14.9 | 2.1 | 4.3 | 19.2 | 14.9 | 10.6 | 17.0 |



Central and provincia hubs well identified.

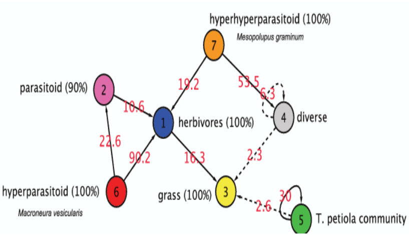
# Food-web network



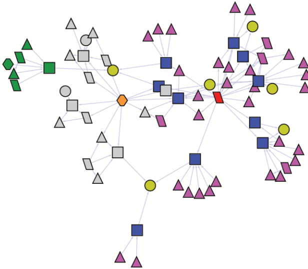
- the food web shows 5 levels of organization: plants (circle), herbivores (box), parasitoids (parallelogram), hyperparasitoids (triangle) and hyper-hyperparasitoids (diamond).
- a trophic link is considered between two insects when one insect is observed within one host
- known properties: hierarchic organization.

Data from Dawah et al. Journal of animal ecology, 1995, and Martinez et al. Ecology, 1999

# Mixnet results for Food-Web network



| | MixNet Classes | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | . | . | 16.3 | . | . | . | . |
| 2 | 10.6 | . | . | . | . | . | . |
| 3 | . | . | . | . | . | . | . |
| 4 | . | . | 2.3 | 6.3 | . | . | . |
| 5 | . | . | 2.6 | . | 30.0 | . | . |
| 6 | 90.2 | 22.6 | . | . | . | . | . |
| 7 | 19.2 | . | . | 53.5 | . | . | . |
| alpha | 14.0 | 44.4 | 8.6 | 22.3 | 8.0 | 1.3 | 1.3 |



The 5 levels are well identified plus a specific community. Local hierarchies are detected.

# Sommaire

Why ?

Exemples

Basic definitions

- Communities and SES
- Similarity transformation

Methods based on a algorithm

Methods based on a optimization criterion

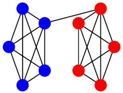
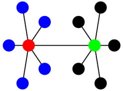
Methods using a probabilistic model

Summary

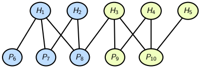
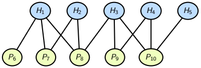
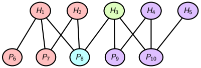
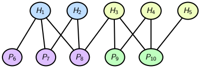
Benchmark

# Two definitions of what is a cluster in a graph

- Modularity or Communities : a cluster is composed of nodes highly connected to members of the same cluster and loosely connected to members of other clusters.
- Structural Equivalence of Actors defined by Lorrain and White : two actors are structurally equivalent if they have identical relational ties to and from all the actors in a network.

| | |
|---|---|
| 2 Communities or modules |  |
| 4 Structurally Equivalent Subsets |  |

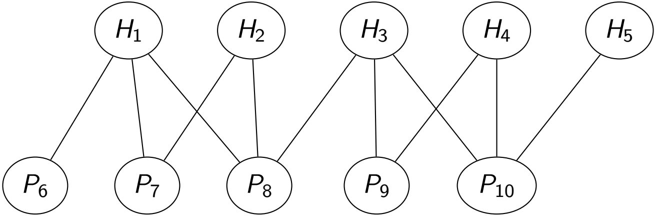
# Example

| | communities | structurally equivalent subsets |
|---|---|---|
| 2 clusters |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |
| 4 clusters |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |

# Basic notations

Let a graph :

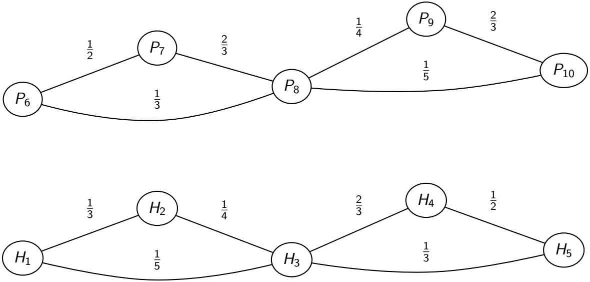
- $G = (V, E)$, $V$ the set of $n$ vertices (or nodes) and $E \subset V \times V$ the set of edges
- $W$ the adjacency matrix (weighted or not)
- $d_i^{(i)}$ and $d_i^{(o)}$ inner and outer degree of node $i$



# Similarity transformation on a graph

The Jaccard's similarity index ($J_{i,j} = \frac{\text{number of nodes connected to i and j}}{\text{number of nodes connected to i or j}}$):

$$S_J = \begin{pmatrix}
- & & & & & & & & & \\
\frac{2}{3} & - & & & & & \text{(sym)} & & & \\
\frac{1}{5} & \frac{1}{4} & - & & & & & & & \\
 & & \frac{2}{3} & - & & & & & & \\
 & & \frac{1}{3} & \frac{1}{2} & - & & & & & \\
 & & & & & - & & & & \\
 & & & & & \frac{1}{2} & - & & & \\
 & & & & & \frac{1}{3} & \frac{2}{3} & - & & \\
 & & & & & & & \frac{1}{4} & - & \\
 & & & & & & & \frac{1}{5} & \frac{2}{3} & -
\end{pmatrix}.$$



# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

- Markov Cluster algorithm (MCL)
- Pons-Latapy distance
- Spectral Clustering
- Edge-Betweeness
- Hierarchical agglomerative clustering algorithm

Methods based on a optimization criterion

Methods using a probabilistic model

Summary

Benchmark

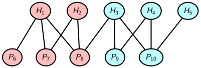
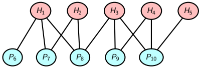
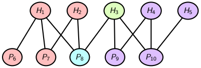
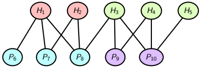
# Markov Cluster algorithm (MCL)

Random walk from nodes to nodes along edges. Probability of a move along an edge proportional to its weight. Transition matrix of the Markov chain: $T = (T_{ij})$, the probability of going from node $i$ to node $j$ in one step.
The MC is assumed to be ergodic (irreducible and aperiodic) → one final state. Several final states needed to obtain several clusters. Thus the MC is modified (by an inflation operation). MCL alternates two operations :

- $T^{(2k)} = (T^{(2k-1)})^e$, progress of the random walk.
- $T^{(2k+1)} = \Gamma_r(T^{(2k)})$, *inflation* operation. $\Gamma_r$ is a term by term $r$ power operator followed by a normalization.

$e$ and $r$ are tuning parameters. The algorithm ends when $T^{(k)}$ is idempotent. Two nodes are classified in the same class if they have the same final state.
*MCL need ergodicity of the Markov Chain, by example by adding self-loops*

# Markov Cluster algorithm (MCL)

| | hight weight on self-loops[a] | low weight on self-loops[b] |
|---|---|---|
| hight $e_k$, low $r_k$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |
| low $e_k$, hight $r_k$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |   $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |

[a] $W_{ii} = 1$, unitary self-loops

[b] $\frac{1}{10}$ weighted self-loops

# MCL

Tunning parameters :

- parameters of speed of Markov Chain in comparison of speed of inflation
- modification of graph (weight of self-loops for example)

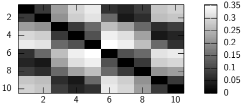
Properties

- MCL detects SES (in a modified graph with self-loops for example)
- Efficient for highly connected graphs, and less efficient for sparse graphs.
- Largely used by the Bioinformatics community...but rare in other scientific communities.

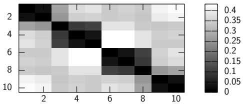
# Pons-Latapy distance

*Not a clustering method !*
Main idea:

- Random walk stopped at $t$ steps
- Distance between nodes = euclidian distance between rows of $T^t$.



$M_1$



$M_{\frac{1}{10}}$

Tunning parameter : $t$ and self-loops added is necessary.

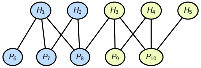
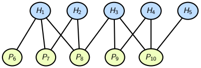
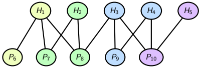
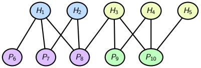
# Spectral Clustering

- Laplacian matrix of the graph $G : L = D_W - W$.
- $G$ has $k$ connected components $\Leftrightarrow$ $L$ has a zero-eigenvalue with multiplicity $k$.
- Each eigenvector is composed of zero and non-zero values (corresponding to the nodes of the connected component).
- $\rightarrow$ Spectral Clustering = k-means procedure in the space generated by the first-$k$ eigenvectors corresponding to the smallest eigenvalues.

Many variants :

- unnormalized Spectral Clustering : first $k$ eigenvectors of $L$ corresponding to $\lambda_1 \leq \lambda_2 \leq ... \lambda_k$.
- Shi-normalized Spectral Clustering : first $k$ eigenvectors of $D_W^{-1} L$, corresponding to $\lambda_1 \leq \lambda_2 \leq ... \lambda_k$.
- Ng-normalized Spectral Clustering : first $k$ eigenvectors of $L_N = I - D_W^{-1/2} W D_W^{-1/2}$, corresponding to $\lambda_1 \leq \lambda_2 \leq ... \lambda_k$.
- Absolute Eigenvalues Spectral Clustering : first $k$ eigenvectors of $I - L_N$, corresponding to $|\lambda_1| \geq |\lambda_2| \geq ... |\lambda_k|$.

# Spectral Clustering

| | Ng-normalized | Absolute Eigenvalues |
|---|---|---|
| 2 clusters |  $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |  $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |
| 4 clusters |  $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |  $H_1$<br>$H_2$<br>$H_3$<br>$H_4$<br>$H_5$<br>$P_6$<br>$P_7$<br>$P_8$<br>$P_9$<br>$P_{10}$ |

# Spectral clustering

Features :

- undirected graphs only
- Absolute Eigenvalues Spectral Clustering is the only SC method that detects SES.

Tunning parameters :

- number of clusters
- variant

# Edge-Betweeness

- Betweeness for a given edge = number of shortest paths using this edge
- quantify the importance of a link to maintain the graph connected
- Link between communities have a higher betweeness than links inside communities.

A divisive algorithm :

- Compute edge-betweeness and cut links with a decreasing betweeness order while the graph is connected
- Apply the algorithm on each connected component

The result is a hierarchical tree of sets.

Features : Detect communities.

Tunning parameters : The number of clusters *ie* the depth of the hierarchical tree.

# Hierarchical agglomerative clustering algorithm

Starting with single node cluster, this is a recursive algorithm :

- Find nearest couple of sets
- Merge couple of sets, compute distances, and apply recursively the algorithm

Feature : Detect communities

Tunning parameters : The number of cluster, and the method to computes distance of merged sets to others.

# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

Methods based on a optimization criterion

- Modularity criterion
- Cut cost

Methods using a probabilistic model

Summary

Benchmark

# Modularity criterion

Modularity of a partition $C$ : $\mathcal{M}_C = \sum_q (e_{qq} - a_q^2)$
($\simeq 0$ if no modularity, $\simeq 1$ if $Q$ unconnected cliques)

- $e_{ql} = \frac{1}{2m} \sum_{ij} W_{ij} \delta_q(i) \delta_l(j)$, proportion of edges between class $q$ and $l$,
- $m =$ total number of edges
- $\delta_q(i)$ is equal to one if $i$ is in the class $q$ and zero if not
- $a_q = \sum_l e_{ql}$ proportion of edges concerning a node of class $q$.

Guimera : Optimization by a Simulated Annealing (SA), with levels of temperature decreasing exponentially. Three moves possible :

- individual move of a node from a class to another
- merge two classes
- split a class into two classes, (SA inside SA)

# Modularity criterion

Features : Find communities

Optimization parameters : Decreasing speed of temperature of the 2 simulated annealing.

*High computation cost*

# Cut cost

- Suppress some edges from $G$ to obtain an unconnected partition of vertices with a minimum modification cost.
- cut cost between two subset of nodes :
  $\mathrm{cut}(V_1, V_2) = \sum_{v_1 \in V_1, v_2 \in V_2} W_{v_1, v_2}$
- cut cost of one partition :
  $\mathrm{cut}(C) = \sum_{q<l} \mathrm{cut}(C_q, C_l) = \frac{1}{2} \sum_{q=1}^{Q} \mathrm{cut}(C_q, V \setminus C_q)$.
- other definitions of the cost are possible...

Obtaining the best Cut partition is NP-hard.
Algorithms:

- heuristics
- greedy algorithms
- simulated annealing

# Sommaire

- Why ?
- Exemples
- Basic definitions
- Methods based on a algorithm
- Methods based on a optimization criterion
- Methods using a probabilistic model
    - SBM
        - Karate Club
        - PPI human cells
    - Continuous Stochastic bloc model (CSBM)
- Summary
- Benchmark

# Stochastic Block Model

Classes of nodes $(\mathcal{C}_q), q = 1, Q$,

Model:

$$P(W_{ij} = 1 \| i \in \mathcal{C}_q, j \in \mathcal{C}_l) = \pi_{ql}$$

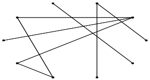
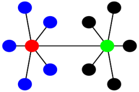
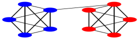
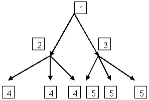
and $P(i \in \mathcal{C}_q) = \alpha_q$

Consistent estimation procedures (ML impossible, MCMC for small graphs $<$ 200 nodes, moments method (only theoretical results), variational method (best method for 200 to 5000 nodes) or based on the degrees for very large graphs: find

- number of classes $Q$ (BIC, AIC, ICL...)
- parameters estimates for $\alpha$ and $\pi$.

Clusters are obtained as a sub-product of the parameters estimation.

# SBM, a versatile model

| Description | Graph | $Q$ | $\pi$ |
|---|---|---|---|
| Erdos |  | 1 | $p$ |
| Hubs |  | 4 | $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ |
| communities |  | 2 | $\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$ |
| Hierarchical |  | 5 | $\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ |

# Generalized SBM

weighted graph

$$(W_{ij} \| i \in \mathcal{C}_q, j \in \mathcal{C}_l) \sim \mathcal{L}_{ql}$$

$\mathcal{L}_{ql}$: Poisson or Normal distributions.

covariates Information available for edges (or nodes) may be used in the model:

$$(W_{ij} \| i \in \mathcal{C}_q, j \in \mathcal{C}_l) \sim \mathcal{P}\left(\lambda_{ql} \exp(\beta^T Y_{ij}\right)$$

with

$Y_{ij}$ vector of covariates for the link $i \leftrightarrow j$.

*Clusters interpretation change with covariables*

- Parameter estimation: Variational algorithm
- Packages Mixer (Bernoulli, R), mixnet (Bernoulli, C), Wmixnet(GSBM, C).

# Karate Club

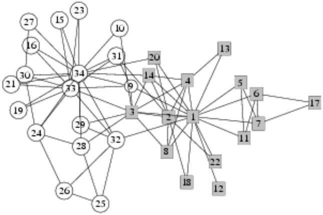


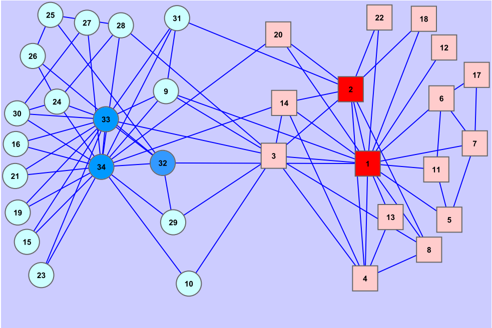
FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

- nodes: members of the club
- edge between two members if they have a social relation external to the club
- after the data collection, a split divided the club in two parts (circles and squares).

Data from W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).

# SBM results

|  | SBM Classes |  |  |  |
|---|---|---|---|---|
| $\widehat{\Pi}$ | 1 | 2 | 3 | 4 |
| 1 | 100 | 53 | 16 | 16 |
| 2 | 53 | 12 | 0 | 7 |
| 3 | 16 | 0 | 8 | 73 |
| 4 | 16 | 7 | 73 | 100 |
| $n \times \widehat{\alpha}$ | 3 | 13 | 16 | 2 |



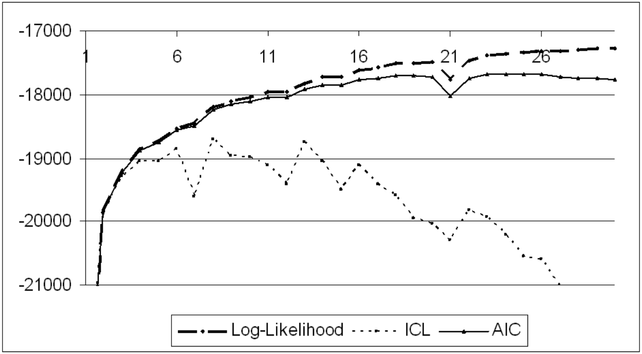
The split is exactly predicted and the leaders role is underlined.

# MS-Interactome data

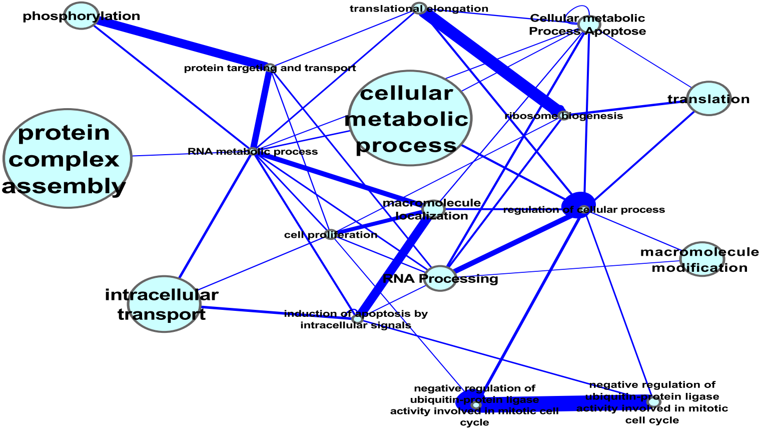
- MS-Interactome (Ewing et al.): first large-scale study of protein-protein interactions in human cells using a mass spectrometry approach.
- 3,494 interactions between 1,561 proteins
- Bait proteins chosen based on known functional annotation and implied disease association.
- One third of the 338 bait proteins are disease-related ones, mainly involved in cancer
- Data previously analyzed by Marras et al. using a two-steps procedure: first a deterministic method allows to find large core and community structures and second a stochastic method (such as mixture model) permits to uncover fine-grained interactome components.
- The following analysis is made using VEM method using package Mixnet.

# Number of groups

Best choices: $Q = 23$ (AIC) and $Q = 8$ (ICL).
x-axis : number of groups.



# Meta-Network obtained with SBM



phosphorylation
translational elongation
Cellular metabolic Process Apoptose
protein targeting and transport
cellular metabolic process
translation
ribosome biogenesis
protein complex assembly
RNA metabolic process
macromolecule localization
regulation of cellular process
cell proliferation
macromolecule modification
RNA Processing
intracellular transport
induction of apoptosis by intracellular signals
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle

# GO-Characteristics of the groups

Description of the first groups. The proteins have been affected to one group if their probability of pertaining to the group is greater than 0.5.

| group | # proteins | # unrecognized proteins | GO Term | Corrected P-Value |
|---|---|---|---|---|
| 1 | 44 | 2 | Cellular metabolic Process & Apoptose | $4.10^{-7}$ |
| 2 | 79 | 11 | RNA Processing | $5.10^{-3}$ |
| 3 | 12 | | cell proliferation | $8.10^{-3}$ |
| 4 | 211 | 24 | intracellular transport | $9.10^{-8}$ |
| 5 | 55 | 11 | macromolecule localization | $1.10^{-4}$ |
| 6 | 4 | | protein targeting and transport | $1.10^{-6}$ |
| 7 | 353 | 57 | Cellular metabolic Process | $5.10^{-12}$ |
| 8 | 111 | 12 | macromolecule modification | $3.10^{-16}$ |
| 9 | 372 | 73 | protein complex assembly | $3.10^{-8}$ |
| 10 | 96 | 14 | phosphorylation | $7.10^{-7}$ |
| 11 | 5 | 2 | negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle | $1.10^{-5}$ |
| 12 | 15 | | negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle | $2.10^{-38}$ |
| 13 | 2 | | RNA metabolic process | $1.10^{-2}$ |
| 14 | 8 | 1 | induction of apoptosis by intracellular signals | $5.10^{-3}$ |
| 15 | 8 | 1 | ribosome biogenesis | $1.10^{-3}$ |
| 16 | 110 | 27 | translation | $4.10^{-25}$ |
| 17 | 2 | | regulation of cellular process | $8.10^{-2}$ |
| 18 | 19 | 1 | translational elongation | $4.10^{-38}$ |
| 19 | 55 | | | |

# More about the groups

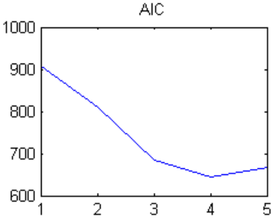
- most of the groups can be identified by at least one GO term with low corrected P-values
- 234 proteins were not recognized by *GO term Finder* → SBM proposes a classification for unknown proteins.
- 17$^{th}$ group composed of two proteins highly related with tumor progression: the Von Hippel Lindau (VHL) tumor suppression protein and MCC, which blocks cell cycle progression.
- group 13, composed of two proteins Tgfb1i4 (transforming growth factor beta 1 induced transcript), which is a growth factor, and RNSP1, which is a part of a post-splicing multiprotein complex regulating exons.

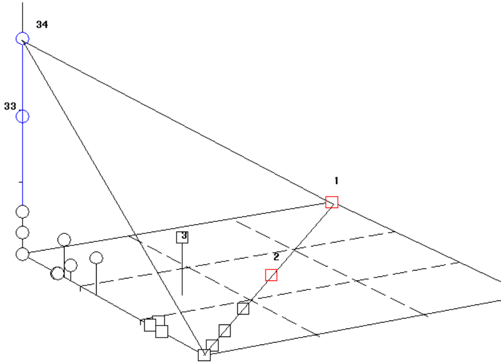
# Continuous Stochastic bloc model (CSBM)

- Each node $i$ : weighted mean of $Q$ Extremal Hypothetical Vertices (EHV)
- weight $Z_i = (z_{i1}, \ldots, z_{iQ})$, $z_{iq} \geq 0$, $\sum_q z_{iq} = 1$.
- $P_{ij} = \sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}$
- $a_{ql} \in [0,1]$: connectivity between EHV $q$ and $l$.
- $X_{ij} \sim B(P_{ij})$
- $X_{ij}$ independent

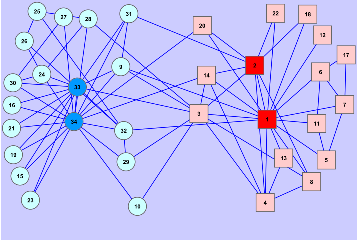
$$P = ZAZ'$$

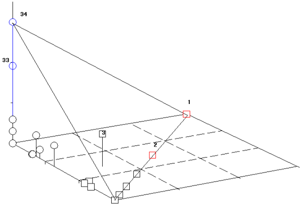
# CSBM results for the Karate Club





| $\widehat{A}$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 100 | 100 | 0 | 0 |
| 2 | 100 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 100 |
| 4 | 0 | 0 | 100 | 100 |



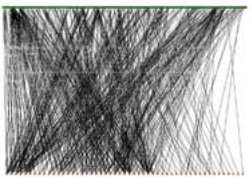


| $\widehat{A}$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 100 | 100 | 0 | 0 |
| 2 | 100 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 100 |
| 4 | 0 | 0 | 100 | 100 |

The split is well predicted, the role of the leaders and the intermediate position of node 3 are enlightened.

# Ecological network

Host-parasite interaction between a young pine tree and the fungi species Armillaria ostoyae (image from C. Vacher web site)



Interaction network between tree species and parasitic fungi species in the French forests (image from C. Vacher web site)

- 543 interactions between 51 forest tree taxa and 154 parasitic fungal species. The network is composed of 205 vertices and 543 edges.
- bipartite graph : tree-fungus interactions are the only possible ones.
- from the database of the French governmental organization in charge of forest health monitoring (the *Département Santé des Forêts (DSF)*) for the 1972-2005 period.
- methods used for data collection described in more detail in Vacher, C., Piou, D., and Desprez-Loustau, M.-L. (2008) *PLoS ONE* **3,** e1740.

# Results

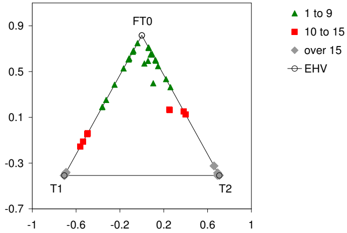
AIC criteria → $Q = 5$

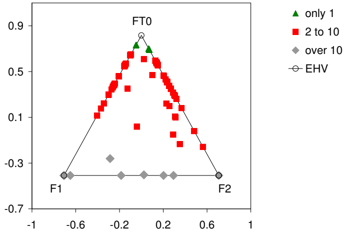
| | FT0 | T1 | T2 | F1 | F2 |
|---|---|---|---|---|---|
| FT0 | 0 | 0 | 0 | 0 | 0 |
| T1 | 0 | 0 | 0 | 0.996 | 0 |
| T2 | 0 | 0 | 0 | 0 | 0.985 |
| F1 | 0 | 0.996 | 0 | 0 | 0 |
| F2 | 0 | 0 | 0.985 | 0 | 0 |

EHV= Extremal Hypothetical Vertex

- EHV0=non connected species
- EHV1=T1 and EHV2=T2 two Extreme Hypothetical Trees
- EHV3=F1 and EHV4=F2 two Extreme Hypothetical Fungus
- the only two connected EHVs are T1 and F1 and T2 and F2

# Triangular representations of tree species and fungal species as a function of their number of interactions

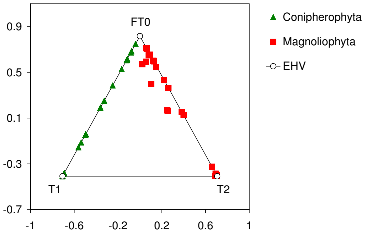


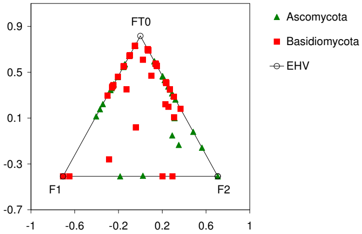


Vertical axis = degree of the vertices

Horizontal axis = differentiation between two classes of species

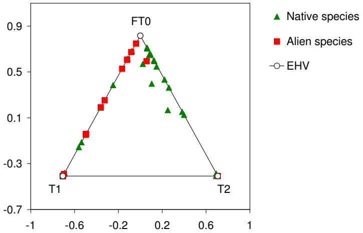
# Triangular representations of tree species and fungal species as a function of their phylogenetic origin

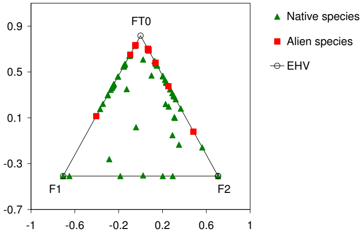




The differentiation along the horizontal axis is due to the phylogenetic origin of Trees. Phylogenetic origin of Fungi does not matter!

# Triangular representations of tree species and fungal species as a function of their introduction status





Aliens are rapidly (less than 600 years) integrated

# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

Methods based on a optimization criterion

Methods using a probabilistic model

Summary

Benchmark

# Summary

| Method | Type [a] | Directed [b] | Weighted [c] | Goal [d] | Tuning parameters |
|---|---|---|---|---|---|
| E.Between. | A | N | N | C | none |
| Cut | O | N | Y | C | Criteria |
| Modularity | O | Y | Y | C | none |
| Spec. Clust. | A | N | Y | C or SHS | method, $k$[e] |
| Hier. Clust. | A | N | Y | C or SHS | method |
| MCL | A | Y | Y | SHS | $r$[f], $e$[6], $\Delta$[g] |
| Pons-Latapy | A | N | Y | SHS | $k$[5], $\Delta$[7] |
| SBM | M | Y | Y | SHS | $k$[5] or none |
| CSBM | M | Y | N | SHS | $k$[5] or none |
| MBCSN[h] | M | N | N | C | $d$[i] and $k$[5] |
| RDPG | M | N | N | C | $d$[9] |

[a] A for algorithm, O for optimization, M for probabilistic model

[b] Y if the method can be applied to a directed graph, N otherwise

[c] Y if the method can be applied to a weighted graph, N otherwise

[d] C for Community research algorithm, SHS for Structural homogeneous subset research algorithm

[e] $k$ is the number of groups

[f] $e$ and $r$ are the importance of transition and inflation step, $\frac{e}{r}$ control the number of groups

[g] Weight of self-loops added for ergodicity

[h] Model-based clustering for social network

[i] $d$ is the dimension of the latent space

# Sommaire

Why ?

Exemples

Basic definitions

Methods based on a algorithm

Methods based on a optimization criterion

Methods using a probabilistic model

Summary

Benchmark

# Simulation design

The simulation model comes from Ecology. It is not a SBM or a CSBM model.

- number of nodes (70-350)
- number of groups (2-10)
- connectedness index (0.3-0.9)
- compartmentalization index (0-0.95)
- nestedness index (0-0.95)

# The corrected rand index

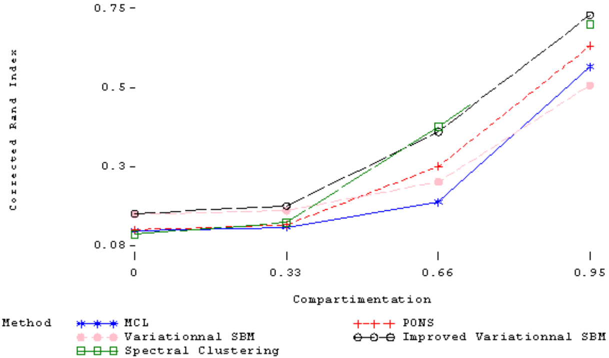
Rand index (in $[0,1]$) between two partitions:

$$R = \frac{\text{number of concordant pairs of nodes}}{\text{number of pairs of nodes}}$$

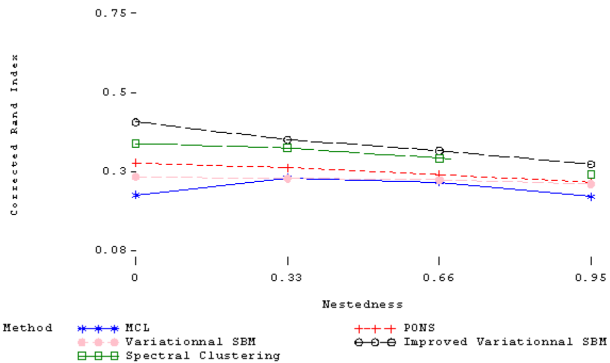
A pair of nodes is "concordant" if the two partitions classify the two nodes in the same way (in a same class or in two different classes).

The Rand index is corrected to have a zero mean when computed between any partition and a random one.

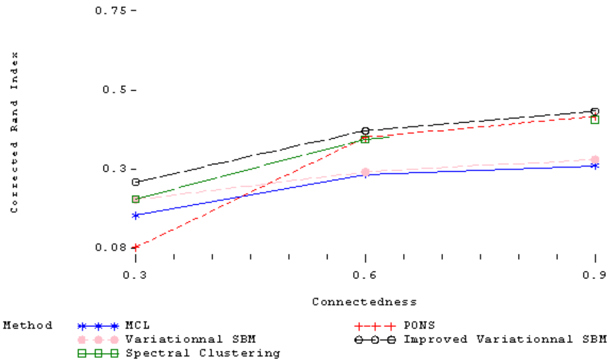
# Effect of Modularity



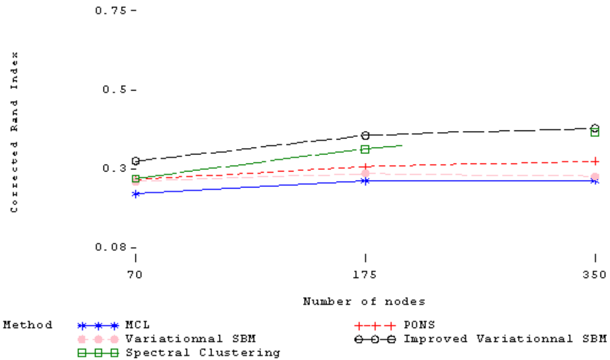
# Effect of Nestedness



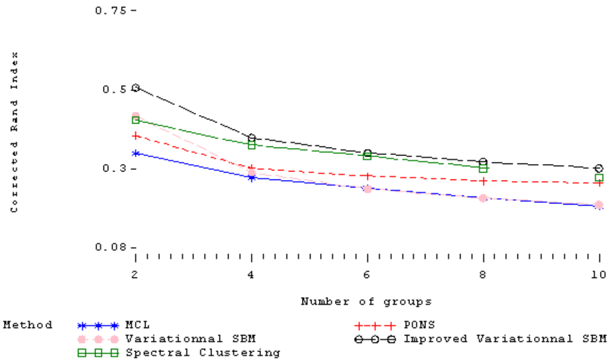
# Effect of Connectedness



# Effect of Number of nodes



# Effect of Number of groups


Corrected Rand Index
Number of groups
Method
MCL
PONS
Variationnal SBM
Improved Variationnal SBM
Spectral Clustering