# Sampling impact on inference structures in networks.

## Application to protein-protein network

Pierre Barbillon, Julien Chiquet, Timothée Tabouy

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay

**NETBIO**

November 9-10, 2017

# Plan

1. Random Graphs & Stochastic Block Model

2. Sampling strategies

3. About ESR1 gene

# Random Graphs

## Generality

- A Graph : $G = (\text{Summits}, \text{Edges})$
- Adjacency matrix : $X = (X_{ij})$
- Models a relationship : $X_{ij}$ lives in $\{0, 1\}$

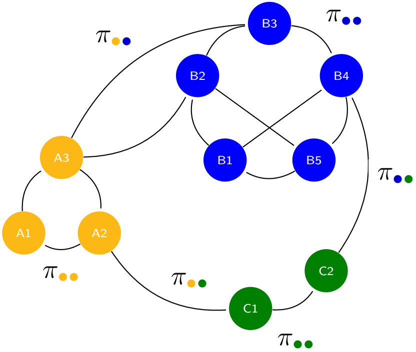
## -> Erdős-Rényi (1959)

## In practice

- Heterogeneity of connections
- Different connectivity profiles

$\rightsquigarrow$ How to model the heterogeneity of the network ?

# Stochastic Block Model




## Stochastic Block Model

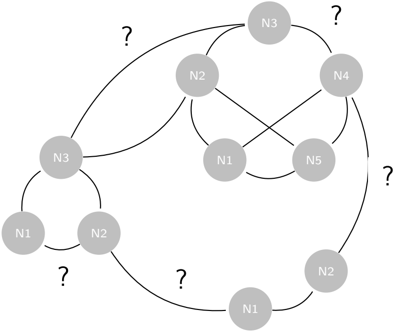
Let $n$ nodes divided into

- $\mathcal{Q} = \{\bullet, \bullet, \bullet\}$ classes
- $\alpha_\bullet = \mathbb{P}(i \in \bullet)$, $\bullet \in \mathcal{Q}, i = 1, \ldots, n$
- $\pi_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \alpha), \quad \forall \bullet \in \mathcal{Q},$$

$$X_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}(\pi_{\bullet\bullet})$$

# Statistical inference




## Stochastic Block Model

Let *n* nodes divided into

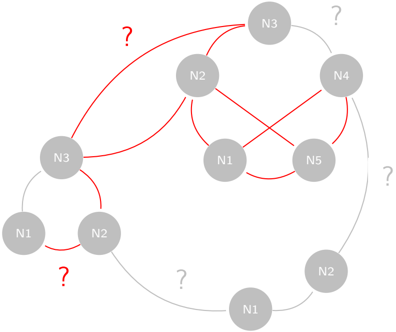
- $\mathcal{Q} = \{\bullet, \bullet, \bullet\}$, card($\mathcal{Q}$) known
- $\alpha_{\bullet} =?$,
- $\pi_{\bullet\bullet} =?$


Nowicki, Snijders, JASA, 2001
Estimation and prediction for stochastic blockstructures.

Daudin, Picard, Robin, Statistics and Computing, 2008
A mixture model for random graphs.

# Inference in the presence of missing data

## Central problem of the thesis




### Stochastic Block Model

*Edges are observed (or not) according to a specific sampling process must be taken into account in the inference*

- Purely random process ?
- Depending on the connectivity ?
- Depending hidden colors ?


Handcock, Gile, The Annals of Applied Statistics, 2010
Modeling social networks from sampled data

Kolaczyk, Springer, 2009
Statistical analysis of network data

# Inference in the presence of missing data

## Sampling matrix

$$(R_{ij}) = \begin{cases} 1 & \text{if } X_{ij} \text{ is observed,} \\ 0 & \text{if } X_{ij} \text{ is not observed.} \end{cases}$$

## Missing at Random (Rubin, 1976)

When the sampling process satisfies the following equation :

$$p(R|X) = p(R|X_{obs}),$$

the inference is done only on the likelihood of the observed data.

# Dyad-centered sampling strategies

1. Random dyad sampling

$$\forall (i,j) \in \{1,\ldots,n\}^2, \quad \mathbb{P}(R_{ij}=1)=\rho.$$

2. Double standard sampling

$$\begin{cases} \mathbb{P}(R_{ij}=1|X_{ij}=1)=\rho_1, \\ \mathbb{P}(R_{ij}=1|X_{ij}=0)=\rho_0. \end{cases}$$

# Node-centered sampling strategies

1. Star sampling

$$\forall i \in \{1, \dots, n\}, \quad \mathbb{P}(S_i = 1) = \rho.$$

2. Star degree sampling

$$\mathbb{P}(S_i = 1 | D_i) = \mathbb{P}(Z \leqslant a + bD_i) \text{ where } D_i = \sum_j X_{ij}.$$

3. Class sampling

$$\mathbb{P}(S_i = 1 | Z_i) = \rho_{Z_i}$$

Where $S_i = \mathbb{1}_{\text{node } i \text{ is sampled}}$

# Recall on binary SBM inference

## Remarks / questions

- The SBM is a latent variable model
- Can you use an EM algorithm ?
  - $\hookrightarrow$ Answer : No

## Complete log-likelihood & Variational hypothesis

$$\log(p_\theta(X,Z)) = \sum_{1\leqslant i<j\leqslant n} \sum_{q,l=1}^{Q} Z_{iq}Z_{jl}\log\{b(X_{ij},\pi_{ql})\} + \sum_{i=1}^{n}\sum_{q=1}^{Q} Z_{iq}\log(\alpha_q)$$

$\rightarrow$ **E Step** is not tractable : $\mathbb{E}_\theta[Z_{iq}Z_{jl}|X]$.

$\rightarrow$ the $Z_i$ are assumed independent (mean-field approximation).

# Results on MAR's inference

## Connectivity matrix inference

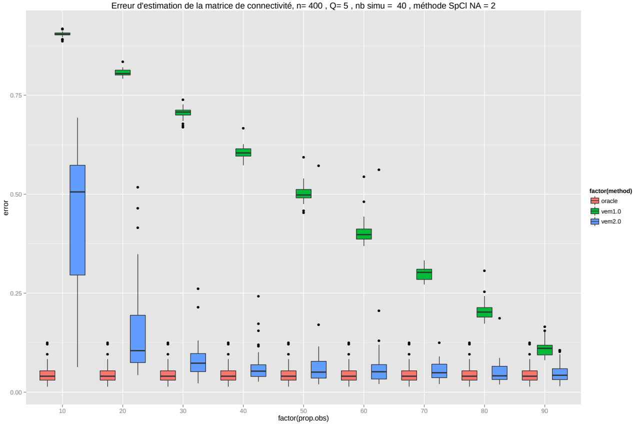



Figure – n=400, Q=5.

# Results on MAR's inference

## Adjusted Rand Index

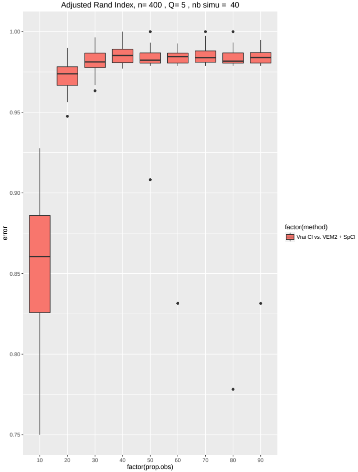



Figure – n=400, Q=5.

# Results on MAR's inference

## Inference of the number of classes

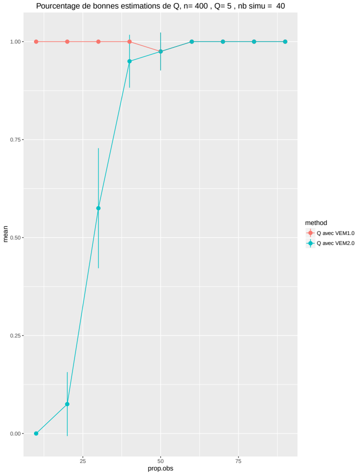



Figure – n=400, Q=5.

# Data analysis

## About ESR1 and the network selected :

Estrogen receptor 1 (ESR1) is a gene that encodes an estrogen receptor protein (ER), a central actor in breast cancer.

→ **Where** : The platform 483 string (Szklarczyk et al., 2015) accessible via http ://www.string-db.org
→ **What** : Protein-Protein Interaction (PPI)
→ **Links** : The value of an edge in this network corresponds to 487 a score obtained by aggregating different types of knowledge (wet-lab experiments, 488 textmining, co-expression data, etc. . . ), reflecting a level of confidence
→ **Number of nodes (neighbour)** : 741 proteins

# Data analysis and inference

## Adjacency Matrix :

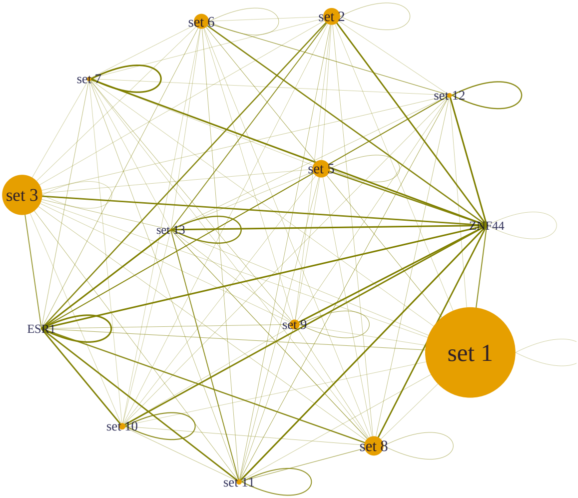
$$\mathbf{A}^{\gamma} = (A^{\gamma})_{ij} = \begin{cases} 1 & \text{if } \omega_{ij} > 1-\gamma, \\ \mathtt{NA} & \text{if } \gamma \leq \omega_{ij} \leq 1-\gamma, \\ 0 & \text{if } \omega_{ij} < \gamma. \end{cases} \quad (1)$$

$\rightarrow$ **Threshold** : $\gamma = 0.25$
$\rightarrow$ **Dyads** : 2546 dyads equal to 1, 264073 dyads equal to 0 and 7551 missing dyads.

## Inference with missing data

- Sampling selected : double standard sampling
- Number of clusters : 15

# Clusters Network



# Connectivity matrix

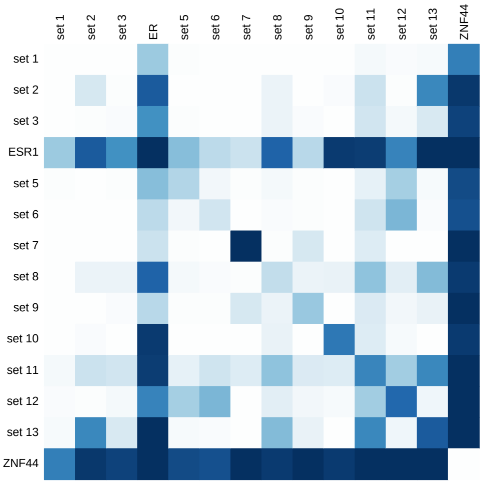



Figure – Matrix of connectivity $\hat{\pi}$ for NMAR inference (double standard) ; intensity of the color is proportional to the probability of connection between blocks.

# Gene Ontology

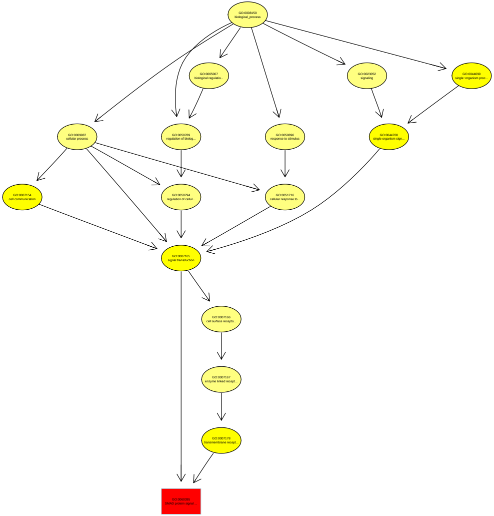



Figure – DAG of ontologies to which genes are annotated if the proteins encoded by these genes have been shown to be involved in a biological process

Thank you for your attention !