missSBM

Inference in Stochastic Block Models from Missing Data

P. Barbillon, J. Chiquet, T. Tabouy

Paris-Saclay, AgroParisTech, INRAE

Last update 09 December, 2021

https://github.com/GrossSBM/missSBM

Resources

R/C++ package

Last stable release on CRAN, development version available on GitHub.

```
install.packages("missSBM")
remotes::install_github("GrossSBM/missSBM@development")
```

```
library(missSBM)
packageVersion("missSBM")
```

```
## [1] '1.0.1'
```

Publications

The missSBM website contains the standard package documentation and a couple of vignettes for the top-level functions.

- Tabouy, T., P. Barbillon, and J. Chiquet (2019). "Variational Inference for Stochastic Block Models from Sampled Data". In: *Journal of the American Statistical Association* 0.ja, pp. 1-20. DOI: 10.1080/01621459.2018.1562934.
- Barbillon, P., J. Chiquet, and T. Tabouy (2022). "misssbm: An r package for handling missing values in the stochastic block model". In: *Journal of Statistical Software*.

Outline

Network data with missing entries

Recommandation system: Epinion

Who-trust-whom online social network of a general consumer review site Epinions.com. Members of the site can decide whether to "trust" each other. All the trust relationships interact and form the Web of Trust which is then combined with review ratings to determine which reviews are shown to the user.

Available at http://www.trustlet.org/datasets/extended_epinions/user_rating.txt.gz

Social networks in ethnobiology: seed exchange network

A limited space area was defined where all the 155 farmers were interviewed. Collected seed exchange between 568 farmers. They belong to different ethnies and speak several dialects

Ecological networks: plant-pollinator nbetwork

Interaction network between plants and pollonitor: how can trust the "0" in network data collected? Rather missing data ?

Companion data set: French political Blogosphere

Single day snapshot of almost 200 political blogs automatically extracted the 14 October 2006 and manually classified by the "Observatoire Présidentielle" project.

```
data("frenchblog2007")
party <- vertex.attributes(frenchblog2007)$party
table(party) %>% kableExtra::kbl() %>% kableExtra::kable_classic()
```

party	Freq
analyst	11
center-left	11
center-rigth	32
far-left	7
far-right	2
green	9
left	57
liberal	25

French blog: graph view

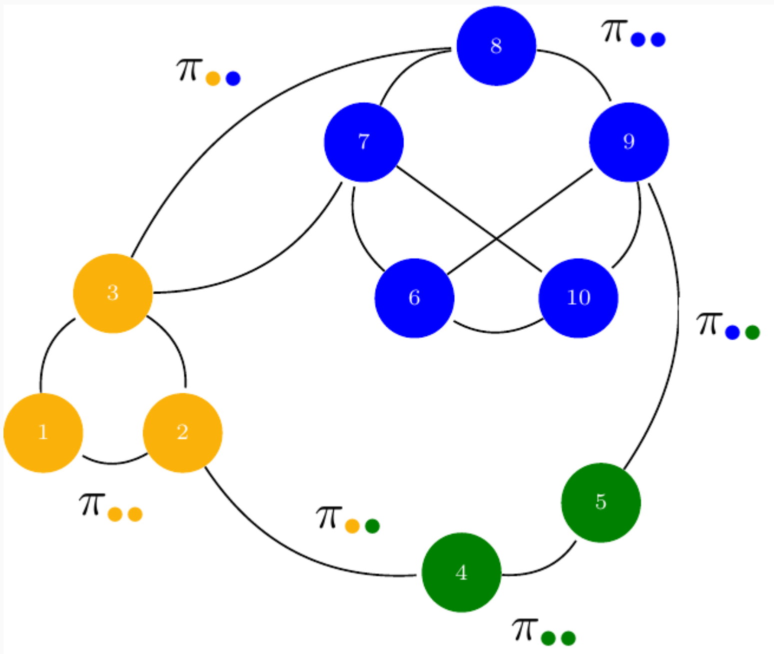
French blog: matrix view

SBM: background

- Probabilistic model for random graph
- Latent variable model
- Variational Inference

Stochastic Block Model

A popular probabilistic model for network data



The binary SBM model

Let

- *Fixed* nodes $\{1, \dots, n\}$
- Unknown colors in $\mathcal{C} = \{{\color{orange}\bullet}, {\color{blue}\bullet}, {\color{green}\bullet}\}$
- $\alpha_\bullet = \mathbb{P}(i \in \bullet)$, $\bullet \in \mathcal{C}$
- $\pi_{{\color{orange}\bullet}{\color{blue}\bullet}} = \mathbb{P}(i \leftrightarrow j | i \in {\color{orange}\bullet}, j \in {\color{blue}\bullet})$

In other words,

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \alpha),$$
$$Y_{ij} \mid \{i \in {\color{orange}\bullet}, j \in {\color{blue}\bullet}\} \sim^{\text{ind}} \mathcal{B}(\pi_{{\color{orange}\bullet}{\color{blue}\bullet}})$$

- Frank, O. and F. Harary (1982). "Cluster inference by using transitivity indices in empirical graphs". In: *J. Am. Stat. Soc.* 77.380, pp. 835–840.
- Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). "Stochastic blockmodels: First steps". In: *Social networks* 5.2, pp. 109–137.

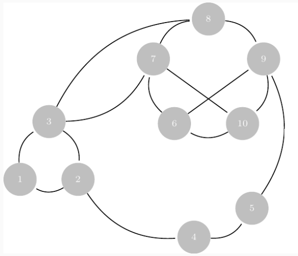
Examples of topology: Community network


```
pi ← matrix(c(0.3,0.02,0.02,0.02,0.3,0.02,0.02,0.02,0.3),3,3)
communities ← igraph::sample_sbm(100, pi, c(25, 50, 25))
plot(communities, vertex.label=NA, vertex.color = rep(1:3,c(25, 50, 25)))
```

Examples of topology: Star network


```
pi <- matrix(c(0.05,0.3,0.3,0),2,2)
star <- igraph::sample_sbm(100, pi, c(4, 96))
plot(star, vertex.label=NA, vertex.color = rep(1:2,c(4,96)))
```

Estimation in the SBM: latent variable model



- *Fixed* nodes $\{1, \dots, n\}$
- latent colors $\mathcal{C} = \{{\color{orange}\bullet}, {\color{blue}\bullet}, {\color{green}\bullet}\}$

Estimate the model parameters and the clustering:

- $\theta = (\boldsymbol{\alpha} = \{\alpha_\bullet), \boldsymbol{\Pi} = (\pi_{{\color{orange}\bullet}{\color{blue}\bullet}})\}$
- Colors of i, i.e. the $\mathbf{Z}_i$

Marginal likelihood

Integration over $\mathcal{Z}$ is intractable: $\mathbf{card}(Q)^n$ terms!

$$p_\theta(\mathbf{Y}_i) = \int_{\mathcal{Z}} \prod_{(i,j)} p_\theta(Y_{ij} | Z_i, Z_j) \, p_\theta(\mathbf{Z}) \mathrm{d}\mathbf{Z}$$

Maximum likelihood for incomplete data model: EM

$$\log p_\theta(\mathbf{Y}) = \mathbb{E}_{p_\theta(\mathbf{Z} \mid \mathbf{Y})}[\log p_\theta(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[p_\theta(\mathbf{Z} \mid \mathbf{Y})], \qquad \text{with } \mathcal{H}(p) = -\mathbb{E}_p(\log(p))$$

EM requires to evaluate (some moments of) $p_\theta(\mathbf{Z} \mid \mathbf{Y})$

Intractable EM: solution(s)

Variants of EM, MCMC/Bayesian approaches

- Nowicki, K. and T. A. B. Snijders (2001). "Estimation and Prediction for Stochastic Blockstructures". In: *J. Am. Stat. Soc.* 96.455, pp. 1077-1087.
- Daudin, J., F. Picard, and S. Robin (2008). "A mixture model for random graphs". In: *Stat. comp.* 18.2, pp. 173-183.
- Latouche, P., É. Birmelé, and C. Ambroise (2012). "Variational Bayesian inference and complexity control for stochastic block models". In: *Stat. Modelling* 12.1, pp. 93-115.
- Peixoto, T. P. (2014). "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models". In: *Physical Review E* 89.1, p. 012804.

Variational approach

Find a proxy $q_\psi(\mathbf{Z}) \approx p_\theta(\mathbf{Z}|\mathbf{Y})$ picked in a convenient class of distribution $\mathcal{Q}$

$$q(\mathbf{Z})^\star \arg\min_{q\in\mathcal{Q}} D\left(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{Y})\right).$$

Küllback-Leibler is a popular choice (error averaged wrt the approximated distribution)

$$KL\left(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{Y})\right) = \mathbb{E}_q\left[\log\frac{q(z)}{p(z)}\right] = \int_{\mathcal{Z}} q(z)\log\frac{q(z)}{p(z)}\mathrm{d}z.$$

Variational EM for SBM

Class of distribution: multinomial

$$\mathcal{Q} = \left\{ q_\psi : \, q_\psi(\mathbf{Z}) = \prod_i q_{\psi_i}(\mathbf{Z}_i), \, q_{\psi_i}(\mathbf{Z}_i) = \mathcal{M}\left(\mathbf{Z}_i; \boldsymbol{\tau}_i\right), \, \psi_i = \{\boldsymbol{\tau}_i\}, \boldsymbol{\tau}_i \in \mathbb{R}^K \right\}$$

Maximize the ELBO (Evidence Lower BOund):

$$J(\theta, \psi) = \log p_\theta(\mathbf{Y}) - KL[q_\psi(\mathbf{Z}) || p_\theta(\mathbf{Z}|\mathbf{Y})] = \mathbb{E}_q[\log p_\theta(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[q_\psi(\mathbf{Z})]$$

Variational EM

- Initialization: get $\mathbf{T}^0 = \{\tau_{ik}^0\}$ with Absolute Spectral Clustering
- M step: update $\boldsymbol{\theta}^h = \{\boldsymbol{\alpha}^h, \boldsymbol{\Pi}^h\}$
- VE step: find the optimal q_ψ, by updating $\psi^h = (\psi_i^h)_i = \mathbf{T}^h = \mathbb{E}_{q^h}(\mathbf{Z})$:

$$\psi^h = \arg\max J(\theta^h, \psi) = \arg\min_\psi KL[q_\psi(\mathbf{Z}) \,||\, p_{\theta^h}(\mathbf{Z} \mid \mathbf{Y})]$$

$$\theta^h = \arg\max J(\theta, \psi^h) = \arg\max_\theta \mathbb{E}_{q_{\psi^h}}[\log p_\theta(\mathbf{Y}, \mathbf{Z})]$$

Variational EM for SBM: ingredients

Variational bound

$$J(\theta, \tau; \mathbf{Y}) = \sum_{(i,j)} \sum_{(k,\ell)} \tau_{ik}\tau_{j\ell} \log b(Y_{ij}, \pi_{k\ell}) + \sum_i \sum_k \tau_{ik} \log(\alpha_k / \tau_{ik})$$

M-step (Analytical)

$$\alpha_k = \frac{1}{n} \sum_i \tau_{ik}, \quad \pi_{k\ell} = \frac{\sum_{(i,j)} \tau_{ik}\tau_{j\ell} Y_{ij}}{\tau_{ik}\tau_{j\ell}} \qquad \left(\boldsymbol{\alpha} = \mathbf{1}_n^\top \mathbf{T}, \quad \boldsymbol{\Pi} = \frac{\mathbf{T}^\top \mathbf{Y} \mathbf{T}}{\mathbf{T}^\top \mathbf{T}} \right)$$

Variational E-step (fixed point)

$$\tau_{ik} \propto \alpha_k \prod_{(i,j)} \prod_{\ell} b(Y_{ij}; \pi_{k\ell})^{\tau_{j\ell}}$$

Model Selection

$$\mathrm{vICL}(K) = \mathbb{E}_q[\log L(\hat{\theta}); \mathbf{Y}, \mathbf{Z}] - \frac{1}{2} \left(\frac{K(K+1)}{2} \log \frac{n(n-1)}{2} + (K-1)\log(n) \right)$$

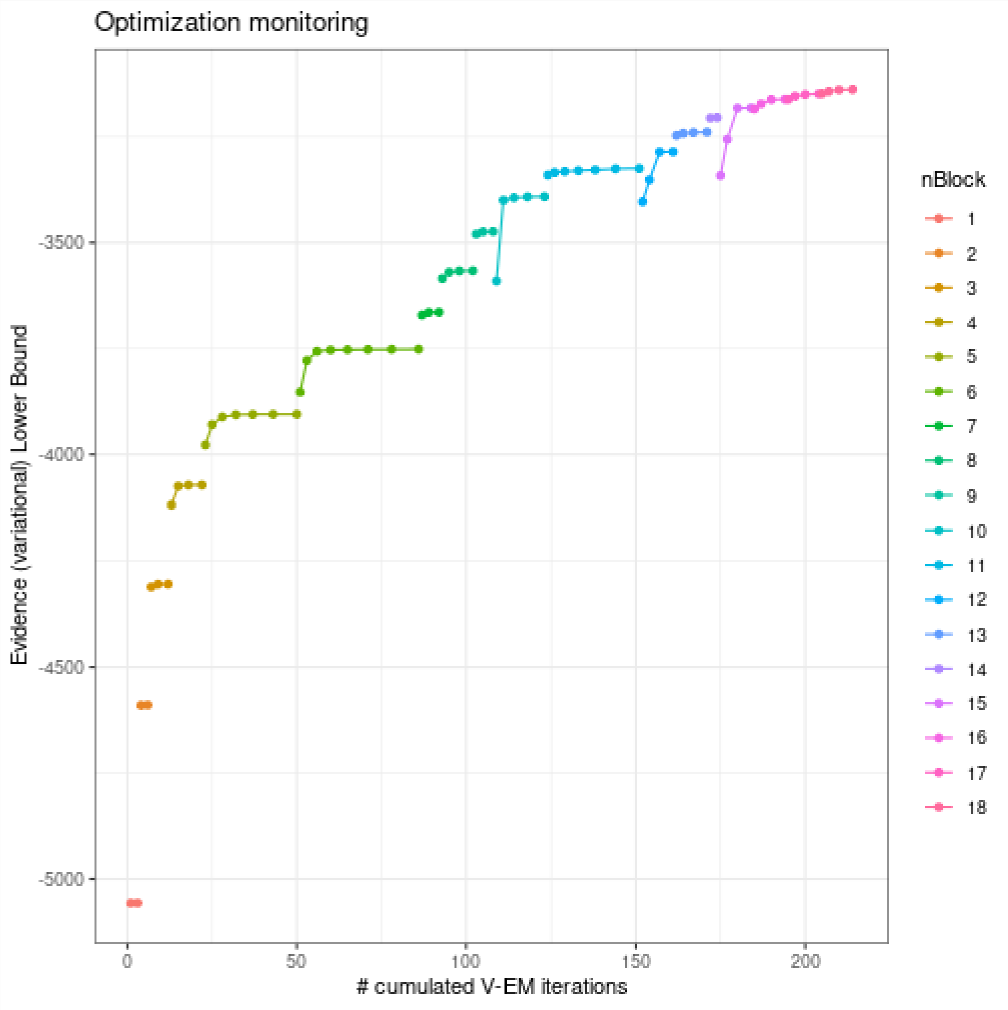
Example: French politcal blogosphere


```r
blog   <- as_adj(frenchblog2007, sparse = FALSE)
blocks <- 1:18
sbm_full <- estimateMissSBM(blog, blocks, "node")
```

```
##
##
##  Adjusting Variational EM for Stochastic Block Model
##
##     Imputation assumes a 'node' network-sampling process
##
##  Initialization of 18 model(s).
##  Performing VEM inference
##      Model with 6 blocks.
     Model with 8 blocks.
     Model with 3 blocks.
     Model with 13 blocks.
     Model with 2 blocks.
     Model with 10 blocks.
     Model with 15 blocks.
     Model with 16 blocks.
     Model with 7 blocks.
     Model with 12 blocks.
     Model with 4 blocks.
```

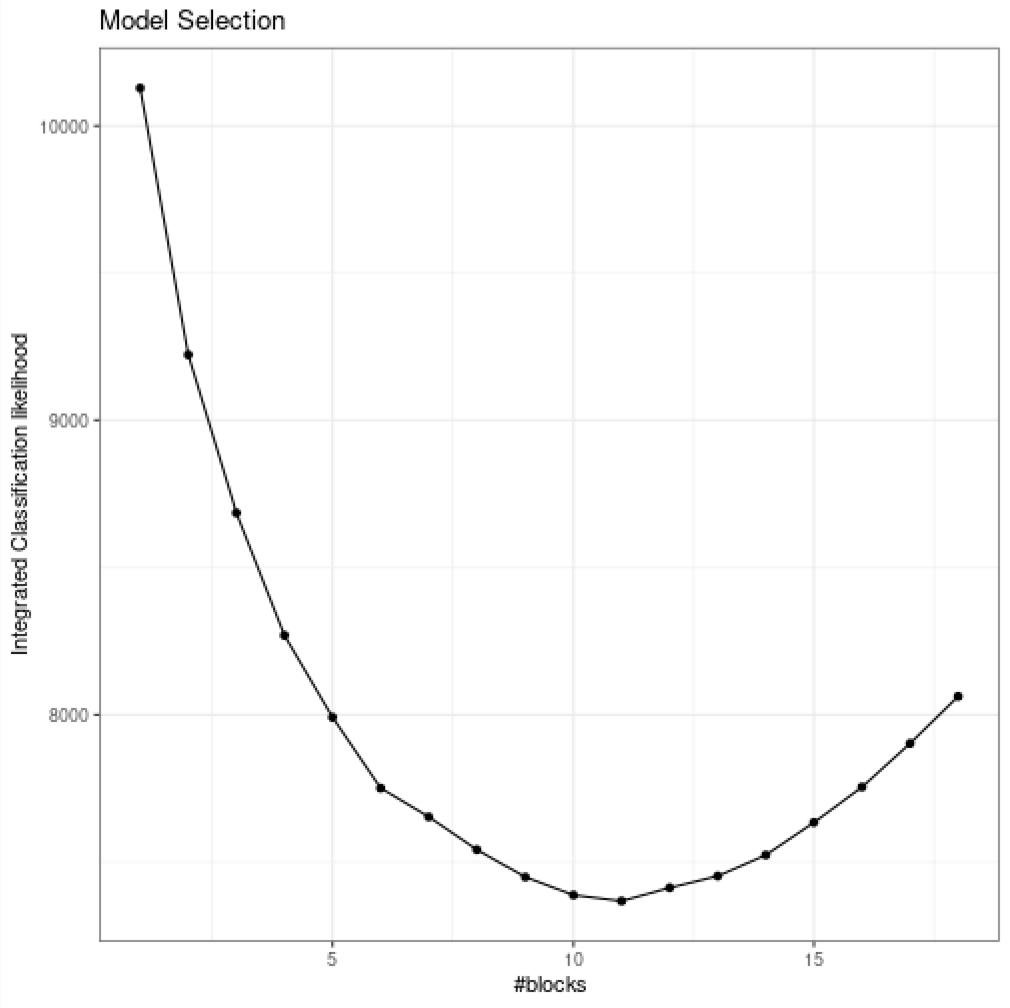
Convergence monitoring (ELBO)


```
plot(sbm_full, "monitoring")
```



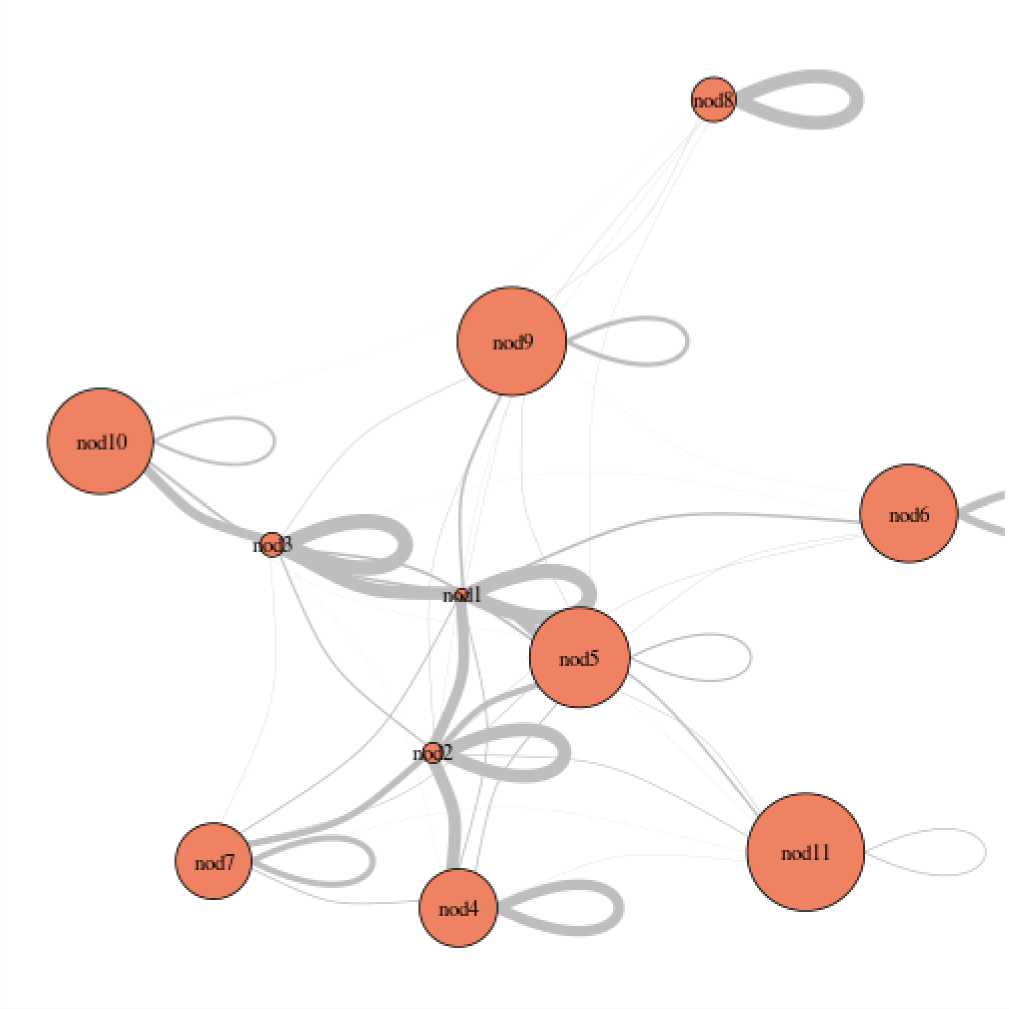
Model Selection (vICL)


```
plot(sbm_full)
```



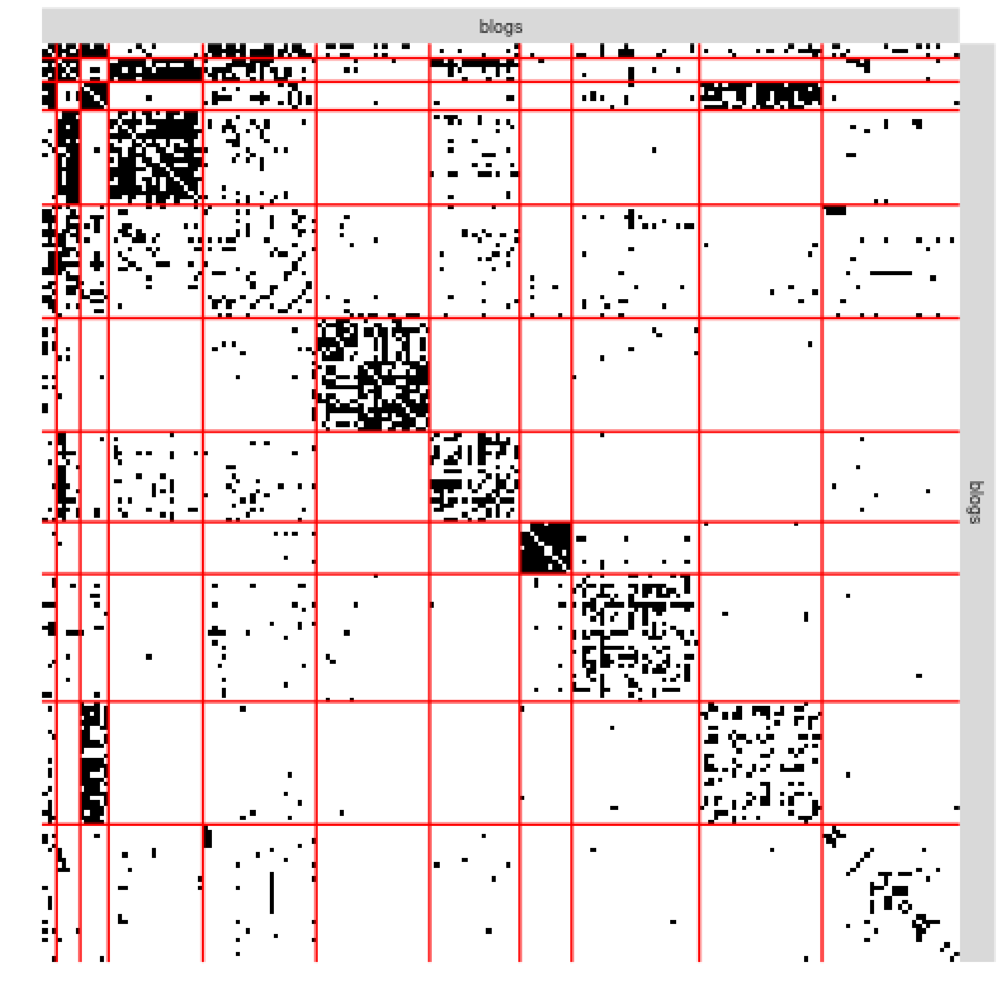
Parameters


```
plot(sbm_full$bestModel, "meso")
```



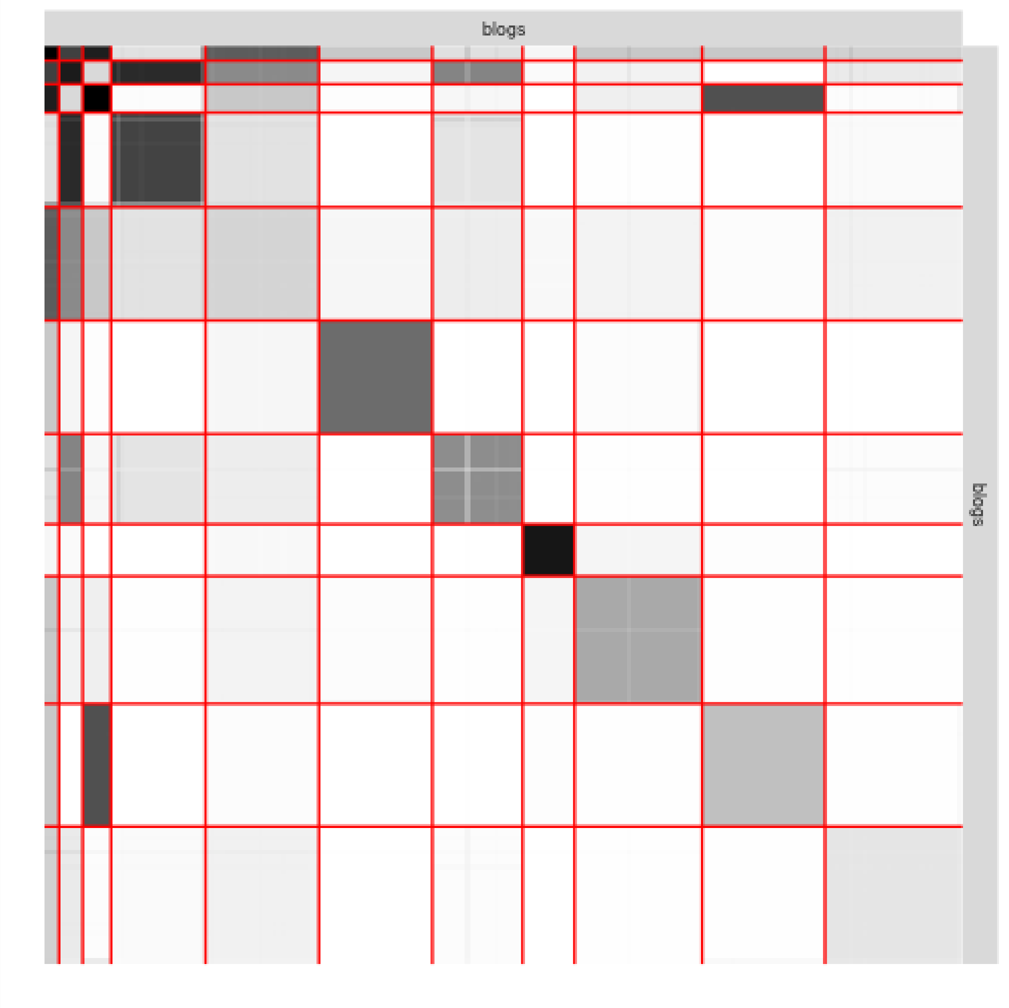
Clustering I


```
plot(sbm_full$bestModel, dimLabels = list(row = "blogs", col = "blogs"))
```



Clustering II


```
plot(sbm_full$bestModel, "expected", dimLabels = list(row = "blogs", col = "blogs"
```



Clustering III


```r
aricode::ARI(sbm_full$bestModel$fittedSBM$memberships, party)
```

```
## [1] 0.4517304
```

```r
aricode::NID(sbm_full$bestModel$fittedSBM$memberships, party)
```

```
## [1] 0.3905426
```

SBM from an observed network

- missing data framework for SBM
- Modeling the observation process
- Inference with missing dyads

Inference of an observed network (missing dyads)

$$
\begin{pmatrix}
 & 1 & \mathtt{NA} & 1 & 0 & \mathtt{NA} & 0 & 0 & 0 & 0 \\
1 & & 0 & 0 & 1 & 0 & 0 & 1 & \mathtt{NA} & 0 \\
\mathtt{NA} & 0 & & \mathtt{NA} & 0 & 0 & 1 & \mathtt{NA} & 1 & 0 \\
1 & 0 & \mathtt{NA} & & 0 & 0 & 0 & \mathtt{NA} & 1 & 0 \\
0 & 1 & 0 & 0 & & 1 & 0 & 0 & 0 & 0 \\
\mathtt{NA} & 0 & 0 & 0 & 1 & & 0 & \mathtt{NA} & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & & 0 & 0 & 0 \\
0 & 1 & \mathtt{NA} & \mathtt{NA} & 0 & \mathtt{NA} & 0 & & \mathtt{NA} & 0 \\
0 & \mathtt{NA} & 1 & 1 & 0 & 1 & 0 & \mathtt{NA} & & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 &
\end{pmatrix}
$$

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

About the sampling

- completely random?
- Depends on the connectivity?
- Depends on hidden colors (groups)?

- Kolaczyk, E. D. (2009). *Statistical analysis of network data, methods and models*. Springer.
- Handcock, M. S. and K. J. Gile (2010). "Modeling Social networks From Sampled Data". In: *The Annals of Applied Statistics* 4.1, pp. 5-25.
- Frisch, G., J. Léger, and Y. Grandvalet (2020). "Learning from missing data with the Latent Block Model". In: *arXiv preprint arXiv:2010.12222*.
- Gaucher, S., O. Klopp, and G. Robin (2021). "Outlier detection in networks with missing links". In: *Computational Statistics and Data Analysis* 164, p. 107308.

Missing data: general framework

Little and Rubin's framework

Let

- $R \sim p_\beta$ be a random process defining the observation (sampling) process
- $Y \sim p_\theta$ be some data split into two subsets Y^m, Y^o ("observed" and "missing")

Little and Rubin [LR14]' define

- **MCAR** (Missing Completely At Random): $R \perp Y$

- **MAR** (Missing At Random): : $R \perp Y^m | Y^o$

- **MNAR** (Missing Not At Random): other cases

Note that MCAR $\subset$ MAR and that in MAR case, inference of θ can be done of Y^o only:

$$
\begin{aligned}
p_{\theta,\beta}(Y^o, R) &= \int p_\theta(Y^o, Y^m) p_\psi(R|Y^o, Y^m) dY^m \\
&= p_\theta(Y^o) p_\beta(R|Y^o)
\end{aligned}
$$

Missing data: SBM case

Setting

- The observation process is given by the sampling matrix

$$(R_{ij}) = \mathbf{1}_{\{Y_{ij} \text{ is observed}\}}$$

- The process is **MAR** if $R \perp Y^m, Z | Y^o$, in which case

$$p_{\theta,\beta}(Y^o, R) = \int p_\theta(Y^o, Y^m, Z) p_\beta(R|Y^o, Y^m, Z) dY^m dZ^m = p_\theta(Y^o) p_\beta(R|Y^o)$$

Typology of observation process

Observation process (a.k.a "sampling design")

Some studied processes

Notation: **M(C)AR**, **MNAR**, $S_i = \mathbf{1}_{\{\text{node i is sampled}\}}$ (i.e., $R_{ij} = 1$ for all j)

Dyad-centered

- **Random dyad sampling**

$$R_{ij} \sim^{iid} \mathcal{B}(\rho)$$

- **Double standard sampling**

$$R_{ij}|Y_{ij} = 1 \sim^{ind} \mathcal{B}(\rho_1)$$
$$R_{ij}|Y_{ij} = 0 \sim^{ind} \mathcal{B}(\rho_0)$$

- **Block dyad sampling**

$$R_{ij}|Z_i, Z_j \sim^{ind} \mathcal{B}(\rho_{Z_i Z_j})$$

Node-centered

- **Node sampling**

$$S_i \sim^{iid} \mathcal{B}(\rho)$$

- **Degree sampling**,

$$S_i|D_i \sim^{ind} \mathcal{B}(\text{logistic}(a + bD_i))$$
$$D_i = \sum_j Y_{ij}$$

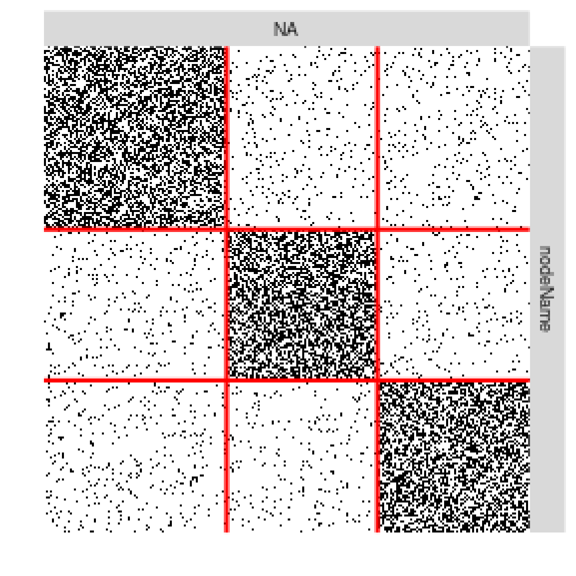
- **Block node sampling**

$$S_i|Z_i \sim^{ind} \mathcal{B}(\rho_{Z_i})$$

Observation proces: illustration

We first generate a community-shape network

```
## SBM parameters
N ← 300 # number of nodes
K ← 3   # number of clusters
alpha ← rep(1,K)/K     # block proportion
pi    ← list(mean = diag(.45,K) + .05 ) # connectivity matrix
## simulate an undirected binary SBM
sbm ← sbm::sampleSimpleSBM(N, alpha, pi)
plot(sbm)
```



Observation process: sample network data

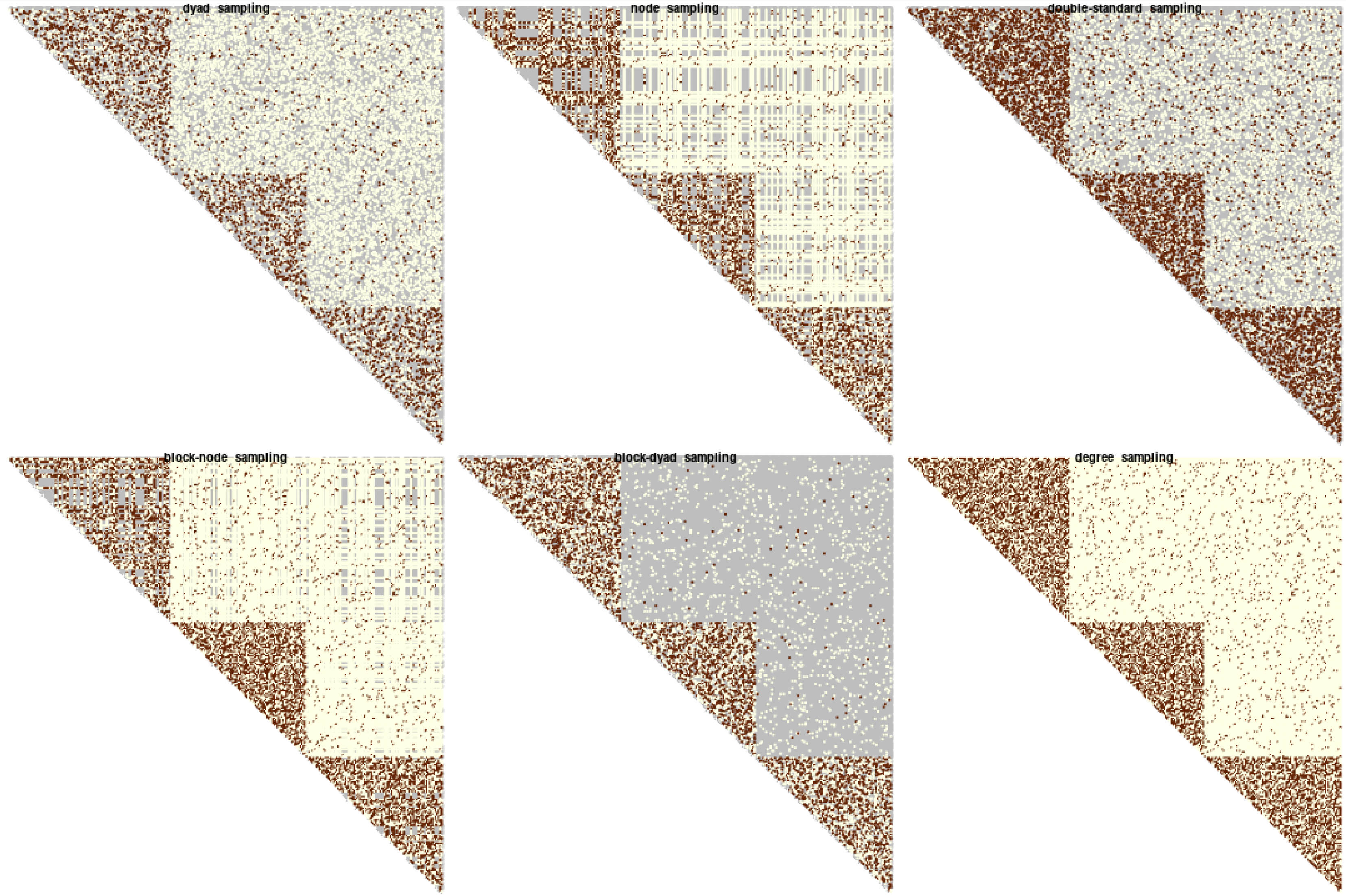
We consider some sampling designs and their associated parameters

```r
sampling_parameters <- list(
   "dyad" = .3,
   "node" = .3,
   "double-standard" = c(0.2, 0.6),
   "block-node" = c(.3, .8, .5),
   "block-dyad" = pi$mean,
   "degree" = c(.1, .2)
 )

observed_networks <- list()

for (sampling in names(sampling_parameters)) {
  observed_networks[[sampling]] <-
     missSBM::observeNetwork(
       adjacencyMatrix = sbm$networkData,
       sampling        = sampling,
       parameters      = sampling_parameters[[sampling]],
       cluster         = sbm$memberships
     )
}
```

Observation process: output



Identifiability

We build on the proof of [Cel+12] for Indentifiability of the SBM (sort marginal probabilities into a Vandermonde matrix which is invertible, so that we can express parameters $\boldsymbol{\pi}, \boldsymbol{\alpha}$ as a function of the original probailities).

SBM observed under MAR samplings (node/dyad centered)

Let $n \geq 2K$ and assume that for any $1 \leq k \leq K$, $\rho > 0$, $\alpha_k > 0$ and the coordinates of $\boldsymbol{\pi}.\boldsymbol{\alpha}$ are pairwise distinct. Then, under dyad (resp. node) sampling, SBM parameters are identifiable w.r.t. the distribution of the observed part of the SBM up to label switching.

SBM observed under block sampling

Let $n \geq 2K$ and assume that for any $1 \leq k \leq K$, $\rho_k > 0$, $\alpha_k > 0$ and the coordinates of $\boldsymbol{\pi}.\boldsymbol{\alpha}$ are pairwise distinct. If the coordinates $\left(\sum_k \pi_{1k}\rho_k\alpha_k, \dots, \sum_k \pi_{Kk}\rho_k\alpha_k\right)$ are pairwise distinct, under block sampling, $\boldsymbol{\theta}$ and β are identifiable w.r.t. the distributions of the SBM and the sampling up to label switching.

Inference of SBM from an observed network: MAR

Setting

We now need to estimate

- The SBM parameters $\boldsymbol{\theta} = \{(\boldsymbol{\alpha}, \boldsymbol{\Pi})\}$
- The sampling parameters β (e.g., $\boldsymbol{\rho}$, or $\boldsymbol{\rho}_k$, etc. depending on the design).

MAR case

Since

$$p_{\theta,\beta}(\boldsymbol{Y}^o, \boldsymbol{R}) = p_\theta(\boldsymbol{Y}^o) p_\beta(\boldsymbol{R}|\boldsymbol{Y}^o),$$

we just have to perform inference on the observed part of the data

$\rightsquigarrow$ "usual" V-EM (with possibility of saving memory footprint par sparsely encoding both $\mathbf{0}$ and **NA**).

Inference of SBM from an observed network: **MNAR**

Variational approximation

To evaluate $\mathbb{E}_{Z,Y^m|Y^o,R}(\,\cdot\,)$, the distribution $p_{\theta,\psi}(Z,Y^m|Y^o,R)$ is approximated by

$$q_\psi(Z,Y^m) = \prod_{i=1}^n m(Z_i;\tau_i) \prod_{Y_{ij}\in Y_{ij}^m} b(Y_{ij};\nu_{ij}) = \prod_{i=1}^n \prod_{k=1}^K (\tau_{ik})^{\mathbf{1}_{\{Z_i=k\}}} \cdot \prod_{Y_{ij}\in Y_{ij}^m} \nu_{ij}^{Y_{ij}} (1-\nu_{ij})^{1-Y_{ij}}$$

where $\psi = \{(\nu_{ij}),(\tau_{ik})\}$ are the variational parameters to be optimized

- τ_{ik} the posterior probabilities, are (almost) generic to any sampling design
- ν_{ij}, the imputation values, are specific to the sampling design.

M-step

- β, the sampling parameters, are specific to the design
- $\theta = (\boldsymbol{\alpha},\boldsymbol{\pi})$ are generic:

$$\hat{\alpha}_k = \frac{1}{n}\sum_i \hat{\tau}_{ik}, \qquad \hat{\pi}_{k\ell} = \frac{\sum_{(i,j)\in Y_{ij}^o} \hat{\tau}_{iq}\hat{\tau}_{j\ell} Y_{ij} + \sum_{(i,j)\in Y_{ij}^m} \hat{\tau}_{iq}\hat{\tau}_{j\ell}\hat{\nu}_{ij}}{\sum_{(i,j)} \hat{\tau}_{iq}\hat{\tau}_{j\ell}}.$$

General Variational EM for MNAR inference

Essentially separate computations for fitting the SBM / the sampling design

Initialize τ^0, ν^0 and β^0

Repeat

$$
\begin{array}{rclrr}
\theta^{(h+1)} & = & \arg\max_\theta J\left(Y^o, R;\ \tau^h, \nu^h, \beta^h, \theta\right) & \text{M-step a)} & \text{SBM} \\
\beta^{h+1} & = & \arg\max_\beta J\left(Y^o, R;\ \tau^h, \nu^h, \beta, \theta^{h+1}\right) & \text{M-step b)} & \text{Sampling} \\
\tau^{h+1} & = & \arg\max_\tau J\left(Y^o, R;\ \tau, \nu^h, \beta^{h+1}, \theta^{h+1}\right) & \text{VE-step a)} & \text{SBM} \\
\nu^{h+1} & = & \arg\max_\nu J\left(Y^o, R; \tau^{h+1}, \nu, \beta^{h+1}, \theta^{h+1}\right) & \text{VE-step b)} & \text{Sampling}
\end{array}
$$

Until $\left\|\theta^{h+1} - \theta^h\right\| < \varepsilon$

where we have the following decomposition:

$$
\begin{aligned}
J(Y^o, R) &= \mathbb{E}_{q_\psi}[\log p_{\theta,\beta}(Y^o, R, Y^m, Z)] + \mathcal{H}(q_\psi(Z, Y^m)) \\
&= \mathbb{E}_{q_\psi}[\log p_\beta(R|Y^o, Y^m, Z)] + \mathbb{E}_{q_\tau}[\log p_\theta(Y^o|Z)] + \mathbb{E}_{q_{\nu,\tau}}[\log p_\theta(Y^m|Z)] \\
&\qquad + \mathcal{H}(q_\tau(Z)) + \mathcal{H}(q_\nu(Y^m))
\end{aligned}
$$

Design specific updates

Example for Block-dyad sampling

Recall that

$$R_{ij}|Z_i, Z_j \sim^{ind} \mathcal{B}(\rho_{Z_i Z_j})$$

Then, the expected log-likelihood w.r.t the variational approximation $\boldsymbol{q}$ is

$$\mathbb{E}_{q_\psi}[\log p_\beta(R|Y^o, Y^m, Z)] = \sum_{(i,j)\in Y^o} \sum_{k,\ell} \tau_{ik}\tau_{j\ell} \log(\rho_{k\ell}) + \sum_{(i,j)\in Y^m} \sum_{k,\ell} \tau_{ik}\tau_{j\ell} \log(\rho_{k\ell}),$$

From which we derive

$$\hat{\rho}_{k\ell} = \frac{\sum_{(i,j)\in Y^o} \tau_{ik}\tau_{j\ell}}{\sum_{(i,j)\in Y} \tau_{ik}\tau_{j\ell}}$$

and

$$\hat{\nu}_{ij} = \mathrm{logistic}\left(\sum_{k,\ell} \tau_{ik}\tau_{j\ell} \log\left(\frac{\pi_{k\ell}}{1-\pi_{k\ell}}\right)\right)$$

Variational Estimators: theoretical guarantees

Consistency & Asymptotic Normality

Inspired by the two following papers:

- [Bic+13] deal with binary SBM under "sparse" conditions
- [BKM17] deal with LBM with distribution in the one-dimensional exponential family fully observed

Theorem [MT20]

Consider an SBM with K blocks and distribution in the *one-dimensional exponential family* under *random dyad sampling* and identifiability conditions (already explicited).

Then, maximum likelihood and variational estimators are *consistent* and *asymptotically normal* with explicit asymptotic variance/covariance matrix.

$\longrightarrow$ Only for MAR sampling !

SBM with covariates and missing data

Consider m external covariates $X_{ij} \in \mathbb{R}^m$ defined at the edge level. For covariates at the node level X_i, we can define a similarity $\phi(X_i, X_j) \to X_{ij}$.

$$Z_i \sim^{\text{iid}} \mathcal{M}(1, \alpha),$$
$$Y_{ij} \mid \{Z_i, Z_j, X_{ij}\} \sim^{\text{ind}} \mathcal{B}(\text{logistic}(\pi_{Z_i Z_j} + \eta^\top X_{ij}))$$

Dyad-centered sampling

Let $\delta \in \mathbb{R}$, $\kappa \in \mathbb{R}^m$. The probability to observe a dyad is

$$\mathbb{P}(R_{ij} = 1 | X_{ij}) = \text{logistic}(\delta + \kappa^T X_{ij}).$$

Node-centered sampling

Let $\delta \in \mathbb{R}$ and $\kappa \in \mathbb{R}^n$. The probability to observe all dyads corresponding to a node is

$$\mathbb{P}(S_i = 1 | X_i) = \text{logistic}(\delta + \kappa^T X_i).$$

These sampling designs are NMAR, however, conditionally to (x) they are MCAR

Illustrations

1. Numerical study of MNAR vs MAR
2. French blogosphere
3. PPI ER (ESR1) ego network

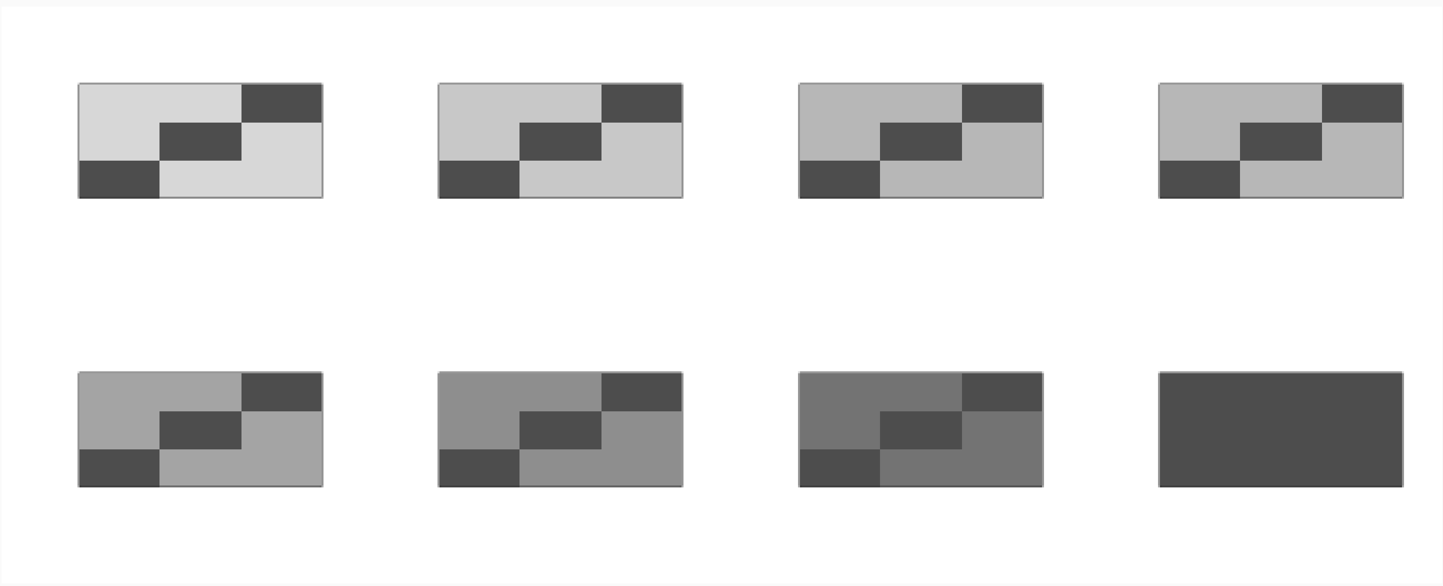
Block-dyad sampling

Consider a community like network:

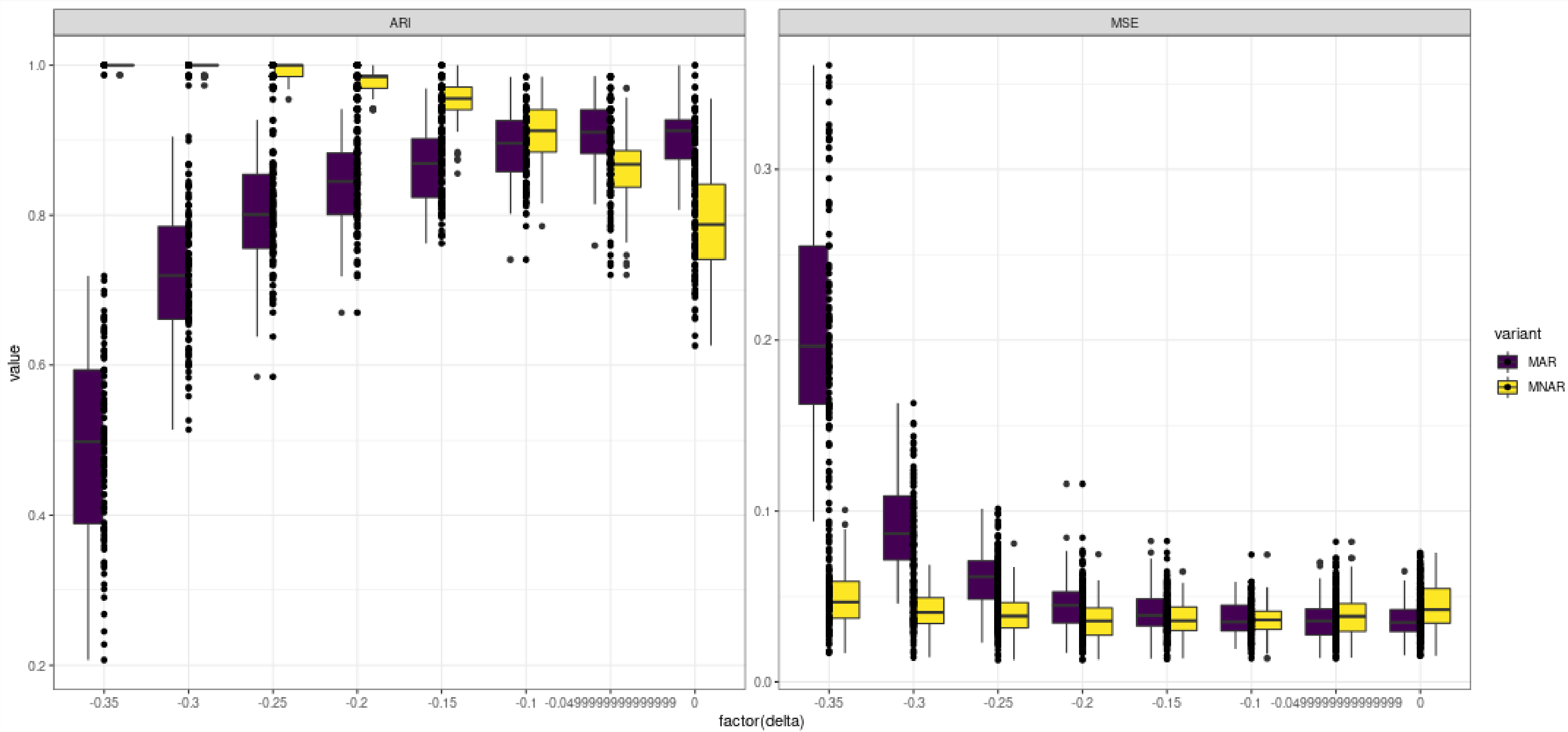
```r
n <- 200
alpha <-  c(1/3,1/3,1/3)
pi    <- .15 + diag(3) * .25
theta <- list(mean = pi)
pi
```

```
##      [,1] [,2] [,3]
## [1,] 0.40 0.15 0.15
## [2,] 0.15 0.40 0.15
## [3,] 0.15 0.15 0.40
```

Define sampling matrices with decreasing agreement with π



Performance of MAR vs MNAR



100 replicates. The closer δ to zero, the closer to the MAR case.

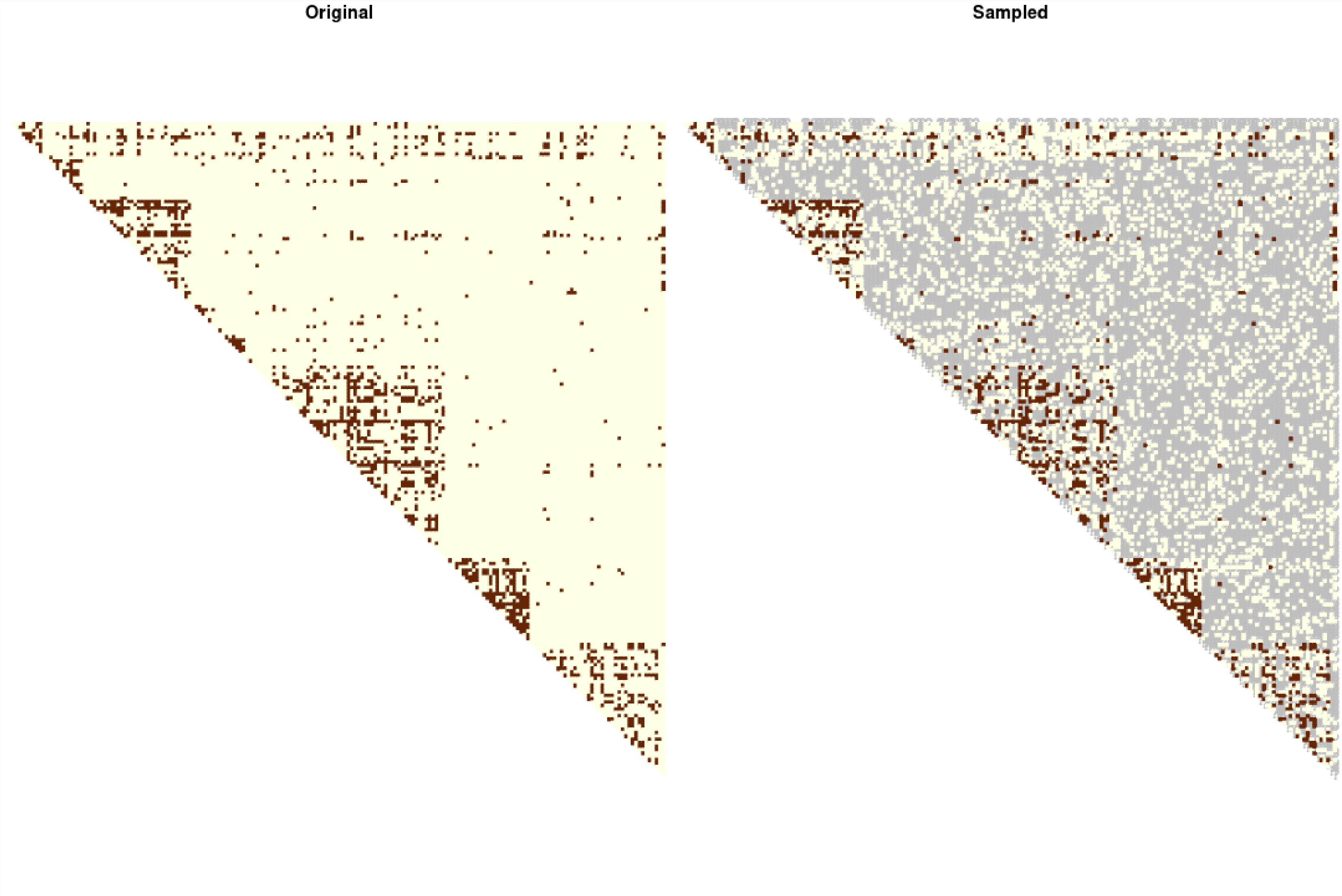
Back to French blogosphere

Control the network observation

- We sample in the original network to get a partly observed blog network
- We sampled more in the highly connected communities.

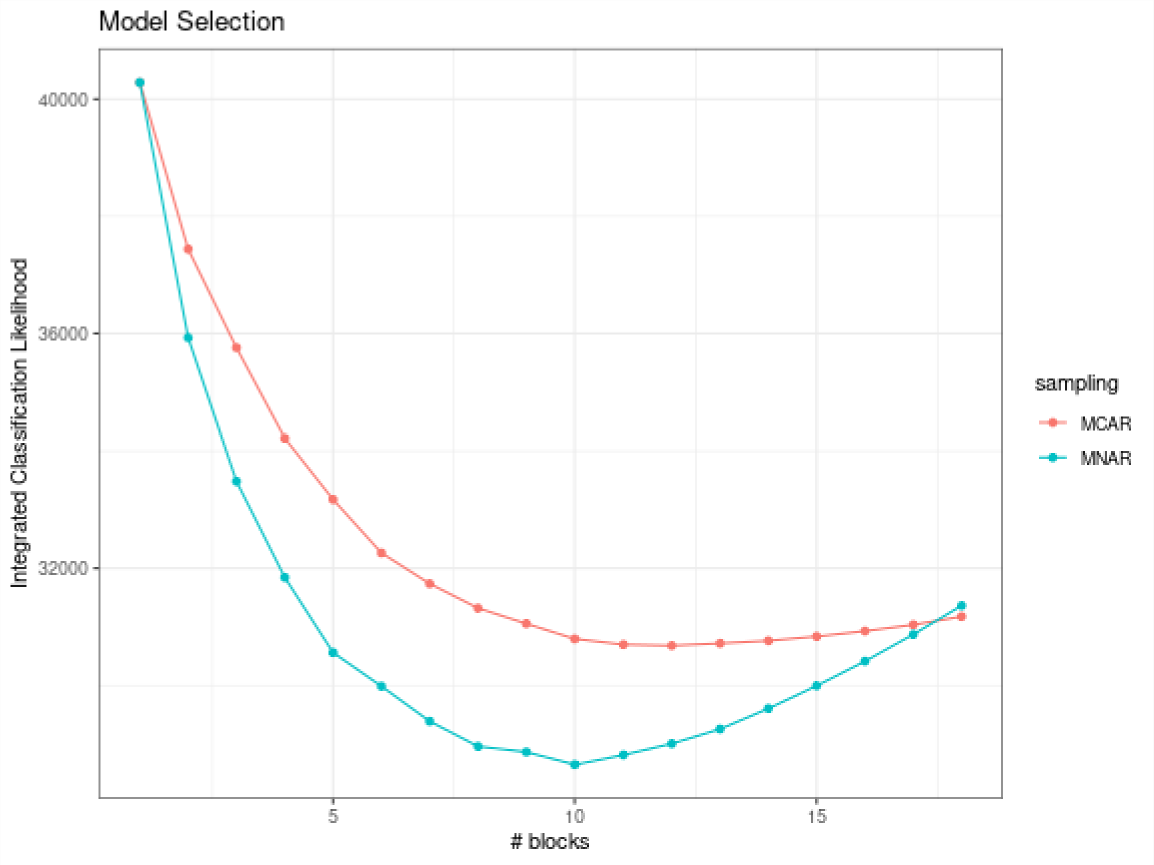
```
samplingParameters <-
  .2 + ifelse(sbm_full$bestModel$fittedSBM$connectParam$mean < .1, 0, .6)
blog_obs <-
  observeNetwork(
    adjacencyMatrix = blog,
    sampling        = "block-dyad",
    parameters      = samplingParameters,
    clusters        = sbm_full$bestModel$fittedSBM$memberships
  )
```

French blogosphere sampled



Compare MAR and NMAR with model selection criterion

```
sbm_mar  ← estimateMissSBM(blog_obs, blocks, "dyad")
sbm_mnar ← estimateMissSBM(blog_obs, blocks, "block-dyad")
```



Validation?

Compare the clustering of the different models with the original *party* classification:

```
ARI(party, sbm_full$bestModel$fittedSBM$memberships)
```

```
## [1] 0.4517304
```

```
ARI(party, sbm_mar$bestModel$fittedSBM$memberships)
```

```
## [1] 0.3501788
```

```
ARI(party, sbm_mnar$bestModel$fittedSBM$memberships)
```

```
## [1] 0.4550551
```

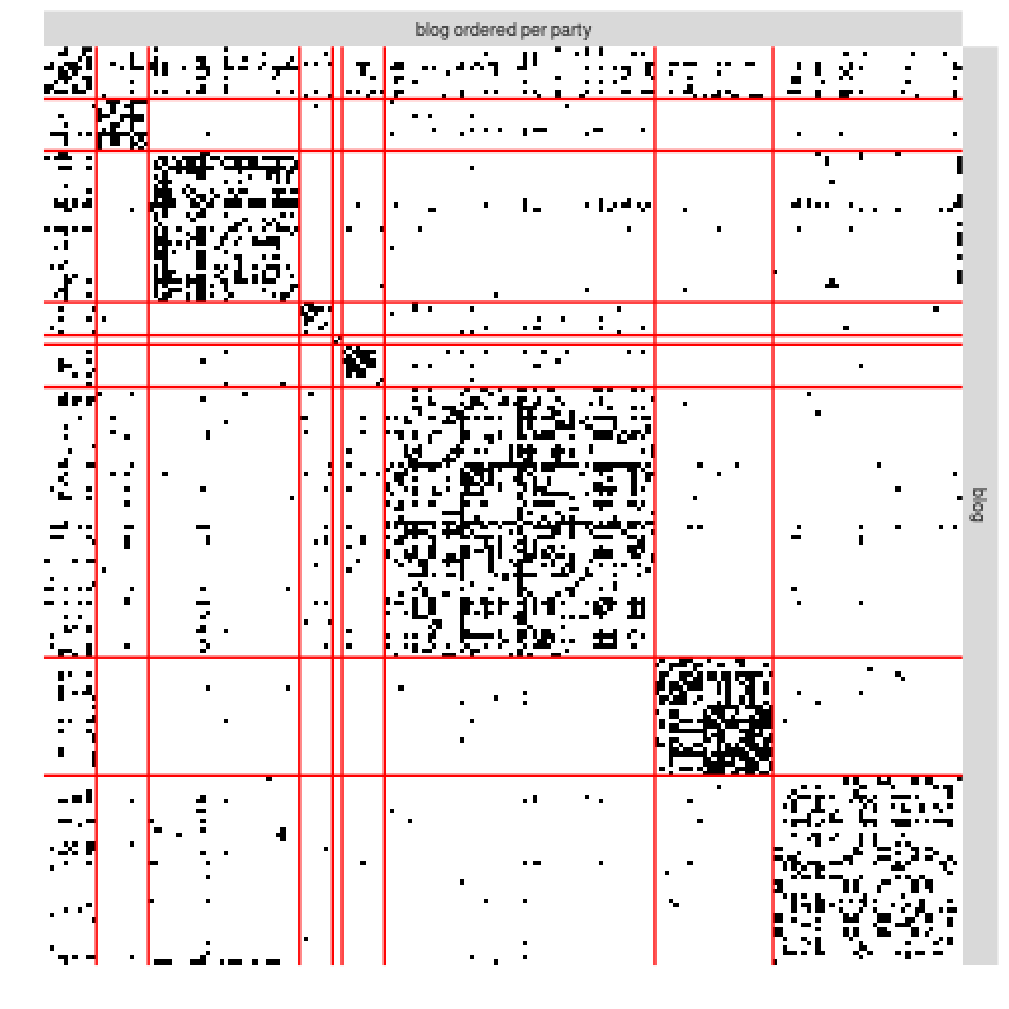
```
ARI(sbm_mnar$bestModel$fittedSBM$memberships, sbm_full$bestModel$fittedSBM$memberships)
```

```
## [1] 0.7758579
```

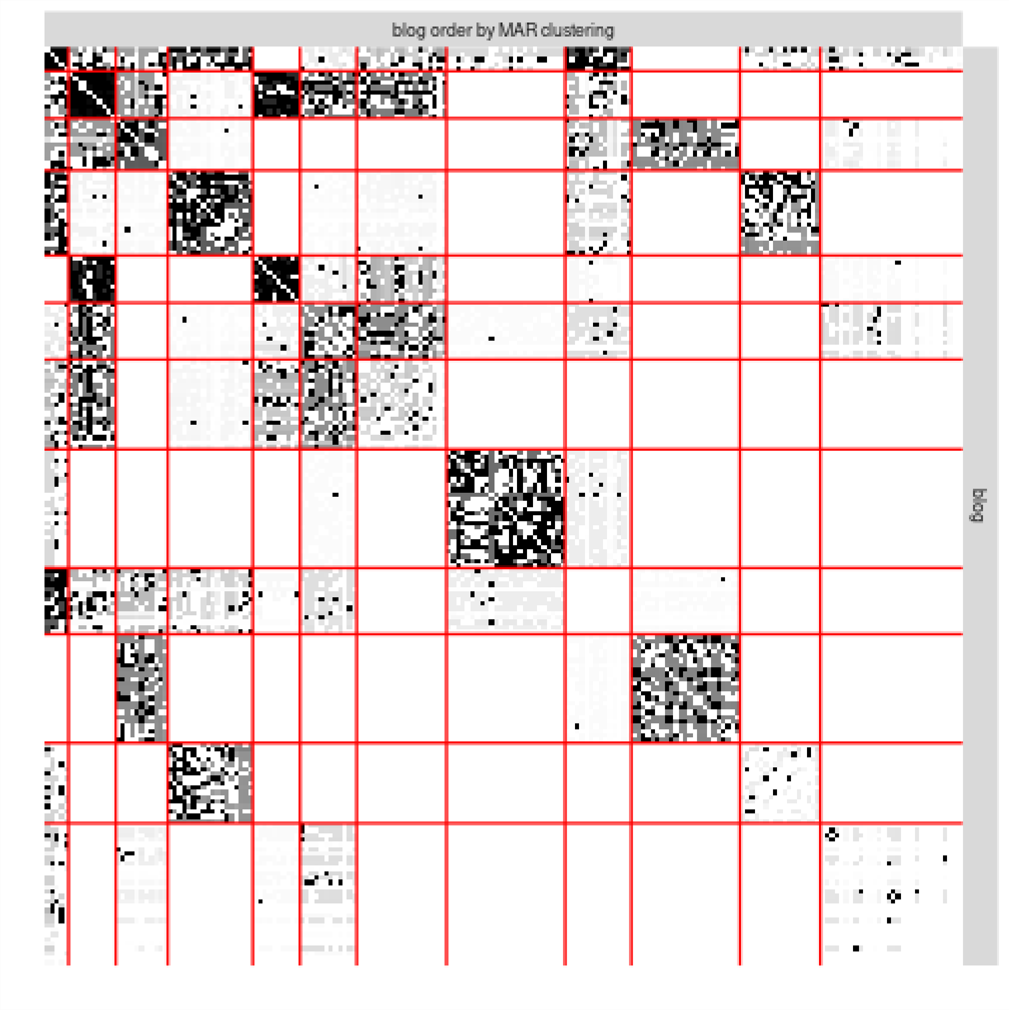
```
ARI(sbm_mnar$bestModel$fittedSBM$memberships, sbm_mar$bestModel$fittedSBM$memberships)
```

```
## [1] 0.4766329
```

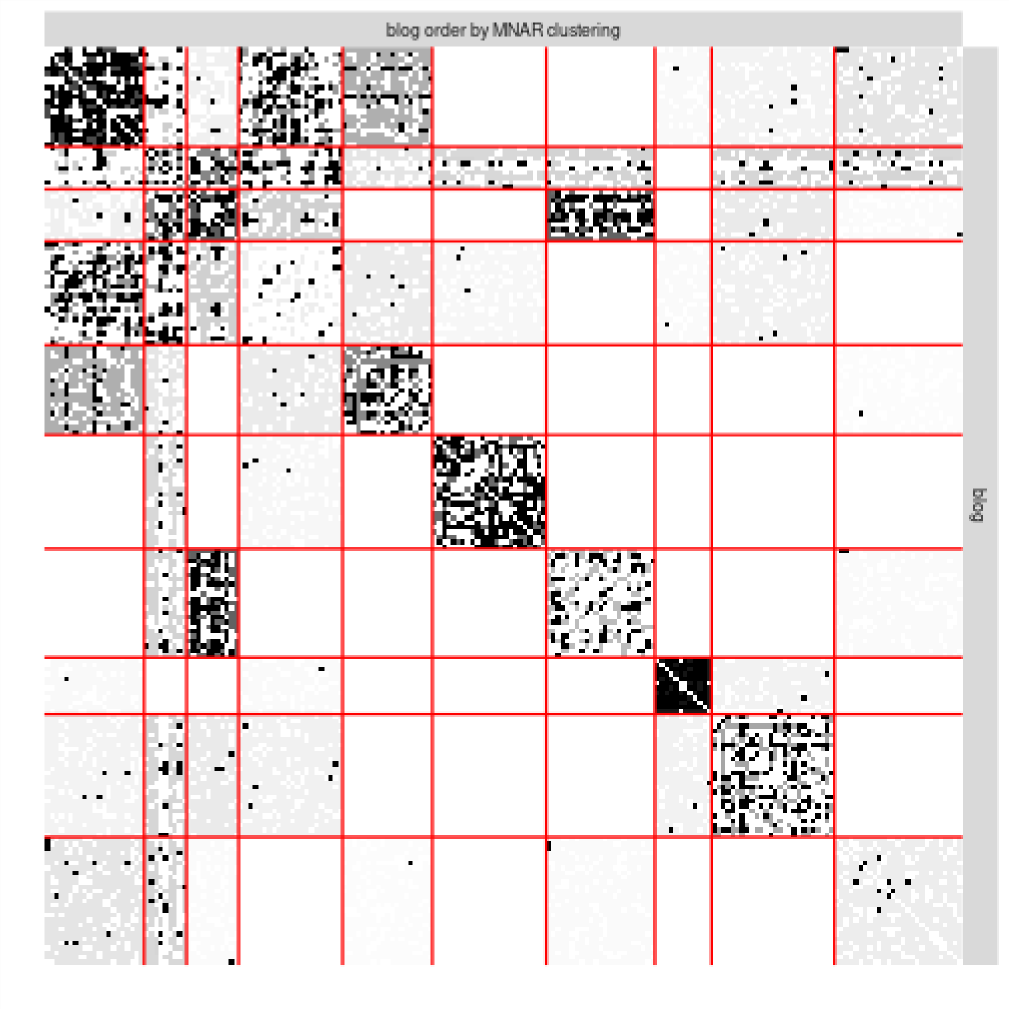
French blogosphere: original classification



French blogosphere: MAR clustering



French blogosphere: MNAR clustering



Protein-Protein Network

The data

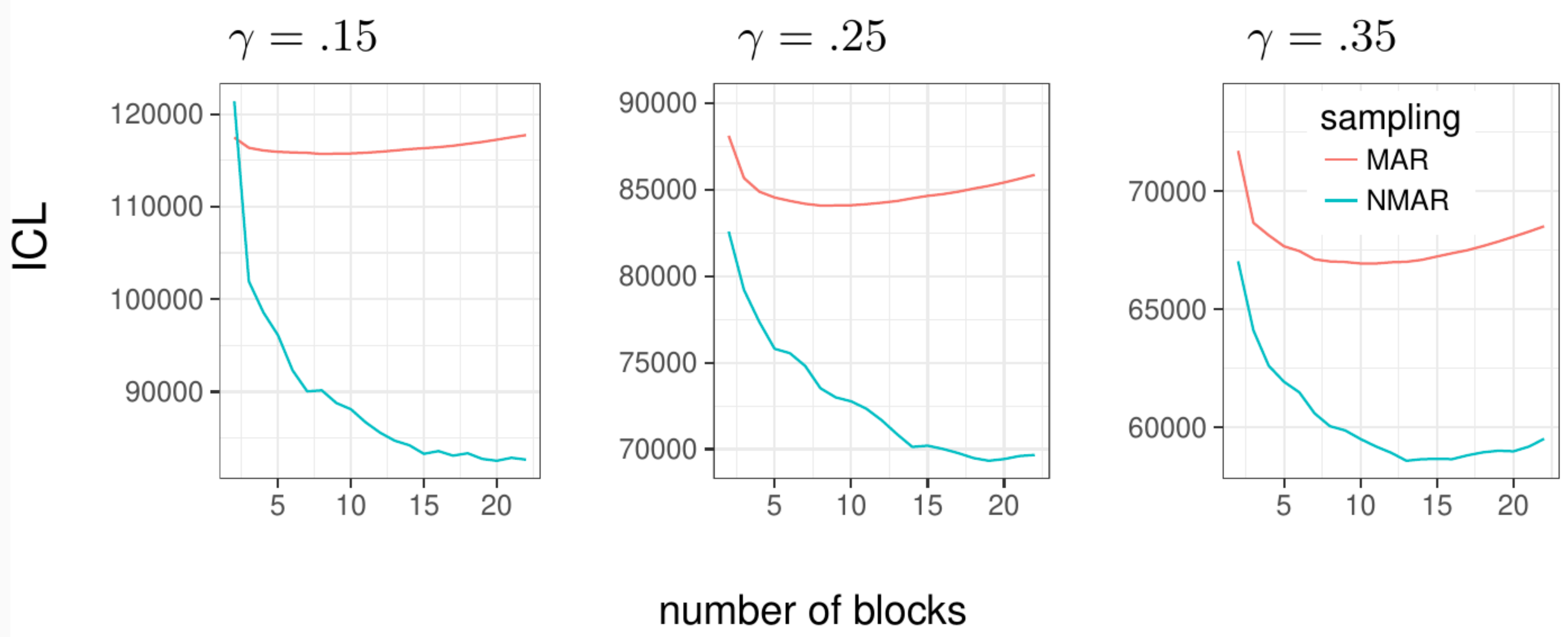
- The PPI network in the neighborhood of ER composed by 741 proteins
- Valued dyads: $\omega_{ij} \in (0, 1]$ reflecting the level of confidence in the interaction
- Binarization of the network with a threshold γ:

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

Questions

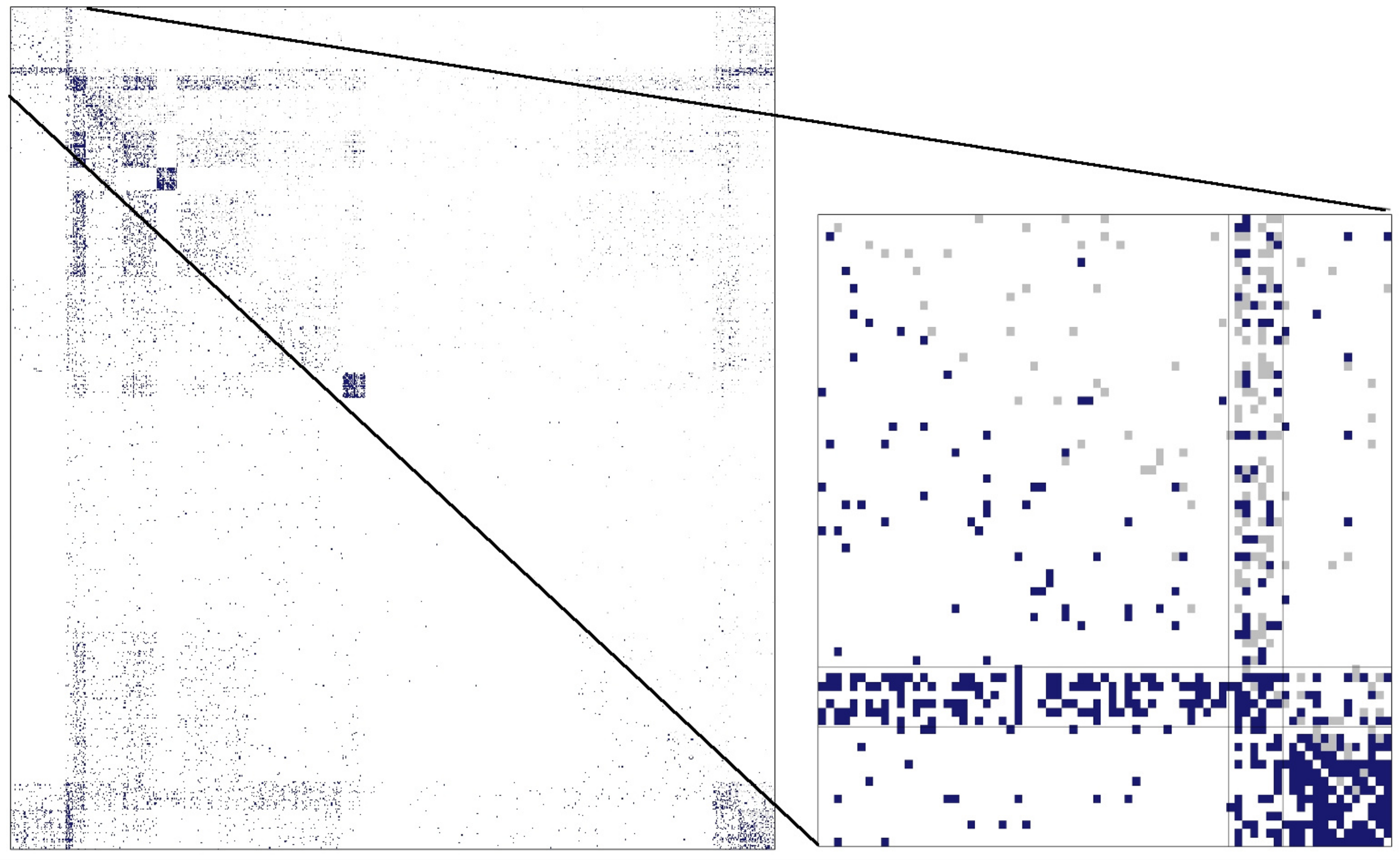
- What γ?
- What sampling design: MAR or NMAR?

Model Selection



- The ICL criterion selects $\gamma = .35$ and MNAR sampling as the one that better fit the data
- Number of selected clusters: 11 (MAR) and 13 (NMAR)
- ARI between NMAR clustering and MAR clustering: $.39$
- MNAR clustering somehow coherent with gene ontology

Imputation



Gene Ontology (GO)

Enrichment analysis *i.e.* identifying classes of genes over-represented in a large set of genes

MNAR founde $\mathbf{13}$ significant biological process founded (MAR: only $\mathbf{1}$)

Conclusion

Perspectives/ongoing

- Sampling
  - study robustness (block-sampling "includes" double-standard?)
- Other models
  - degree-corrected SBM
  - (ZI)-Poisson emission law
  - Simple SBM $\longrightarrow$ Bipartite SBM (aka Latent block models)
- Other algorithms
  - SGD algorithms + Pytorch framework
  - 'Exact' ICL maximization (with É. Côme)

Advertisement

https://computo.sfds.asso.fr, a new journal promoting reproducible research

THANK YOU

References

Barbillon, P., J. Chiquet, and T. Tabouy (2022). "misssbm: An r package for handling missing values in the stochastic block model". In: *Journal of Statistical Software*.

Bickel, P., D. Choi, X. Chang, et al. (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *Ann. Stat.* 41.4, pp. 1922-1943.

Brault, V., C. Keribin, and M. Mariadassou (2017). "Consistency and Asymptotic Normality of Latent Blocks Model Estimators". working paper or preprint.

Celisse, A., J. Daudin, L. Pierre, et al. (2012). "Consistency of maximum-likelihood and variational estimators in the stochastic block model". In: *Electron. J. Stat.* 6, pp. 1847-1899.

Little, R. J. and D. B. Rubin (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Mariadassou, M. and T. Tabouy (2020). "Consistency and asymptotic normality of stochastic block models estimators from sampled data". In: *Electronic Journal of Statistics* 14.2, pp. 3672-3704.

Tabouy, T., P. Barbillon, and J. Chiquet (2019). "Variational Inference for Stochastic Block Models from Sampled Data". In: *Journal of the American Statistical Association* 0.ja, pp. 1-20. DOI: 10.1080/01621459.2018.1562934.