EXPLORATION OF DENSE SBM VIA A RANDOM WALK

Presented by **THUY VO**
Supervised by CHI TRAN

Aussois Summer School 13-18 Juin 2021

Motivations

The motivation of this work is to discover the structure and the topology of a hidden network: drug users, MSM,...

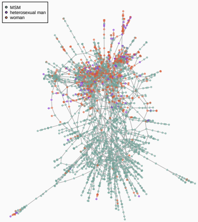


Figure 1: Sexual contacts in a population in Cuba[1]

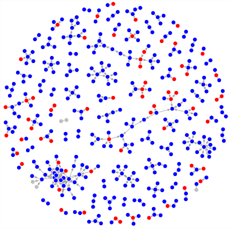


Figure 2: RDS on the HCV population[2]

➥ detect the identities of hidden individuals by exploring the graphs.

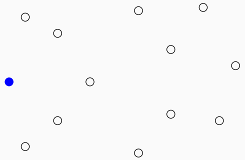
➥ Proposed methods: Respondent Driven Sampling (RDS)[3],...

1. Clémençon et al. (2015)
2. Jauffret-Roustide et al. en cours (2020)
3. Respondent Driven Sampling: a new approach to the study of hidden populations; Heckathorn (1997)

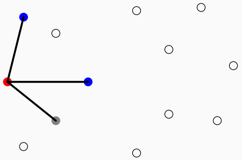
Respondent Driven Sampling

There are c coupons distributed at each turn of the interview.

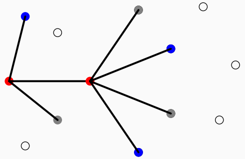
- interviewed
- having coupon but have not been interviewed yet
- have been named but without coupon



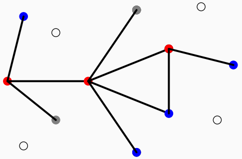
Step 0



Step 1



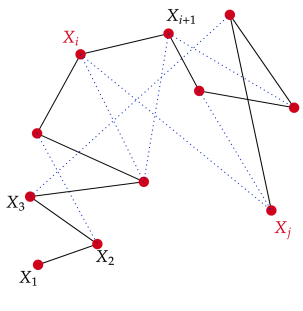
Step 2



Step 3

Random walk: RDS with c=1

★ Denote $X^{(n)} = (X_1, ..., X_n)$ the explored nodes after n steps.



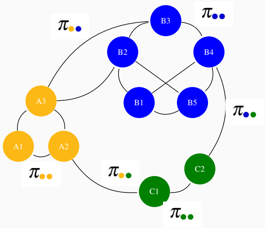
★ $H_n = (V_n, E_n)$ the path of nodes visited by the random walk:

$$V_n = \{X_1, ..., X_n\} \text{ and } E_n = \cup_{i=1}^{n-1}\{X_i, X_{i+1}\}$$

★ $G_n = G(X^{(n)}, H_n, \kappa)$: the subgraph discovered.

Stochastic Block Model (SBM) and schema of the observations on graphs

Stochastic Block Model[1]



Q blocks (classes)
$\alpha = (\alpha_1, ..., \alpha_Q)$ proportions of blocks
$\pi = (\pi_{qr})_{q,r \in [\![1,Q]\!]}$ probabilities of connection

★ The observations:

- the random walk $X^{(n)}$;
- the types $Z = (Z_1, ..., Z_n)$;
- the adjacency matrix: $Y = (Y_{ij})_{i,j \in \{1,...,n\}}$.

★ The parameter to estimate: $\theta = (\alpha, \pi)$.

1. The graph of SBM is draw by Julien Chiquet and Catherine Matias.

Graphon is a symmetric function: $\kappa : [0,1]^2 \mapsto [0,1]$.

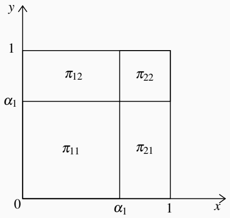
➥ Associate to the finite graph of n vertices a graphon by its adjacency matrix $(Y_{ij})_{1 \leq i,j \leq n}$:

$$\kappa \ : \ (x,y) \mapsto \mathbf{1}_{Y_{\lceil nx \rceil, \lceil ny \rceil}=1}.$$

➥ When the size of the graph is "infinite":

- Erdös-Rényi $\kappa \equiv p$;
- SBM(Q, α, π): $I = (I_1, ..., I_Q)$ a partition of [0,1] such that $|I_q| = \alpha_q$. Then

If $x \in I_q, \ y \in I_r, \quad \kappa(x,y) = \pi_{qr}$.



Metric on the space of graphs, graphons

☆ For a graph G of n vertices and F a graph of $k \leq n$ vertices, we define:

$$t(F, G) = \frac{|\text{inj}(F, G)|}{(n)_k}$$

and for the graphon κ:

$$t(F, \kappa) = \int_{[0,1]^k} \prod_{\{i,j\} \in E(F)} \kappa(x_i, x_j) dx_1 \ldots dx_k.$$

☆ Let $(F_i)_{i \in \mathbb{N}^*}$ be an enumeration of all the finite graphs. We define:

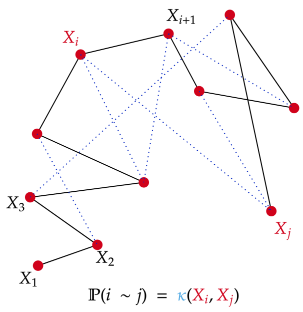
$$d_{sub}(G, \kappa) = \sum_{i \geq 1} \frac{1}{2^i} |t(F_i, G) - t(F_i, \kappa)|$$

Prop: When the size of graph G_n tends to infinity, the SBM graph converges to an SBM graphon for the distance d_{sub}.

The random walk on a graphon

★ $X^{(n)} = (X_1, \ldots, X_n)$ a RW on κ, $X_i \in [0,1]$ with the transition kernel:

$$P(x, dy) = \frac{\kappa(x,y)dy}{\int_0^1 \kappa(x,v)dv}$$



★ $X^{(n)}$ admits a stationary measure:

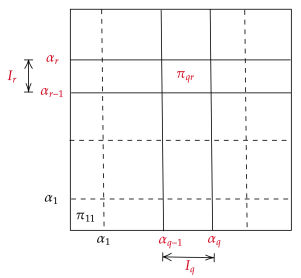
$$m(dx) = \frac{\int_0^1 \kappa(x,v)dv}{\int_0^1 \int_0^1 \kappa(u,v)du\ dv}\ dx.$$

★ $G_n = G(X^{(n)}, H_n, \kappa)$ constructed by $X^{(n)}$ and the graphon κ.

The random walk on a dense SBM

For the SBM(Q, α, π), the associated graphon is:

$$\kappa(x,y) = \sum_{q=1}^{Q}\sum_{r=1}^{Q} \pi_{qr}\ \mathbf{1}_{I_q}(x)\mathbf{1}_{I_r}(y).$$



Prop:
The random walk $X^{(n)}$ on the graphon κ admits unique invariant measure:

$$m(dx) = \frac{\int_0^1 \kappa(x,v)dv}{\int_0^1\int_0^1 \kappa(u,v)du\ dv}\ dx = \frac{\sum_{q=1}^{Q}\left(\sum_{r=1}^{Q}\pi_{qr}\alpha_r\right)\mathbf{1}_{I_q}(x)}{\sum_{q=1}^{Q}\sum_{r=1}^{Q}\pi_{qr}\alpha_q\alpha_r} dx. \qquad (1)$$

Limit of explored subgraph

Proposition[1]

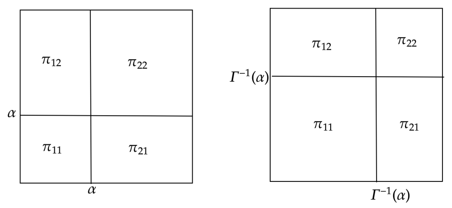
$$\lim_{n\to\infty} d_{sub}\left(G_n, \kappa_{\Gamma^{-1}}\right) = 0, \qquad \text{a.s.}$$

where Γ is distribution function of m and Γ^{-1} is the generalized inverse of Γ and

$$\kappa_{\Gamma^{-1}}(x, y) = \kappa(\Gamma^{-1}(x), \Gamma^{-1}(y)).$$

For $Q = 2$:

$$\Gamma(\alpha) = \frac{(\pi_{11}\alpha + \pi_{12}(1-\alpha))\alpha}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2}$$



➥ How can we estimate κ from the subgraph G_n?.

1. Dense graph limits under Respondent Driven Sampling; Athreya and Röllin. Annals of Applied Probability (2016).

Estimation methods

Complete observations

- Method 1: Maximum likelihood
- Method 2: De-biased graphon

Incomplete observations

- Method 3: Maximum likelihood by SAEM
- Method 4: De-biased graphon by VEM

Complete observations

★ Suppose that $X^{(n)}, Z, Y$ are observed:

$$N_n^q = \text{number of nodes of type } q$$
$$N_n^{q\leftrightarrow r} = \text{number of edges of type } qr.$$

For the SBM without bais:

$$\mathcal{L}(Z_i, Y_{ij}; i,j \in X^{(n)}; \theta) = \frac{n!}{N_n^1! \cdots N_n^Q!} \prod_{q=1}^{Q} \alpha_q^{N_n^q} \times \prod_{\substack{1\leq i,j\leq n \\ i\neq j}} \pi_{Z_i Z_j}^{Y_{i,j}} (1-\pi_{Z_i Z_j})^{(1-Y_{i,j})}$$

★Without biases, the classical MLE:

$$\widehat{\alpha}_q^{\text{class}} = \frac{N_n^q}{n}, \qquad \widehat{\pi}_{qr}^{\text{class}} = \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r}, \qquad \widehat{\pi}_{qq}^{\text{class}} = \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q - 1)}.$$

Method 1: Complete observations + ML

★ With the biases:

$$\mathcal{L}(Z_i, Y_{ij}; i,j \in X^{(n)}; \theta) = \frac{\prod_{i=1}^{n} \alpha_{Z_i}}{\prod_{i=1}^{n-1} \sum_{q=1}^{Q} \pi_{Z_i q}\alpha_q} \times \prod_{\substack{1 \le i,j \le n \\ i \ne j}} \pi_{Z_i Z_j}^{Y_{ij}} (1 - \pi_{Z_i Z_j})^{1-Y_{ij}}, \tag{2}$$

Proposition

The ML estimator $\widehat{\theta} = (\widehat{\pi}, \widehat{\alpha})$ is solution of:

$$\frac{N_n^q}{\widehat{\alpha}_q} - \sum_{p=1}^{Q} \frac{(N_n^p - \mathbf{1}_{Z_n=p})\widehat{\pi}_{pq}}{\sum_{q'=1}^{Q} \widehat{\pi}_{pq'}\widehat{\alpha}_{q'}} = \frac{N_n^r}{\widehat{\alpha}_r} - \sum_{p=1}^{Q} \frac{(N_n^p - \mathbf{1}_{Z_n=p})\widehat{\pi}_{pr}}{\sum_{q'=1}^{Q} \widehat{\pi}_{pq'}\widehat{\alpha}_{q'}};$$

$$\frac{N_n^{q\leftrightarrow q}}{\widehat{\pi}_{qq}} - \frac{N_n^{q\not\leftrightarrow q}}{1-\widehat{\pi}_{qq}} - \frac{(N_n^q - \mathbf{1}_{Z_n=q})\widehat{\alpha}_q}{\sum_{q'=1}^{Q} \widehat{\pi}_{qq'}\widehat{\alpha}_{q'}} = 0;$$

$$\frac{N_n^{q\leftrightarrow r}}{\widehat{\pi}_{qr}} - \frac{N_n^{q\not\leftrightarrow r}}{1-\widehat{\pi}_{qr}} - \frac{(N_n^q - \mathbf{1}_{Z_n=q})\widehat{\alpha}_r}{\sum_{q'=1}^{Q} \widehat{\pi}_{qq'}\widehat{\alpha}_{q'}} - \frac{(N_n^r - \mathbf{1}_{Z_n=r})\widehat{\alpha}_q}{\sum_{q'=1}^{Q} \widehat{\pi}_{rq'}\widehat{\alpha}_{q'}} = 0 \quad \text{if } q \ne r.$$

Method 2: Complete observation + de-biased graphon(1/2)

★ By Athreya & Röllin: $G_n \longrightarrow \kappa_{\Gamma-1}$, where $\kappa_{\Gamma-1} =: \kappa_{\widetilde{\theta}}$ and $\widetilde{\theta} := (\widetilde{\alpha}, \pi)$.

The classical estimator for $\widetilde{\alpha}, \pi$ (neglecting the biases):

$$\widehat{\lambda}_q^n := \frac{N_n^q}{n};$$
$$\widehat{\pi}_{qr}^n := \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r} \quad \text{for} \quad q \neq r \quad \text{and} \quad \widehat{\pi}_{qq}^n := \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q - 1)}.$$

★ $\widehat{\chi}_n(x, y)$ the graphon associated to $(\widehat{\lambda}^n, \widehat{\pi}^n)$.

Proposition

(i) When $n \to +\infty$,

$$\lim_{n\to+\infty} d_{sub}(G_n, \widehat{\chi}_n) = 0. \tag{3}$$

(ii) The limit $\widehat{\chi}_n$ is then the biased graphon $\kappa_{\Gamma-1}$.

$$\lim_{n\to+\infty} d_{\mathrm{sub}}(\widehat{\chi}_n, \kappa_{\Gamma-1}) = 0. \tag{4}$$

Method 2: Complete observation + de-biased graphon(2/2)

★ The 2-stage estimation:

1st step: Estimate $\widetilde{\theta} = (\widetilde{\alpha}, \pi)$:

- $\widehat{\pi}^n$ is a consistent estimator of π:

$$\lim_{n \to +\infty} \widehat{\pi}^n = \pi_{qr},$$

- and $\widehat{\lambda}_q^n$ is a consistent estimator of $\widetilde{\alpha}$:

$$\lim_{n \to +\infty} \widehat{\lambda}_q^n = \Gamma\big(\sum_{r=1}^{q} \alpha_r\big) - \Gamma\big(\sum_{r=1}^{q-1} \alpha_r\big) = \widetilde{\alpha}_q.$$

2nd step: Correct the estimator $\widetilde{\theta}$ to obtain θ

A consistent estimator of α_q is

$$\widehat{\alpha}_q^n = \Gamma_n^{-1}\big(\sum_{r=1}^{q} \widehat{\lambda}_r^n\big) - \Gamma_n^{-1}\big(\sum_{r=1}^{q-1} \widehat{\lambda}_r^n\big). \tag{5}$$

In the case $Q = 2$, an estimator for α_1 is $\widehat{\alpha}_1^n = \Gamma_n^{-1}(\widehat{\lambda_1^n})$.

Method 3 (1/2): Incomplete observations + SAEM

Suppose that we observe only Y_{ij} and Z_i are unknown.

★ The incomplete likelihood:

$$\mathcal{L}(Y_{ij}; i,j \in [\![1,n]\!]; \theta) = \sum_{q_1,\cdots q_n=1}^{Q} \Bigg[\prod_{i=1}^{n} \mathbf{1}_{Z_i=q_i} \frac{\prod_{i=1}^{n} \alpha_{q_i}}{\prod_{i=1}^{n-1} \sum_{q=1}^{Q} \pi_{q_i q} \alpha_q} \times \prod_{\substack{1 \le i,j \le n \\ i \neq j}} b(Y_{ij}, \pi_{q_i q_j}) \Bigg],$$

➥ The sum of $q \in \{1, .., Q\}$ is not tractable.

➥ Use the SAEM approach the MLE numerically.

Method 3 (2/2): Incomplete observations + SAEM

Given $\theta^{(k-1)} = (\alpha^{(k-1)}, \pi^{(k-1)})$, at the iteration k^{eme}:

★ Step 1: Choose the appropriate proposal Z;

We follow the variational approach of Daudin et al.[1]: choose Z_i by a multinomal distribution of parameter τ_{iq},

$$\tau_{iq} \propto \frac{\alpha_q}{\sum_{\ell=1}^{Q} \pi_{q\ell}\alpha_\ell} \prod_{i \neq j} \prod_{\ell=1}^{Q} b(Y_{ij}, \pi_{q\ell})^{\tau_{j\ell}}. \tag{6}$$

★ Step 2: Stochastic approximation, update the quantity:

$$\mathcal{Q}^{(k)}(\theta) = \mathcal{Q}^{(k-1)}(\theta) + s_k \left(\log \mathcal{L}(Z_i^{(k)}, Y_{ij}, \theta) - \mathcal{Q}^{(k-1)}(\theta) \right);$$

★ Step 3: Maximization:

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{Q}^{(k)}(\theta).$$

1. Coupling a stochastic approximation version of EM with an MCMC procedure; Kuhn and Lavielle. ESAIM:ps (2004).

Method 4a: Incomplete observations (Z is unobserved and $X^{(n)}$ is observed) + graphon de-biasing

Suppose that $(Z_1, ..., Z_n)$ are unobserved, but the positions $(X_1, ..., X_n)$ are observed.

Step 1: Neglecting the sampling biases and using the variational EM algorithm (VEM):

- Using EM algorithm to estimate (λ, π);
- Choosing the types Z_i based on the information of $X^{(n)}$.

Step 2: Estimate the cumulative distribution function Γ_n, then deduce the estimator $\widehat{\alpha}^n$ of α and thus the estimator of κ:

$$\widehat{\kappa}_n(x, y) := \sum_{q=1}^{Q} \sum_{r=1}^{Q} \widehat{\pi}_{qr}^n \mathbf{1}_{[\sum_{k=1}^{q-1} \widehat{\alpha}_k^n, \sum_{k=1}^{q} \widehat{\alpha}_k^n)}(x) \mathbf{1}_{[\sum_{k=1}^{r-1} \widehat{\alpha}_k^n, \sum_{k=1}^{r} \widehat{\alpha}_k^n)}(y). \tag{7}$$

Method 4b: Incomplete observations (Z is unobserved and $X^{(n)}$ is observed) + graphon de-biasing

When $Z = (Z_1, ..., Z_n)$ and $X^{(n)} = (X_1, ..., X_n)$ are unobserved:

$$\widetilde{\alpha}_q = \frac{\alpha_q \bar{\pi}_q}{\bar{\pi}}, \quad \text{for all } q \in \{1, \dots Q\} \quad \Leftrightarrow \widetilde{\alpha} = \frac{\alpha \odot (\pi\alpha)}{\alpha^T \pi \alpha},$$

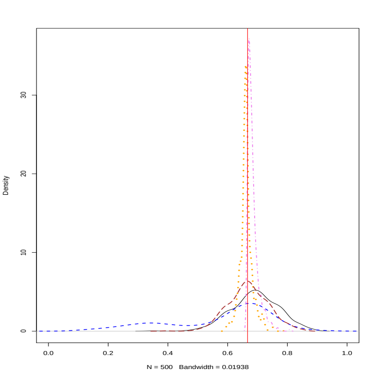
➥ Estimator $\widehat{\alpha}$ for the vector $\alpha = (\alpha_1, \dots \alpha_Q)$ can be obtained from solving the equation:

$$(\widehat{\alpha}^T \widehat{\pi} \widehat{\alpha}) \widehat{\lambda} = \widehat{\alpha} \odot (\widehat{\pi} \widehat{\alpha}).$$

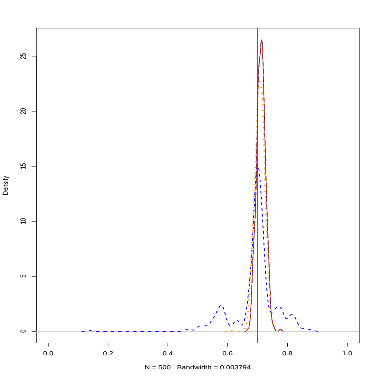
It leads to solve the optimization problem

$$\min_{x \in S} \|(x^T \widehat{\pi} x) \widehat{\lambda} - x \odot (\widehat{\pi} x)\|,$$

where $S = \left\{x = (x_1, \cdots, x_Q) \in [0;1]^Q : x_1 + ... + x_Q = 1\right\}$.



(a)



(b)

Figure 3: *E*stimation by the complete data for a graph of $n = 60$ vertices with $Q = 2$ classes and parameters $\alpha_1 = 2/3$, $\pi_{11} = 0.7$, $\pi_{12} = \pi_{21} = 0.4$ and $\pi_{22} = 0.8$. 500 such graphs are simulated and the empirical distributions of the estimators are represented here with the true parameters in red line. (a): estimator of α, (b):estimator of π_{11}.

Simulations:

Parameters	Complete likelihood	SAEM	De-biased graphon	De-biasing & SAEM	De-biasing & alg. eq.
π_{11}	$3.52\ 10^{-4}$	$5.25\ 10^{-3}$	$3.52\ 10^{-4}$	$3.54\ 10^{-4}$	$3.54\ 10^{-4}$
π_{12}	$4.99\ 10^{-4}$	$5.14\ 10^{-3}$	$4.99\ 10^{-4}$	$6.65\ 10^{-4}$	$4.99\ 10^{-4}$
π_{22}	$1.41\ 10^{-3}$	$1.45\ 10^{-2}$	$1.41\ 10^{-3}$	$1.42\ 10^{-3}$	$1.41\ 10^{-3}$
α	$7.01\ 10^{-3}$	$3.80\ 10^{-2}$	$6.80\ 10^{-4}$	$5.31\ 10^{-4}$	$4.51\ 10^{-3}$

Table 1: *Mean square errors.*

Merci de votre attention !!!