Estimation of dense stochastic block models visited by random walks

Viet-Chi Tran, Thuy Vo Thi Phuong

Estimation of dense stochastic block models visited by random walks

Viet Chi Tran[*] Thi Phuong Thuy Vo[†]

June 15, 2020

Abstract

We are interested in recovering information on a stochastic block model from the subgraph discovered by an exploring random walk. Stochastic block models correspond to populations structured into a finite number of types, where two individuals are connected by an edge independently from the other pairs and with a probability depending on their types. We consider here the dense case where the random network can be approximated by a graphon. This problem is motivated from the study of chain-referral surveys where each interviewee provides information on her/his contacts in the social network. First, we write the likelihood of the subgraph discovered by the random walk: biases are appearing since hubs and majority types are more likely to be sampled. Even for the case where the types are observed, the maximum likelihood estimator is not explicit any more. When the types of the vertices is unobserved, we use an SAEM algorithm to maximize the likelihood. Second, we propose a different estimation strategy using new results by Athreya and Roellin. It consists in de-biasing the maximum likelihood estimator proposed in Daudin et al. and that ignores the biases.

Keywords: random graph; graphon; random walk exploration; sampling bias; EM estimation; stochastic approximation expectation-maximization; incomplete likelihood; respondent driven sampling; chain-referral survey.
AMS Classification: 62D05; 05C81; 05C80; 60J20; secondary: 82C20

Acknowledgements: The authors thank Jean-Stéphane Dhersin, Sophie Donnet, Stéphane Robin and Timothée Tabouy for discussions. This work was supported by the GdR GeoSto 3477, by the ANR Econet (ANR-18-CE02-0010) and by the Chair "Modlisation Mathmatique et Biodiversit" of Veolia Environnement-Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X.

1 Introduction

A way to infer a random structure such as the graph of a social network and discover its properties is to explore it with random walks (e.g. [25]). This mathematical idea can be put into practice to reveal hidden populations such as drug users by using referral chain sampling where each new person provides information on her/his contacts: see for example the snowball sampling [13] or the 'respondent-driven sampling' (RDS) introduced by Heckathorn [14] (see also the PhD thesis of the second author [30]). These methods were first used to estimate the size of the hidden population or to infer population means, under the assumption that subjects' network degree determines their probability of being sampled, see Volz and Heckathorn [31] (see also [20]). Because the inclusion probability of a subject is complicated to compute, due to the dependencies associated with the graph and the fact that the sampling should

[*]Tran Viet Chi, LAMA, Univ Gustave Eiffel, UPEM, Univ Paris Est Creteil, CNRS, F-77447, Marne-la-Valle, France; E-mail: chi.tran@u-pem.fr

[†]Vo Thi Phuong Thuy, Univ. Paris 13, CNRS, UMR 7539 - LAGA, 99 avenue J.-B. Clément, F-93430 Villetaneuse, France; E-mail: phuongthuywz@gmail.com

be in practice without replacement, an important numerical literature on the subject has followed (see e.g. [11, 12, 24]). Gile [10] proposed an improved estimator for population means taking into account the without replacement sampling, and Rohe established critical threshold for the design effects [26]. Because of privacy restrictions, the social-network information is usually only a tree, as each interviewee has been 'invited' into the survey by a previously interviewed subject. Crawford, Wu and Heimer [8] use a Bayesian approach to integrate over the missing edge between recruited individuals.
It appears that the information gathered in chain-referral surveys can also be used in estimating the social network itself or at least properties associated with its topology. Recent surveys allow to gather connectivity information for recruited members: see for example the Rolls et al. [27] and Jauffret-Roustide et al. [29]. Interviewees are asked for a description of their contacts, and for a first name or a nickname. This information allows to reconstruct partially the social network and obtain a subgraph that is not a tree. It is then natural to wonder how much information on the total graph can be recovered from the observation of the subgraph obtained by the chain-referral sampling. Of course, biases have been emphasized as individuals of high degrees (hubs) are sampled with higher probability and 'common profiles' are much more likely to be discovered (e.g. [18]). This motivates the present paper. To fix the framework of study, we consider a particular class of random graphs, namely the Stochastic Block Models (SBM) that are popular models for social networks (see [15] and the review [1]). For this parametric model, inferring the distribution of the random graph boils down to a finite dimensional parameter estimation. Also, for simplification, we consider here a model of random walk on the continuous version of the SBM graph, namely the SBM graphon that is introduced in the next paragraph. Two estimations strategies are considered in this paper. First, we establish the likelihood of a random walk exploring this structure, and which accounts for the sampling biases. Two cases are classically considered, depending on whether the types of the visited nodes are observed or not. Even in the case of a complete observation, the maximum likelihood estimator has no explicit form. When the types of the vertices are unobserved, we adapt the Stochastic Approximation Expectation-Maximization algorithm (SAEM) as introduced in [6, 19]. Second, we propose a new estimation using new theoretical probabilistic results by Athreya and Roellin [3] who compute an exact formula for the bias. We provide a consistent estimator in the case of complete observations and a de-biasing strategy for the usual maximum likelihood estimator of Daudin et al. [9] in the case where the types of the explored nodes are unknown.

We consider as a toy model a Stochastic Block Model graphon with Q classes. Graphons, considered here as symmetric integrable functions from $[0,1]^2$ to $\mathbb{R}$, can be seen as limit of dense graphs (see e.g. [21]). Recall that SBM graphs are a generalization of Erdös-Rnyi graphs, where each node i is characterized by a type, $Z_i \in \{1,\dots,Q\}$, with Q the number of different possible values. The random variable (r.v.) Z_i are assumed independent and identically distributed (i.i.d.) with $\mathbb{P}(Z_i = q) = \alpha_q > 0$. The graph is non oriented. Each pair of nodes $\{i,j\}$ is connected independently with a probability $\pi_{Z_i,Z_j} \in (0,1)$ that depends only on the types. When the number of vertices of the graph tends to infinity, it is known that the dense graph converges to a limiting continuous object called graphon, see e.g. [4, 5, 21]. Let us recall the definition of the SBM graphon.

For the sequel, we introduce the partition of $[0,1]$ defined by

$$I_q = \big[\sum_{k=1}^{q-1}\alpha_k, \sum_{k=1}^{q}\alpha_k\big), \qquad q \in \{1,\dots Q\}. \tag{1}$$

The SBM graphon is the function from $[0,1]^2$ to $[0,1]$ defined as follows:

$$\kappa(x,y) = \sum_{q=1}^{Q}\sum_{r=1}^{Q} \pi_{qr}\ \mathbf{1}_{I_q}(x)\mathbf{1}_{I_r}(y). \tag{2}$$

Heuristically, we can see $[0,1]$ as a continuum of vertices, and κ is the limit of the adjacency matrix of the graph in the sense that $\kappa(x,y)$ measures the probability of connection between x and y.

We consider a random walk on the graphon κ, i.e. the process $X = (X_m)_{m\geq 1}$ with values in $[0,1]$ and transition kernel:

$$K_\kappa(x,dy) = \frac{\kappa(x,y)dy}{\int_0^1 \kappa(x,v)dv} = \frac{\sum_{q=1}^Q \Big(\sum_{r=1}^Q \pi_{qr}\ \mathbf{1}_{I_r}(y)\Big)\mathbf{1}_{I_q}(x)\ dy}{\sum_{q=1}^Q \Big(\sum_{r=1}^Q \pi_{qr}\alpha_r\Big)\mathbf{1}_{I_q}(x)}. \tag{3}$$

This random walk is the analogous of the classical random walk on a graph that jumps from a vertex to one of its neighbouring vertices chosen uniformly at random.
From the exploration of this random walk, we can construct a subgraph of the 'nodes' visited. Assume that we observe n steps of the random walk, i.e. $X^{(n)} = (X_1, \dots, X_n)$. The associated path (up to its nth step) is a subgraph (chain) $H_n = (V_n, E_n)$ with vertices $V_n = \{X_1, \dots X_n\}$ and edges $E_n = \cup_{m=1}^{n-1}\{X_m, X_{m+1}\}$. This chain is completed by sampling independently edges between vertices that are not already connected with probability according to their types. Following the notation of Athreya and Roellin [3], we denote by $G_n := G(X^{(n)}, \kappa, H_n)$ the random graph, which is completed from H_n w.r.t. the graphon κ:

Definition 1. *The vertices of $G_n = G(X^{(n)}, \kappa, H_n)$ are the nodes $X^{(n)}$, and the edges are as follows. Let i and j be two vertices.*

- *If there is an edge between i and j in H_n, $i \sim_{H_n} j$ then there is also an edge between these nodes in G_n: $i \sim_{G_n} j$.*
- *If there is no edge between i and j in H_n, we connect i and j in G_n with probability $\kappa(X_i, X_j)$.*

This subgraph G_n is the RDS graph. We assume that this is the model generating our data and that the observation corresponds to a realization of G_n. In the sequel, we denote the parameter of the SBM by $\theta = (\alpha_1, \dots \alpha_Q, \pi_{qr}; q, r \in \{1, \dots Q\})$. Our purpose is to estimate θ using the subgraph G_n. In the literature, the estimation of SBM graphs has been extensively studied, but often in a framework where the number of nodes is known. In particular, variational EM approaches have been used in many cases where types are unknown, see [9, 28, 22]. The estimation of SBM graphs, when the total population size is unknown and when we only have a subgraph obtained by a chain-referral method, is not studied to our knowledge. We develop in this paper two approaches that we compare in a final numerical section (Section 5).

First, it is possible to write the likelihood of G_n. Here, because graph is explored through an RDS random walk, our likelihood differs from the likelihoods in these papers: it accounts both on the transitions of the random walk and on the connectivity of vertices given their types. We study in Section 3 the maximum likelihood estimator (MLE) in our setting for both cases, when the nodes types are observed or not. Even when the observation is complete, the maximum likelihood estimator does not have an explicit form. When the types are unknown, we adapt to our likelihood the variational EM approach of [9].
The second approach developed in Section 4 is inspired by the recent work of Athreya and Roellin [3]. These authors showed that when we observe the random walk sufficiently long ($n \to +\infty$), the sequence of graphs $(G(H_n, \kappa))_{n\geq 1}$ converges to a biased graphon of κ. Based on their probabilistic result, a natural estimator of the biased graphon turns out to be the MLE in the 'classical' case studied by [9]. Based on this estimator that is not consistent in our case, we propose a new consistent estimator of κ.

2 Probabilistic setting

In this section, we give some important properties of the RDS Markov chain $X^{(n)}$, in particular on its long term behaviour. Then we explain the biases that appear when estimating the graphon κ from the RDS subgraph G_n.

2.1 Exploration by a random walk

Assumption 1. *In all the paper, we assume that κ is the graphon of an SBM graph (see (2)) and that κ is connected, i.e. that for all measurable subset $A \subset [0,1]$ such that $|A| \in (0,1)$,*

$$\int_A \int_{A^c} \kappa(x,y)dx\ dy > 0.$$

Proposition 1. *Under Assumptions 1, the random walk $X = (X_n)_{n\geq 1}$ admits a unique invariant probability measure*

$$m(dx) = \frac{\int_0^1 \kappa(x,v)dv}{\int_0^1 \int_0^1 \kappa(u,v)du\ dv}\ dx = \frac{\sum_{q=1}^Q \Big(\sum_{r=1}^Q \pi_{qr}\alpha_r\Big)\mathbf{1}_{I_q}(x)\ dx}{\sum_{q=1}^Q \sum_{r=1}^Q \pi_{qr}\ \alpha_q\ \alpha_r}. \tag{4}$$

The general proof is given in [3, Prop. 4.1] but for the case of SBM graphons, the result is easy to prove.

From expression (4), we see that the stationary measure $m(dx)$ put more weight on the intervals I_q corresponding to frequent types (large α_q) or hubs ($\pi_{q\cdot}$ close to one). Because $m(dx)$ is not the uniform measure, we expect biases in how the graphon κ is discovered by G_n.

2.2 Convergence of dense graphs

We are interested in the case where $n \to +\infty$. Then, the (dense) RDS graph G_n might converge to a graphon, and it is natural to compare the possible limit to the graphon κ on which the random walk moves. Let us recall briefly some topological facts. We refer the interested reader to [21].

Let us give first some notations. For integers n and $k \leq n$, $[\![1,n]\!] = \{1, 2\cdots n\}$ and $(n)_k = n(n-1)\cdots(n-k+1)$. For a graph G, $E(G)$ denotes the edges of G and $i \sim_G j$ means that $\{i,j\} \in E(G)$. We can define the subgraph F density in G by:

$$t(F,G) = \frac{\#\{\text{injections from } F \text{ to } G\}}{(n)_k} = \frac{1}{(n)_k} \sum_{(i_1,\cdots i_k)\in[\![1,n]\!]} \prod_{\{\ell,\ell'\}\in E(F)} \mathbf{1}_{i_\ell \sim_G i_{\ell'}} \tag{5}$$

where $\sum_{(i_1,\cdots i_k)\in[\![1,n]\!]}$ is a sum ranging over all vectors $(i_1,\cdots i_k)$ with mutually different coordinates in $[\![1,n]\!]$. This notion of subgraph density can be generalized to a graphon κ by:

$$t(F,\kappa) = \int_{[0,1]^k} \prod_{\{\ell,\ell'\}\in E(F)} \kappa(x_\ell, x_{\ell'})dx_1\cdots dx_k. \tag{6}$$

Let $\mathcal{F}$ denote the class of isomorphism classes on finite graphs and let $(F_i)_i \geq 1$ be a particular enumeration of $\mathcal{F}$. Then, the distance of two graphs G and G' is:

$$d_{\text{sub}}(G,G') = \sum_{i\geq 0} \frac{1}{2^i} \big|t(F_i,G) - t(F_i,G')\big| \tag{7}$$

The convergence of the large graphs to graphons can be expressed with this distance [21, Chapter 11].

2.3 Biases in the discovery of κ

Let us denote by Γ the cumulative distribution function of $\pi(dx)$:

$$\Gamma(x) = \frac{\sum_{q=1}^Q \sum_{r=1}^Q (\pi_{qr}\alpha_r)\Big[\min\big(\alpha_q,\ x - \sum_{k=1}^{q-1}\alpha_k\big)\Big]_+}{\sum_{q=1}^Q \sum_{r=1}^Q \pi_{qr}\ \alpha_q\ \alpha_r} \tag{8}$$

Athreya and Roellin [3] have proved that the graphon discovered by the RDS is biased:

Proposition 2 (Corollary 2.2 [3]). *We have under Assumptions 1 that:*

$$\lim_{n\to+\infty} d_{\mathrm{sub}}\big(G_n, \kappa_{\Gamma^{-1}}\big) = 0,$$

where the generalised inverse of Γ *is*

$$\Gamma^{-1}(v) = \inf\{u \in [0,1] : \Gamma(u) \geq v\},$$

and where for all $x, y \in [0,1]$,

$$\kappa_{\Gamma^{-1}}(x,y) = \kappa\big(\Gamma^{-1}(x), \Gamma^{-1}(y)\big). \tag{9}$$

This proposition, that is true not only for SBM graphons but also in more general cases, as developed in [3], says that the topology of the subgraph discovered by the RDS is biased compared with the true underlying structure (κ) because the random walk visits more likely the nodes with high degrees (hubs) and the frequent types.

Example 1. *When* $Q = 2$, *the graphon is given:*

$$\kappa(x,y) = \begin{cases} \pi_{11}, & 0 \leq x, y \leq \alpha; \\ \pi_{12}, & \alpha < x \leq 1 \quad or \quad \alpha < y \leq 1; \\ \pi_{22}, & otherwise. \end{cases}$$

This function is represented in Fig. 1 The invariant probability measure is:

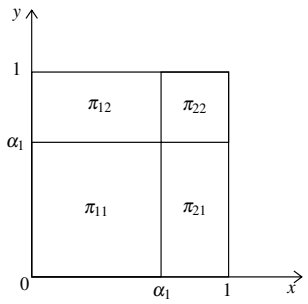

Figure 1: *F*unction $\kappa(x,y)$ for an SBM graphon with $Q = 2$ classes.

$$m(dx) = \frac{(\pi_{11}\alpha + \pi_{12}(1-\alpha))\mathbf{1}_{x\in[0,\alpha]}(x) + (\pi_{12}\alpha + \pi_{22}(1-\alpha))\mathbf{1}_{x\in(\alpha,1]}(x)}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2}dx.$$

Then the cumulative distribution of m *is:*

$$\begin{aligned}\Gamma(x) = &\frac{(\pi_{11}\alpha + \pi_{12}(1-\alpha))x}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2}\mathbf{1}_{x<\alpha} \\ &+ \left[\frac{\pi_{11}\alpha^2 + \pi_{12}(1-\alpha)\alpha}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2} + \frac{(\pi_{12}\alpha + \pi_{22}(1-\alpha))(x-\alpha)}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2}\right]\mathbf{1}_{x\geq\alpha}.\end{aligned}$$

The biased graphon $\kappa_{\Gamma^{-1}}$ *is here:*

$$\kappa_{\Gamma^{-1}}(x,y) := \begin{cases} \pi_{11}, & if \quad (x,y) \in [0,\Gamma(\alpha)] \times [0,\Gamma(\alpha)]; \\ \pi_{22}, & if \quad (x,y) \in [\Gamma(\alpha),1] \times [\Gamma(\alpha),1]; \\ \pi_{12}, & otherwise; \end{cases} \tag{10}$$

with

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

It can be seen that $\Gamma(\alpha) = \alpha$ *when* $(1-\alpha)(\pi_{12} - \pi_{22}) = \alpha(\pi_{12} - \pi_{11})$. *This is satisfied for example when* $\pi_{11} = \pi_{12} = \pi_{22}$ *(Erdös-Rnyi) or when* $\alpha = 1/2$ *and* $\pi_{11} = \pi_{22}$ *(both types are symmetric).*

2.4 Empirical cumulative distribution

As seen in the previous paragraph, the bias linked with the discovery of the graphon κ by the RDS subgraph G_n is expressed in term of the cumulative distribution Γ of the stationary distribution m of $X^{(n)}$. In the sequel, the empirical cumulative distribution of m will be useful and we recall here some facts:

$$\Gamma_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} \qquad \text{and} \qquad \Gamma_n^{-1}(y) = \inf\big\{x \in [0,1] \ : \ \Gamma_n(x) \geq y\big\}. \tag{12}$$

Lemma 3. *Γ_n and Γ_n^{-1} converge a.s. uniformly to Γ and Γ^{-1} respectively.*

Proof. The almost sure pointwise convergence of Γ_n to Γ is a consequence of the ergodic theorem. Then, the a.s. uniform convergence is obtain by the Glivenko-Cantelli theorem.
Let us prove the uniform convergence of Γ_n^{-1} to Γ^{-1}. Because all the α_q's are positive, Γ is a nondecreasing and piecewise affine bijection and the inverse bijection Γ^{-1} is also nondecreasing and piecewise affine. Let $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ sufficiently large so that for all $n \geq n_0$, $\|\Gamma_n - \Gamma\|_\infty \leq \varepsilon$. Let $y \in [0,1]$. For $n \geq n_0$,

$$\big|\Gamma_n^{-1}(y) - \Gamma^{-1}(y)\big| \leq C\big|\Gamma(\Gamma_n^{-1}(y)) - y\big|.$$

Because the jumps of Γ_n are a.s. of size $1/n$, we necessarily have that $y - \varepsilon \leq \Gamma(\Gamma_n^{-1}(y)) \leq y + \varepsilon + \frac{1}{n}$. Thus,

$$\big|\Gamma_n^{-1}(y) - \Gamma^{-1}(y)\big| \leq C\big(\frac{1}{n} + \varepsilon\big),$$

which proves the uniform convergence of Γ_n^{-1} to Γ^{-1}. □

3 Likelihood estimation

In this section, we write the likelihood of G_n and compute the MLE of the parameters θ. Here our likelihood is specific to the RDS exploration. The MLE does not have an explicit formula and we explain how to compute it numerically. Then, we study the case where the types Z_i of the nodes are unobserved. Notice that the estimation in this Section 3 makes only use of the connectivity information carried by the random variables Y_{ij}. The estimators here do not depend on the positions X_i. The types Z_i may be known or unobserved.

Let us introduce some notations. We define by N_n^q, $q \in \{1, ..., Q\}$ the number of vertices of type q sampled by the Markov chain. For $q, r \in \{1, ..., Q\}$ we also define by:

$$N_n^{q \leftrightarrow r} = \text{Card}\big\{(i,j) \quad | \quad i,j \in X^{(n)},\ Z_i = q,\ Z_j = r,\ Y_{i,j} = 1\big\};$$
$$N_n^{q \nleftrightarrow r} = \text{Card}\big\{(i,j) \quad | \quad i,j \in X^{(n)},\ Z_i = q,\ Z_j = r,\ Y_{i,j} = 0\big\}$$

the number of couples of types (q,r) that are connected (resp. not connected).

3.1 Complete observations

Assume that we observe a subset of explored nodes $X^{(n)} = (X_1, \dots X_n) \subset [0,1]^n$ discovered by the RDS, with their classes and connections: $(Z_i, Y_{ij}; X_i, X_j \in X^{(n)}, i \neq j) \in \{1, \cdots Q\}^n \times \{0,1\}^{n(n-1)}$.

Proposition 4. *The complete likelihood of the observations is*

$$\mathcal{L}(Z, Y, X, \theta) = \prod_{q=1}^{Q} \left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}} (1-\pi_{qq})^{N_n^q(N_n^q-1)/2} \prod_{q \neq r} \left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}} (1-\pi_{qr})^{N_n^q N_n^r}$$
$$\times \prod_{q=1}^{Q} \frac{\alpha_q^{N_n^q}}{(\sum_{q'=1}^{Q} \pi_{qq'}\alpha_{q'})^{N_n^q - \mathbf{1}_{Z_n = q}}}. \tag{13}$$

Proof. We have that

$$\mathcal{L}(Z_i, Y_{ij}; i,j \in X^{(n)}; \theta) = \alpha_{Z_1} \prod_{m=1}^{n-1} \frac{\pi_{Z_m Z_{m+1}} \alpha_{Z_{m+1}}}{\sum_{q=1}^{Q} \pi_{Z_m q} \alpha_q} \times \prod_{\substack{i,j: X_i, X_j \in X^{(n)}, \\ \{X_i, X_j\} \notin H_n}} \pi_{Z_i Z_j}^{Y_{i,j}} (1 - \pi_{Z_i Z_j})^{(1 - Y_{i,j})},$$

where the first product corresponds to the likelihood of the types sampled along the Markov chain, and the second product corresponds to the likelihood of edges between vertices that are not visited successively by the Markov chain. Thus:

$$\mathcal{L}(Z_i, Y_{ij}; i,j \in X^{(n)}; \theta) = \frac{\prod_{i=1}^{n} \alpha_{Z_i}}{\prod_{i=1}^{n-1} \sum_{q=1}^{Q} \pi_{Z_i q} \alpha_q} \times \prod_{\substack{i,j \in [\![1,n]\!] \\ X_i, X_j \in X^{(n)}}} b(Y_{ij}, \pi_{Z_i Z_j}), \tag{14}$$

where $b(Y_{ij}, \pi_{Z_i Z_j}) = \pi_{Z_i Z_j}^{Y_{ij}} (1 - \pi_{Z_i Z_j})^{1 - Y_{ij}}$. Finally, rewritting the above likelihood using N_n^q, $N_n^{q \leftrightarrow r}$, we obtain (13). □

Proposition 5. *The MLE $\widehat{\theta} = (\widehat{\alpha}, \widehat{\pi})$ is the solution of the following system of equations:*

$$\sum_{m=1}^{n} \frac{\mathbf{1}_{Z_m = q}}{\alpha_q} - \sum_{m=1}^{n-1} \frac{\pi_{Z_m q}}{\sum_{q'=1}^{Q} \pi_{Z_m q'} \alpha'_q} = 0; \tag{15}$$

$$\sum_{m=1}^{n-1} \left(\frac{\mathbf{1}_{(Z_m, Z_{m+1}) = (qr)}}{\pi_{qr}} - \frac{\alpha_r \mathbf{1}_{Z_m = q}}{\sum_{q'=1}^{Q} \pi_{q q'} \alpha_{q'}} \right) + \sum_{\substack{i,j: X_i, X_j \in X^{(n)} \\ \{X_i, X_j\} \notin H_n}} \left(\frac{Y_{i,j}}{\pi_{qr}} - \frac{1 - Y_{i,j}}{1 - \pi_{qr}} \right) \mathbf{1}_{(Z_i, Z_j) = (qr)} = 0. \tag{16}$$

Proof. The log likelihood of the observations is:

$$\begin{aligned} \log \mathcal{L} = & \sum_{q=1}^{Q} \Big(N_n^q \log \alpha_q - (N_n^q - \mathbf{1}_{Z_n = q}) \log \big(\sum_{q'=1}^{Q} \pi_{q q'} \alpha_{q'} \big) \Big) \\ & + \sum_{q=1}^{Q} \Big(N_n^{q \leftrightarrow q} \log \big(\frac{\pi_{qq}}{1 - \pi_{qq}} \big) + \frac{N_n^q (N_n^q - 1)}{2} \log(1 - \pi_{qq}) \Big) \\ & + \sum_{q=1}^{Q} \Big(\sum_{r \neq q} N_n^{q \leftrightarrow r} \log \big(\frac{\pi_{qr}}{1 - \pi_{qr}} \big) + N_n^q N_n^r \log(1 - \pi_{qr}) \Big) \end{aligned}$$

When we take the derivative of function $\log \mathcal{L}$ with respect to the parameters, we obtain:

$$\frac{N_n^q}{\alpha_q} - \sum_{p=1}^{Q} \frac{N_n^p \pi_{pq}}{\sum_{q'=1}^{Q} \pi_{p q'} \alpha_{q'}} = 0; \tag{17}$$

$$\frac{N_n^{q \leftrightarrow r}}{\pi_{qr}} - \frac{N_n^{q \nleftrightarrow r}}{1 - \pi_{qr}} - N_n^q \frac{\alpha_r}{\sum_{q'=1}^{Q} \pi_{q q'} \alpha_{q'}} = 0. \tag{18}$$

The identifiability of the model is a result by Allman et al. [2]. Since the likelihood is differentiable, there exists a sequence of solutions of (17) that converge to the true parameter θ. □

Remark 6. *Notice that in absence of bias, the classical likelihood, as obtained in Daudin et al. [9] is:*

$$\begin{aligned}\mathcal{L}^{\text{class}}(Z_i, Y_{ij}; \theta) &= \prod_{i=1}^{n} \alpha_{Z_i} \times \prod_{i,j \in (X_n)} b(Y_{ij}, \pi_{Z_i Z_j}) \\ &= \prod_{q=1}^{Q} \alpha_q^{N_n^q} \times \prod_{q=1}^{Q} \left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}} (1-\pi_{qq})^{N_n^q(N_n^q-1)/2} \prod_{q\neq r} \left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}} (1-\pi_{qr})^{N_n^q N_n^r}.\end{aligned} \tag{19}$$

The difference between (19) *and* (14) *is the first product which corresponds of the likelihood of the node types. In the classical case, these types are chosen independently whereas here they are discovered by the successive states of the Markov chain. In this classical case, the MLE has an explicit formula:*

$$\widehat{\alpha}_q^{\text{class}} = \frac{N_n^q}{n}, \qquad \widehat{\pi}_{qr}^{\text{class}} = \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r}, \qquad \widehat{\pi}_{qq}^{\text{class}} = \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q-1)}. \tag{20}$$

Here, for the likelihood (13), the MLE which solves (15) is not explicit any more. In Section 3.1.1, we detail in the case of two classes ($Q = 2$) the computation of the MLE.

3.1.1 Case where $Q = 2$

Let us solve the likelihood equations when $Q = 2$. The parameter is then $\theta = (\alpha, \pi_{11}, \pi_{12}, \pi_{22})$. Define $\widehat{\theta} = (\widehat{\alpha}, \widehat{\pi_{11}}, \widehat{\pi_{12}}, \widehat{\pi_{22}})$ the estimator of θ. Then the estimators $\widehat{\theta}$ is the solution of

$$\frac{N_n^1}{\widehat{\alpha}} - \frac{N_n^1\widehat{\pi_{11}}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} - \frac{N_n^2\widehat{\pi_{12}}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{21}$$

$$\frac{N_n^2}{1-\widehat{\alpha}} - \frac{N_n^1\widehat{\pi_{12}}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} - \frac{N_n^2\widehat{\pi_{22}}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{22}$$

$$\frac{N_n^{1\leftrightarrow 1}}{\widehat{\pi_{11}}} - \frac{N_n^{1\nleftrightarrow 1}}{1-\widehat{\pi_{11}}} - \frac{N_n^1\widehat{\alpha}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} = 0; \tag{23}$$

$$\frac{N_n^{1\leftrightarrow 2}}{\widehat{\pi_{12}}} - \frac{N_n^{1\nleftrightarrow 2}}{1-\widehat{\pi_{12}}} - \frac{N_n^1(1-\widehat{\alpha})}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} = 0; \tag{24}$$

$$\frac{N_n^{2\leftrightarrow 1}}{\widehat{\pi_{12}}} - \frac{N_n^{1\nleftrightarrow 2}}{1-\widehat{\pi_{12}}} - \frac{N_n^2\widehat{\alpha}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{25}$$

$$\frac{N_n^{2\leftrightarrow 2}}{\widehat{\pi_{22}}} - \frac{N_n^{2\nleftrightarrow 2}}{1-\widehat{\pi_{22}}} - \frac{N_n^2(1-\widehat{\alpha})}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0. \tag{26}$$

Proposition 7. *The MLE* $\widehat{\theta} = (\widehat{\alpha}, \widehat{\pi_{11}}, \widehat{\pi_{12}}, \widehat{\pi_{22}})$ *can be expressed as a function of* $\widehat{\pi}_{12}$*:*

$$\widehat{\pi_{11}} = \frac{(N_n^{1\leftrightarrow 1} + N_n^{1\leftrightarrow 2} - N_n^1) - (N_n^1 N_n^2 - N_n^1 + N_n^{1\leftrightarrow 1})\widehat{\pi_{12}}}{(\frac{N_n^1(N_n^1-1)}{2} - N_n^1 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^1(N_n^1-1)}{2} + N_n^1 N_n^2 - N_n^1)\widehat{\pi_{12}}}, \tag{27}$$

$$\widehat{\pi_{22}} = \frac{(N_n^{2\leftrightarrow 2} + N_n^{1\leftrightarrow 2} - N_n^2) - (N_n^{2\leftrightarrow 2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}}{(\frac{N_n^2(N_n^2-1)}{2} - N_n^2 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^2(N_n^2-1)}{2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}}, \tag{28}$$

$$\widehat{\alpha} = \frac{\widehat{\beta}}{1+\widehat{\beta}}, \tag{29}$$

with

$$\widehat{\beta} = \frac{(N_n^1 - N_n^2)\widehat{\pi_{12}} + \sqrt{(N_n^1 - N_n^2)^2 \widehat{\pi_{12}}^2 + 4N_n^1 N_n^2 \widehat{\pi_{11}}\widehat{\pi_{22}}}}{2N_n^2\widehat{\pi_{11}}}, \tag{30}$$

and where $\widehat{\pi}_{12}$ is one of the root of

$$\begin{aligned}\widehat{\pi_{12}}^2 =& \frac{(N_n^{1\leftrightarrow 1}+N_n^{1\leftrightarrow 2}-N_n^1)-(N_n^1N_n^2-N_n^1+N_n^{1\leftrightarrow 1})\widehat{\pi_{12}}}{(\frac{N_n^1(N_n^1-1)}{2}-N_n^1+N_n^{1\leftrightarrow 2})-(\frac{N_n^1(N_n^1-1)}{2}+N_n^1N_n^2-N_n^1)\widehat{\pi_{12}}}\\ &\times\frac{(N_n^{2\leftrightarrow 2}+N_n^{1\leftrightarrow 2}-N_n^2)-(N_n^{2\leftrightarrow 2}+N_n^1N_n^2-N_n^2)\widehat{\pi_{12}}}{(\frac{N_n^2(N_n^2-1)}{2}-N_n^2+N_n^{1\leftrightarrow 2})-(\frac{N_n^2(N_n^2-1)}{2}+N_n^1N_n^2-N_n^2)\widehat{\pi_{12}}}\\ &\times\frac{(N_n^{1\leftrightarrow 2}-N_n^1N_n^2\widehat{\pi_{12}})^2}{[(N_n^{1\leftrightarrow 2}-N_n^1)-(N_n^1N_n^2-N_n^1)\widehat{\pi_{12}}][(N_n^{1\leftrightarrow 2}-N_n^2)-(N_n^1N_n^2-N_n^2)\widehat{\pi_{12}}]}.\end{aligned} \tag{31}$$

Proof. Multiply (23) by $\widehat{\pi_{11}}$ and (24) by $\widehat{\pi_{12}}$, and sum them up, we have

$$N_n^{1\leftrightarrows 1}\frac{\widehat{\pi_{11}}}{1-\widehat{\pi_{11}}}+N_n^{1\leftrightarrows 2}\frac{\widehat{\pi_{12}}}{1-\widehat{\pi_{12}}}=N_n^{1\leftrightarrow 1}+N_n^{1\leftrightarrow 2}-N_n^1. \tag{32}$$

Similarly, from equations (25) and (26), we deduce

$$N_n^{1\leftrightarrows 2}\frac{\widehat{\pi_{12}}}{1-\widehat{\pi_{12}}}+N_n^{2\leftrightarrows 2}\frac{\widehat{\pi_{22}}}{1-\widehat{\pi_{22}}}=N_n^{1\leftrightarrow 2}+N_n^{2\leftrightarrow 2}-N_n^2. \tag{33}$$

Also, the system of equations (23)-(26) gives

$$\left(\frac{N_n^{1\leftrightarrow 1}}{\widehat{\pi_{11}}}-\frac{N_n^{1\leftrightarrows 1}}{1-\widehat{\pi_{11}}}\right)\left(\frac{N_n^{2\leftrightarrow 2}}{\widehat{\pi_{22}}}-\frac{N_n^{2\leftrightarrows 2}}{1-\widehat{\pi_{22}}}\right)=\left(\frac{N_n^{1\leftrightarrow 2}}{\widehat{\pi_{12}}}-\frac{N_n^{1\leftrightarrows 2}}{1-\widehat{\pi_{12}}}\right)^2. \tag{34}$$

Notice that $N^{1\leftrightarrows 2}+N^{1\leftrightarrow 2}=N_n^1N_n^2$, $N^{1\leftrightarrows 1}+N^{1\leftrightarrow 1}=\frac{N_n^1(N_n^1-1)}{2}$ and $N^{2\leftrightarrows 2}+N^{2\leftrightarrow 2}=\frac{N_n^2(N_n^2-1)}{2}$ and we consider $\widehat{\pi_{12}}$ as a parameter. Solving the system (32)-(33) for $\widehat{\pi_{11}}$, $\widehat{\pi_{22}}$ provides the two first equations of (27). Using this, (34) is equivalent to:

$$\frac{(N_n^{1\leftrightarrow 2}-N_n^1N_n^2\widehat{\pi_{12}})^2}{[(N_n^{1\leftrightarrow 2}-N_n^1)-(N_n^1N_n^2-N_n^1)\widehat{\pi_{12}}][(N_n^{1\leftrightarrow 2}-N_n^2)-(N_n^1N_n^2-N_n^2)\widehat{\pi_{12}}]}\frac{\widehat{\pi_{11}}\widehat{\pi_{22}}}{\widehat{\pi_{12}}^2}=1. \tag{35}$$

This gives the (31).

For the estimator of α, let us denote $\beta:=\frac{\alpha}{(1-\alpha)}$. Then equations (21) and (22) are the same and equivalent to

$$\frac{N_n^1}{\widehat{\pi_{11}}\hat{\beta}+\widehat{\pi_{12}}}=\frac{N_n^2\hat{\beta}}{\widehat{\pi_{12}}\hat{\beta}+\widehat{\pi_{22}}}. \tag{36}$$

The unique positive solution is $\hat{\beta}$ and provides in turn $\widehat{\alpha}$. □

Let us explain how the preceding proposition allows us to compute numerically the MLE $\widehat{\theta}$.

First: there might be several solutions of (31), see Fig. 2. For each of them, we compute the corresponding estimators of π_{11}, π_{22} and α, which allows us to obtain the corresponding likelihood of the observations. We choose the set of estimators that provides the best likelihood for our observations.

Second: to solve numerically the equation (31), we use the bisection method with the following constraints:

- The equation (31) has 4 excluded values that make the denominator zero:

$$\bar{\pi}_{12}^1=\frac{N_n^{1\leftrightarrow 2}-N_n^2}{N_n^1N_n^2-N_n^2}\qquad\bar{\pi}_{12}^2=\frac{N_n^{1\leftrightarrow 2}-N_n^1}{N_n^1N_n^2-N_n^1} \tag{37}$$

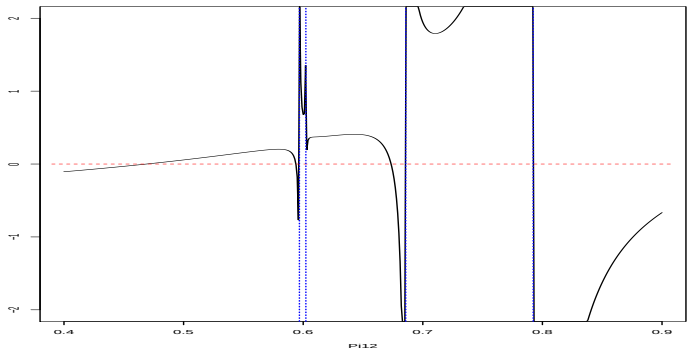

Figure 2: *E*quation (31) can be rewritten as $\phi(\pi_{12}) = 0$. The function ϕ is represented graphically on the figure above as a function of π_{12}. The vertical dotted lines correspond to the excluded values $\bar{\pi}^1_{12}, \dots \bar{\pi}^4_{12}$ given in (37).

$$\bar{\pi}^3_{12} = \frac{\frac{N^1_n(N^1_n-1)}{2} - N^1_n + N^{1\leftrightarrow 2}_n}{\frac{N^1_n(N^1_n-1)}{2} - N^1_n + N^1_n N^2_n}; \qquad \text{and} \qquad \bar{\pi}^4_{12} = \frac{\frac{N^2_n(N^2_n-1)}{2} - N^2_n + N^{1\leftrightarrow 2}_n}{\frac{N^2_n(N^2_n-1)}{2} - N^2_n + N^1_n N^2_n}$$

It is observed that $\max(\bar{\pi}^1_{12}, \bar{\pi}^2_{12}) < \min(\bar{\pi}^3_{12}, \bar{\pi}^4_{12})$. And if $N^1_n < N^2_n$, we have them ordered: $\bar{\pi}^1_{12} < \bar{\pi}^2_{12} < \bar{\pi}^3_{12} < \bar{\pi}^4_{12}$.

- All the estimators $\widehat{\pi_{11}}, \widehat{\pi_{12}}, \widehat{\pi_{22}}$ and $\widehat{\alpha}$ take values in the interval $(0, 1)$.

Taking care of the points above, we solve (31) with the bisection method on a grid that includes the excluded points $\{\bar{\pi}^i_{12}, i \in \{1, 2, 3, 4\}\}$.
For each root of (31), corresponding to a possible value of $\widehat{\pi}_{12}$, we compute the corresponding estimators of π_{11}, π_{22}.

For the numerical simulations, we refer the reader to Section 5.

3.2 Incomplete observations: SAEM Algorithm

Here, we assume that the types $(Z_i)_{i=1,\dots,n}$ are unobserved. In this case, the likelihood of the observed data $(Y_{ij};\ i, j \in [\![1, n]\!])$ is obtained by summing the complete-data likelihood (14) over all the possible values of the unobserved variables Z:

$$\mathcal{L}(Y_{ij}; i, j \in [\![1, n]\!]; \theta) = \sum_{q_1,\cdots q_n=1}^{Q} \Big[\prod_{i=1}^{n} \mathbf{1}_{Z_i=q_i} \frac{\prod_{i=1}^{n} \alpha_{q_i}}{\prod_{i=1}^{n-1} \sum_{q=1}^{Q} \pi_{q_i q} \alpha_q} \times \prod_{i,j: X_i, X_j \in X^{(n)})} b(Y_{ij}, \pi_{q_i q_j}) \Big], \qquad (38)$$

Unfortunately, this sum is not tractable and it is classical to use the Expectation-Maximization (EM) algorithm to compute the maximum likelihood. Here we follow the steps in [9] by adapting the expression to our setting with the likelihood (13).

Let us sum up the EM algorithm (see e.g. [6, 7, 19]). Given the observed data: the Markov chain $X^{(n)}$, the connections $(Y_{ij},\ i, j \in X^{(n)})$ and the number of blocks Q and the current estimator θ, and given the value $\theta^{(k-1)}$ at the $(k-1)^{th}$ iteration of the EM, on the k^{th} step, we compute the conditional expectation of the log-likelihood $\mathcal{L}(Z|X, Y, \theta^{(k)})$ given X, Y for the current fit $\theta^{(k)}$. Here there is no explicit expression for the latter likelihood because the exact distribution of Z given X, Y is unknown and this we need to approximate it numerically by using an SAEM algorithm [6, 19], proceeding as follows.

3.2.1 The SAEM algorithm

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

Step 1: Choosing the appropriate $Z^{(k)}$
- Simulate a candidate Z^c following the proposal distribution $q_{\theta^{(k-1)}}(.|Z^{(k-1)})$. The choice of proposal distribution is discussed in Section 3.2.2, where we use a variational approach.
- Calculate the acceptance probability

$$\omega(Z^{(k-1)}, Z^c) := \min\left\{1, \frac{\mathcal{L}(Z^c, Y, \theta^{(k-1)}) \cdot q_{\theta^{(k-1)}}(Z^{(k-1)}|Z^c)}{\mathcal{L}(Z^{(k-1)}, Y, \theta^{(k-1)}) \cdot q_{\theta^{(k-1)}}(Z^c|Z^{(k-1)})}\right\}; \tag{39}$$

- Accept the candidate Z^c with probability ω: $\mathbb{P}(Z^{(k)} = Z^c) = \omega$ and $\mathbb{P}(Z^{(k)} = Z^{(k-1)}) = 1 - \omega$.

Step 2: Stochastic approximation Update the quantity

$$\mathcal{Q}^{(k)}(\theta) = \mathcal{Q}^{(k-1)}(\theta) + s_k\left(\log \mathcal{L}(Z_i^{(k)}, Y_{ij}, \theta) - \mathcal{Q}^{(k-1)}(\theta)\right), \tag{40}$$

with the initialization $\mathcal{Q}^{(0)}(\theta) := \mathbb{E}[\log \mathcal{L}(Z, Y, \theta^{(0)})]$ and $(s_k)_{k\in\mathbb{N}}$ is a positive decreasing step sizes sequence satisfying $\sum_{k=1}^{\infty} s_k = \infty$ and $\sum_{k=1}^{\infty} s_k^2 < \infty$.

Step 3: Maximization Choose $\theta^{(k)}$ to be the value of θ that maximizes $\mathcal{Q}^{(k)}$

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{Q}^{(k)}(\theta). \tag{41}$$

Kuhn and Lavielle studied the convergence of the sequence $\theta^{(k)}$ in [19].

3.2.2 Variational approach

For the proposal distribution $q_{\theta^{(k-1)}}(. \mid Z^{(k-1)})$ of $Z^{(k)}$, we follow Daudin et al. [9], who use a variational approach. Let us recall the main idea of this approach. The general strategy has been described in Jordan et al. [17] or Jaakkola [16].

Recall the likelihood $\mathcal{L}(Y, \theta)$ of the incomplete data (38). The idea of the variational approach is to replace the likelihood by a lower bound:

$$\mathcal{J}(R_{Y,\theta}) = \mathcal{L}(Y, \theta) - \mathrm{KL}(R_{Y,\theta}(Z), \mathcal{L}(Z|Y, \theta)), \tag{42}$$

where $\mathrm{KL}(\mu, \nu) := \int d\mu \log\left(\frac{d\mu}{d\nu}\right)$ is the Kullback-Leibler divergence of distributions μ and ν, and where $R_{Y,\theta}(Z)$ is an approximation of the conditional likelihood $\mathcal{L}(Z|Y, \theta)$. When $R_{Y,\theta}$ is a good-approximation of $\mathcal{L}(Z|Y, \theta)$, $\mathcal{J}(R_{Y,\theta})$ is very closed to $\mathcal{L}(Y, \theta)$.

Here, Z takes discrete values in $\{1, ..., Q\}$. Then,

$$\begin{aligned}
\mathcal{J}(R_{Y,\theta}) &= \log \mathcal{L}(Y,\theta) - \sum_{(Z_1,...,Z_n)\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log \frac{R_{Y,\theta}(Z)}{\mathcal{L}(Z|Y,\theta)} \\
&= \log \mathcal{L}(Y,\theta) - \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log R_{Y,\theta}(Z) + \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log \mathcal{L}(Z|Y,\theta) \\
&= \log \mathcal{L}(Y,\theta) - \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log R_{Y,\theta}(Z) + \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log \mathcal{L}(Z,Y,\theta) \\
&\qquad\qquad - \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log \mathcal{L}(Y,\theta) \\
&= \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log \mathcal{L}(Z,Y,\theta) - \sum_{Z\in\{1,...,Q\}^n} R_{Y,\theta}(Z) \log R_{Y,\theta}(Z).
\end{aligned}$$

Following [9], we restrict to distributions $R_{Y,\theta}$ that belong to the family of multinomial probability distributions parameterized by $\tau = (\tau_1, \cdots \tau_Q)$, as approximated conditional distribution of Z given Y and θ. If we look for the parameter τ that maximizes (42), we will hence obtain the best approximation of $\mathcal{L}(Z|Y,\theta)$ among the multinomial distributions. We will chose the latter to be the proposal distribution for Z in the Step 1 of the SAEM algorithm.

If $\mathbf{1}_{Z_i}$ follows the multinomial distribution $\mathcal{M}(1; (\tau_{i1}, ..., \tau_{iq}))$, with $\tau_{iq} = \mathbb{P}(Z_i = q|Y,\theta)$, for $i \in \{1, ..., n\}, q \in \{1, ..., Q\}$ then,

$$R_{Y,\theta}(Z) = \prod_{i=1}^{n} \tau_{i,Z_i}. \tag{43}$$

As a consequence, $\mathcal{J}(R_X)$ is rewritten as

$$\begin{aligned}
\mathcal{J}(R_{Y,\theta}) = \sum_{Z\in\{1,...Q\}^n} \left\{ \prod_{j=1}^{n} \tau_{j,Z_j} \left(\sum_{i=1}^{n} \log \alpha_{Z_i} - \sum_{i=1}^{n-1} \log(\sum_{q=1}^{Q} \pi_{Z_i q}\alpha_q) + \sum_{i,j:X_i,X_j\in X^{(n)}} \log b(Y_{ij}; \pi_{Z_i Z_j}) \right) \right\} \\
- \sum_{Z\in\{1,...Q\}^n} \prod_{j=1}^{n} \tau_{j,Z_j} \left(\sum_{i=1}^{n} \log \tau_{i,Z_i} \right).
\end{aligned}$$

We aim at calculating the parameter $\hat{\tau}$ that maximizes the lower bound of $\mathcal{L}(Y,\theta)$. Then the proposal distribution $q_{\theta^{(k-1)}}(. \mid Z^{(k-1)})$ for updating the types will be given by (43) with the parameters $\widehat{\tau}$ given in the next proposition:

Proposition 8. *Given α, π, the optimal parameter*

$$\hat{\tau} := \arg\max_{\tau} \mathcal{J}(R_{Y,\theta}), \tag{44}$$

with constraint $\sum_{q=1}^{Q} \tau_{iq} = 1, \forall i \in \{1, ..., n\}$, satisfies the fixed point relation

$$\tau_{iq} \propto \frac{\alpha_q}{\sum_{\ell=1}^{Q} \pi_{q\ell}\alpha_\ell} \prod_{i\neq j} \prod_{\ell=1}^{Q} b(Y_{ij}, \pi_{q\ell})^{\tau_{j\ell}}. \tag{45}$$

Proof. To simplify $\mathcal{J}(R_{Y,\theta})$, we have

$$\begin{aligned}\sum_{Z\in\{1,\dots Q\}^n}\prod_{i=1}^n\tau_{i,Z_i}\sum_{i=1}^n\log\alpha_{Z_i} &= \sum_{Z\in\{1,\dots Q\}^n}\sum_{i=1}^n\prod_{\substack{j=1\\j\neq i}}^n\tau_{j,Z_j}(\tau_{i,Z_i}\log\alpha_{Z_i})\\ &=\sum_{i=1}^n\sum_{Z_i=1}^Q\tau_{i,Z_i}\log\alpha_{Z_i}\sum_{Z_1,\dots,Z_n\setminus Z_i}\prod_{j\neq i}\tau_{j,Z_j}=\sum_{i=1}^n\sum_{q=1}^Q\tau_{i,q}\log\alpha_q\prod_{j\neq i}\left(\sum_{Z_j=1}^Q\tau_{j,Z_j}\right)\\ &=\sum_{i=1}^n\sum_{q=1}^Q\tau_{iq}\log\alpha_q.\end{aligned}$$

Similarly,

$$\sum_{Z\in\{1,\dots Q\}^n}\prod_{j=1}^n\tau_{j,Z_j}\left(\sum_{i=1}^n\log\tau_{i,Z_i}\right)=\sum_{i=1}^n\sum_{q=1}^Q\tau_{iq}\log\tau_{iq}.$$

In addition,

$$\begin{aligned}\sum_Z\prod_{j=1}^n\tau_{j,Z_j}\sum_{i=1}^{n-1}\log(\sum_{q=1}^Q\pi_{Z_i,q}\alpha_q) &= \sum_{i=1}^{n-1}\sum_{Z\setminus Z_i}(\prod_{j=1}^n\tau_{j,Z_j})\log\left(\sum_{q=1}^Q\pi_{Z_i,q}\alpha_q\right)\tau_{i,Z_i}\\ &=\sum_{i=1}^{n-1}\sum_{q=1}^Q\log\left(\sum_{q=1}^Q\pi_{Z_i,q}\alpha_q\right)\tau_{i,Z_i},\end{aligned}$$

and

$$\begin{aligned}\sum_Z\prod_{k=1}^n\tau_{k,Z_k}\sum_{i<j}\log b(Y_{ij},\pi_{Z_i,Z_j}) &= \sum_{i<j}\sum_{Z\setminus\{Z_i,Z_j\}}(\prod_{k\neq i,j}\tau_{k,Z_k})\sum_{Z_i,Z_j}b(Y_{ij},\pi_{Z_i,Z_j})\tau_{j,Z_i}\tau_{j,Z_j}\\ &=\sum_{i<j}\sum_{q,r=1}^Q\tau_{iq}\tau_{jr}b(Y_{ij},\pi_{qr}).\end{aligned}$$

In conclusion,

$$\mathcal{J}(R_{Y,\theta})=\sum_{i=1}^n\sum_{q=1}^Q\tau_{iq}\log\alpha_q-\sum_{i=1}^n\sum_{q=1}^Q\tau_{iq}\log\tau_{iq}+\frac{1}{2}\sum_{i\neq j}\sum_{q,r=1}^Q\tau_{iq}\tau_{jr}\log b(Y_{ij},\pi_{qr})-\sum_{i=1}^{n-1}\sum_{q=1}^Q\log\left(\sum_{r=1}^Q\pi_{qr}\alpha_r\right)\tau_{iq}. \tag{46}$$

To solve the optimization problem $\arg\max_\tau\mathcal{J}(R_{Y,\theta})$ with constraint $\sum_{q=1}^Q\tau_{iq}=1$, we use the method of Lagrange multipliers, that is finding the optimal parameters τ,λ that maximize the Lagrangian function $\mathcal{L}\text{ag}(\tau,\lambda):=\mathcal{J}(R_{Y,\theta})+\sum_{i=1}^n\lambda_i(\sum_{q=1}^Q\tau_{iq}-1)$, where λ_i is the Lagrange multiplier. Take the derivative of $\mathcal{L}ag$ w.r.t. λ_i and τ, we have

$$\begin{cases}\dfrac{\partial\mathcal{L}\text{ag}}{\partial\lambda_i}=\sum_{q=1}^Q\tau_{iq}-1\\ \dfrac{\partial\mathcal{L}\text{ag}}{\partial\tau_{iq}}=\log\alpha_q-\log\tau_{iq}+\lambda_i-1-\log\sum_{r=1}^Q\pi_{qr}\alpha_r+\dfrac{1}{2}\sum_{j\neq i}\sum_{r=1}^Q\tau_{jr}\log b(Y_{ij},\pi_{qr})+\dfrac{1}{2}\sum_{j\neq i}\sum_{r=1}^Q\tau_{jr}\log b(Y_{ji},\pi_{rq})\end{cases}$$

The optimal solution must satisfy $\dfrac{\partial \mathcal{L}\text{ag}}{\partial \lambda_i} = \dfrac{\partial \mathcal{L}\text{ag}}{\partial \tau_{iq}} = 0$, which implies

$$\log \tau_{iq} = \log \alpha_q + \lambda_i - 1 - \log \sum_{r=1}^{Q} \pi_{qr}\alpha_r + \sum_{j\neq i}\sum_{r=1}^{Q} \tau_{jr} \log b(Y_{ij}, \pi_{qr}).$$

In another word,

$$\tau_{iq} = e^{\lambda_i - 1} \frac{\alpha_q}{\sum_{r=1}^{Q} \pi_{qr}\alpha_r} \prod_{i\neq j} \prod_{r=1}^{Q} b(Y_{ij}, \pi_{qr})^{\tau_{jr}}. \tag{47}$$

□

In the case $Q = 2$, it turns out the problem is more simple since for each $i \in \{1, ..., n\}, \tau_{i1} + \tau_{i2} = 1$. For sake of simplification, we denote by τ_i instead of τ_{i1}. Hence, $\tau_{i2} = 1 - \tau_{i1} = 1 - \tau_i$.

Proposition 9. *When $Q = 2$, the variational parameter τ_i has formula:*

$$\tau_i = \frac{\phi_i(\tau)}{1 + \phi_i(\tau)} =: \Phi_i(\tau), \tag{48}$$

where

$$\phi_i(\tau) := \frac{\alpha}{1-\alpha} \frac{\alpha\pi_{21} + (1-\alpha)\pi_{22}}{\alpha\pi_{11} + (1-\alpha)\pi_{12}} \prod_{j\neq i} \left(\frac{b(Y_{ij}, \pi_{12})}{b(Y_{ij}, \pi_{22})} \right)^{1/2} \prod_{j\neq i} \left(\frac{b(Y_{ij}, \pi_{11}) b(Y_{ij}, \pi_{22})}{b(Y_{ij}, \pi_{12})^2} \right)^{\tau_j/2}. \tag{49}$$

Proof. We solve directly the optimization problem $\max_\tau \mathcal{J}(R_{Y,\theta})$ without using the Lagrangian multiplier λ. The quantity $\mathcal{J}(R_{Y,\theta})$ is written explicitly as:

$$\begin{aligned}
\mathcal{J}(R_{Y,\theta}) =& \sum_{i=1}^{n} (\tau_i \log \alpha + (1-\tau_i)\log(1-\alpha)) - \sum_{i=1}^{n} (\tau_i \log \tau_i + (1-\tau_i)\log(1-\tau_i)) \\
&+ \frac{1}{2} \sum_{i\neq j} [\tau_i\tau_j \log b(Y_{ij}, \pi_{11}) + \tau_i(1-\tau_j) \log b(Y_{ij}, \pi_{12}) + (1-\tau_i)\tau_j \log b(Y_{ij}, \pi_{21}) \\
&+ (1-\tau_i)(1-\tau_j) \log b(Y_{ij}, \pi_{22})] - \sum_{i=1}^{n-1} [\tau_i \log(\alpha\pi_{11} + (1-\alpha)\pi_{12}) + (1-\tau_i)\log(\alpha\pi_{21} + (1-\alpha)\pi_{22}].
\end{aligned}$$

Take the derivative of $\mathcal{J}(R_{Y,\theta})$ w.r.t. τ_i,

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \tau_i} &= \log \frac{\alpha}{1-\alpha} + \log \frac{1-\tau_i}{\tau_i} + \frac{1}{2} \sum_{j\neq i} \left\{ \tau_j \log \frac{b(Y_{ij}, \pi_{11})}{b(Y_{ij}, \pi_{21})} + (1-\tau_j) \log \frac{b(Y_{ij}, \pi_{12})}{b(Y_{ij}, \pi_{22})} \right\} - \log \frac{\alpha\pi_{11} + (1-\alpha)\pi_{12}}{\alpha\pi_{21} + (1-\alpha)\pi_{22}} \\
&= \log \frac{\alpha}{1-\alpha} - \log \frac{\tau_i}{1-\tau_i} - \log \frac{\alpha\pi_{11} + (1-\alpha)\pi_{12}}{\alpha\pi_{21} + (1-\alpha)\pi_{22}} + \frac{1}{2} \sum_{j\neq i} \tau_j \log \frac{b(Y_{ij}, \pi_{11}) b(Y_{ij}, \pi_{22})}{b(Y_{ij}, \pi_{12})^2} + \frac{1}{2} \sum_{j\neq i} \log \frac{b(Y_{ij}, \pi_{12})}{b(Y_{ij}, \pi_{22})}.
\end{aligned}$$

Then the variational parameter τ_i is the solution of equation $\frac{\partial \mathcal{J}}{\partial \tau_i} = 0$, which gives

$$\frac{\tau_i}{1-\tau_i} = \frac{\alpha}{1-\alpha} \times \frac{\alpha\pi_{11} + (1-\alpha)\pi_{12}}{\alpha\pi_{21} + (1-\alpha)\pi_{22}} \times \prod_{j\neq i} \left(\frac{b(Y_{ij}, \pi_{12})}{b(Y_{ij}, \pi_{22})} \right)^{1/2} \prod_{j\neq i} \left(\frac{b(Y_{ij}, \pi_{11}) b(Y_{ij}, \pi_{22})}{b(Y_{ij}, \pi_{12})^2} \right)^{\tau_j/2} = \phi_i(\tau). \tag{50}$$

It implies that $\tau_i = \frac{\phi_i(\tau)}{1+\phi_i(\tau)} = \Phi_i(\tau)$. □

3.2.3 Proposal distribution for the Step 1 of SAEM

For the sake of simplicity, we treat here the case $Q = 2$, but generalization is straightforward. Using the previous results, we can now detail the Step 1 of the SAEM algorithm. Given the parameters $\theta^{(k-1)}$, the types $Z^{(k-1)}$ and the data $(Y_{ij}; i, j \in [\![1, n]\!])$, we proceed as follows.

Step 1: We compute the parameters $\tau_i^{(k)}$ as in Proposition 9. The parameters in (49) are given by $\theta^{(k-1)}$ and the terms $b(Y_{ij}, \pi_{11}^{(k-1)})$, $b(Y_{ij}, \pi_{12}^{(k-1)})$ and $b(Y_{ij}, \pi_{22}^{(k-1)})$ are computed with the types $Z^{(k-1)}$.

Step 2: We simulate a candidate $Z^c \in \{1, 2\}^n$ for Z such that $Z_i^c - 1$ follows the law $\mathcal{B}er(\tau_i)$. Recall that the acceptance probability is

$$\mu(Z^{(k-1)}, Z^c) := \min\left\{1, \frac{\mathcal{L}_{\text{complete}}(Z^c, Y, \theta^{(k-1)}) q_{\theta^{(k-1)}}(Z^{(k-1)}|Z^c)}{\mathcal{L}_{\text{complete}}(Z^{(k-1)}, Y, \theta^{(k-1)}) q_{\theta^{(k-1)}}(Z^c|Z^{(k-1)})}\right\}, \tag{51}$$

where the complete likelihood with respect to α, π, Z, Y is

$$\mathcal{L}_{\text{complete}}(Z, Y, \theta) = \prod_{q=1}^{Q} \left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}} (1-\pi_{qq})^{N_n^q(N_n^q-1)/2} \prod_{q\neq r} \left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}} (1-\pi_{qr})^{N_n^q N_n^r} \times \prod_{q=1}^{Q} \frac{\alpha_q^{N_n^q}}{(\sum_{q'=1}^{Q} \pi_{qq'}\alpha_{q'})^{N_n^q - \mathbf{1}_{Z_n=q}}}.$$

and

$$q_{\theta^{(k-1)}}(Z^c|Z^{(k-1)}) = \prod_{i=1} \tau_i^{2-Z_i^c}(1-\tau_i)^{Z_i^c-1}; \quad q_{\theta^{(k-1)}}(Z^{(k-1)}|Z^c) = \prod_{i=1} \tau_i^{2-Z_i^{(k-1)}}(1-\tau_i)^{Z_i^{(k-1)}-1}.$$

4 Estimation via biased graphon and 'classical likelihood'

In Section 3, the MLE are computed but they do not have explicit formula in the case of RDS exploration. We thus investigate other estimators. The most natural one is the graphon estimator corresponding to (20). It turns out that we can study the asymptotic bias of this estimator thanks to the result of Athreya and Roellin [3]. Here, we need some to have the knowledge on the positions X_i of the Markov chain $X^{(n)}$. The types Z_i may be observed or not.

4.1 Complete observations

Assume in this section that we observe $X^{(n)} = (X_1, \dots X_n)$, the types $(Z_i)_{i\in\{1,\dots n\}}$ and the adjacency matrix $(Y_{ij})_{i,j\in\{1,\dots n\}}$ of the subgraph $G_n = G(X^{(n)}, \kappa, H_n)$.

It is natural that G_n converges to an SBM graphon of parameters $\gamma = (\gamma_1, ..., \gamma_Q)$ and the connection probabilities $\rho = (\rho_{qr})_{q,r\in\{1,\dots,Q\}}$:

$$\chi_\infty(x, y) = \sum_{q=1}^{Q}\sum_{r=1}^{Q} \rho_{qr}\mathbf{1}_{J_q}(x)\mathbf{1}_{J_r}(y).$$

where $J = (J_1, ..., J_Q)$ is a partition of $[0, 1]$ defined by

$$J_q = \big[\sum_{k=1}^{q-1}\gamma_k, \sum_{k=1}^{q}\gamma_k\big), \qquad q \in \{1, \dots Q\}. \tag{52}$$

The parameters γ correspond to the frequencies of the types and the parameters ρ give the probabilities of connection. Thus, a natural estimator for χ_∞ is given by:

Definition 2. *Denote by*

$$\widehat{\gamma}_q^n := \frac{N_n^q}{n}; \quad \widehat{\rho}_{qr}^n := \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r} \quad \textit{for} \quad q \neq r \quad \textit{and} \quad \widehat{\rho}_{qq}^n := \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q-1)}. \tag{53}$$

an estimator of (γ, ρ)*. The graphon associated to these estimators is defined as:*

$$\widehat{\chi}_n(x,y) := \sum_{q=1}^{Q}\sum_{r=1}^{Q} \widehat{\rho}_{qr}^n \mathbf{1}_{J_q^n}(x)\mathbf{1}_{J_r^n}(y), \tag{54}$$

with $J_q^n = \big[\sum_{k=1}^{q-1}\widehat{\gamma}_k^n, \sum_{k=1}^{q}\widehat{\gamma}_k^n\big)$*,* $q \in \{1, \dots Q\}$*.*

We notice that this estimator corresponds to the MLE in the 'classical case' (see (20)). Thanks to the Proposition 2 (due to [3]), we can study the asymptotic limit of $\widehat{\chi}_n$.

4.1.1 Limit of $\widehat{\chi}_n$

We have two empirical approximations of the limiting graphon χ_∞: the graph G_n and the graphon $\widehat{\chi}_n$. These two approximations are asymptotically equal:

Proposition 10. *We have under Assumption 1 that:*
(i) when $n \to +\infty$*,*

$$\lim_{n\to+\infty} d_{sub}(G_n, \widehat{\chi}_n) = 0. \tag{55}$$

(ii) The limit of the empirical graphon $\widehat{\chi}_n$ *is thus the biased graphon* $\kappa_{\Gamma^{-1}}$*.*

$$\lim_{n\to+\infty} d_{\mathrm{sub}}(\widehat{\chi}_n, \kappa_{\Gamma^{-1}}) = 0. \tag{56}$$

Proof. We postpone the proof of Proposition 10 (i) to the Section 4.1.2. For the point (ii), we have:

$$d_{\mathrm{sub}}(\widehat{\chi}_n, \kappa_{\Gamma^{-1}}) \leq d_{\mathrm{sub}}(\widehat{\chi}_n, G(H_n,\kappa)) + d_{\mathrm{sub}}(G(H_n,\kappa), \kappa_{\Gamma^{-1}}).$$

The first term in the right hand side is upper bounded by C/n by Proposition 10. The second term is the Proposition 2 shown in [3, Corollary 2.2]. □

As a consequence, using the result of Athreya and Roellin [3] (see Proposition 2), we obtain:

Proposition 11. *Under Assumptions 1,*
(i) $\widehat{\rho}$ *is a consistent estimator of* π*, and for* $q, r \in [\![1, Q]\!]$*,*

$$\lim_{n\to+\infty} \widehat{\rho}_{qr}^n = \pi_{qr}, \qquad \textit{and} \qquad \lim_{n\to+\infty} \widehat{\gamma}_q^n = \Gamma\Big(\sum_{r=1}^{q}\alpha_r\Big) - \Gamma\Big(\sum_{r=1}^{q-1}\alpha_r\Big) =: \gamma_q. \tag{57}$$

It follows that a consistent estimator of α_q *is*

$$\widehat{\alpha}_q^n = \Gamma_n^{-1}\Big(\sum_{r=1}^{q}\widehat{\gamma}_r^n\Big) - \Gamma_n^{-1}\Big(\sum_{r=1}^{q-1}\widehat{\gamma}_r^n\Big). \tag{58}$$

(ii) In the special case of $Q = 2$*, an estimator of* α_1 *is* $\widehat{\alpha}_1^n = \Gamma_n^{-1}(\widehat{\gamma}_1^n)$*.*

Proof. Let us consider point (i). The limit for $\widehat{\gamma}_q^n$ follows from the ergodic theorem. Indeed, we can write that

$$\widehat{\gamma}_q^n = \frac{N_n^q}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i^{(n)} \in]\sum_{r=1}^{q-1}\alpha_r, \sum_{r=1}^{q}\alpha_r]}.$$

The ergodic theorem for the Markov chain $(X^n)_n$ says that

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i^{(n)}\in I_q} = \mathbb{E}_m[\mathbf{1}_{X_1\in I_q}] = \Gamma(\sum_{r=1}^{q}\alpha_r) - \Gamma(\sum_{r=1}^{q-1}\alpha_r) = \gamma_q.$$

It remains to prove that $\widehat{\rho}_{qr}^n$ is a consistent estimator of π_{qr}. Rewrite $\widehat{\rho}_{qr}^n$ as

$$\widehat{\rho}_{qr}^n = \frac{N_n^{q\leftrightarrow r}/n^2}{\frac{N_n^q}{n}\frac{N_n^r}{n}} = \frac{1}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n}\frac{1}{n^2}N_n^{q\leftrightarrow r}.$$

Recall that the subgraph G_n is constructed from the Markov chain $X^{(n)}$ and that each pair of non-consecutive vertices X_i and X_j are connected with probability $\kappa(Z_i, Z_j)$ depending on theirs types and independently of the others edges. Let us focus on the number of edges $N_n^{q\leftrightarrow r}$: two cases have to be distinguished.

Case 1, $q \neq r$**:** The number of edges of types (q, r) is

$$N_n^{q\leftrightarrow r} = \sum_{i=1}^{n-1} \mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r} + \sum_{\substack{1\le i,j\le n \\ \{X_i,X_j\}\notin E(H_n)}} \mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_r}.$$

Then,

$$\widehat{\rho}_{qr}^n = \frac{1}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n n}\left(\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r}\right) + \frac{1}{n^2}\sum_{\substack{1\le i,j\le n \\ \{X_i,X_j\}\notin E(H_n)}} \frac{\mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_r}}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n} \tag{59}$$

By the ergodic theorem for Markov chain $X^{(n)}$, we have

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r} = \mathbb{E}_m[\mathbf{1}_{X_0\in I_q, X_1\in I_r}] = \gamma_q\pi_{qr} < +\infty.$$

Since $\lim_{n\to+\infty}\widehat{\gamma}_q^n = \gamma_q > 0$ in probability, there exists a constant $c > 0$ such that $c \le \inf_{q\in\{1,\dots Q\}}\gamma_q$ and

$$\lim_{n\to+\infty}\mathbb{P}\left(\frac{1}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n n}\left(\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r}\right) \le \frac{1}{c^2 n}\left(\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r}\right)\right) = 1,$$

and hence the first term in the right hand side of (59) converges to 0 in probability.

Consider now the second term in the r.h.s. of (59). Let us define the function

$$f(G_n) = \frac{1}{n^2}\sum_{\substack{1\le i,j\le n \\ \{X_i,X_j\}\notin E(H_n)}} \mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_r},$$

then f is a function of the $n(n-1)/2-(n-1)=(n-1)(n-2)/2$ random edges on n vertices. We see that

$$\mathbb{E}[f(G_n)]=\mathbb{E}\big[\frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_r}\big]=\frac{(n-1)(n-2)}{n^2}\pi_{qr}\gamma_q\gamma_r.$$

We have

$$\begin{aligned}\mathbb{P}\big(\big|\frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}&\frac{\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_r}}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n}-\pi_{qr}\big|>\varepsilon\big)\\ &\le\mathbb{P}\big(\frac{1}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n}\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon-\big|\frac{1}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n}\mathbb{E}[f(G_n)]-\pi_{qr}\big|\big)\\ &=\mathbb{P}\big(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon\widehat{\gamma}_q^n\widehat{\gamma}_r^n-|\mathbb{E}[f(G_n)]-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\pi_{qr}|\big)\\ &=\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon\widehat{\gamma}_q^n\widehat{\gamma}_r^n-\pi_{qr}\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|\right)\end{aligned}$$

For $c<\inf_{q\in\{1,\dots Q\}}\gamma_q$,

$$\begin{aligned}&\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon\widehat{\gamma}_q^n\widehat{\gamma}_r^n-\pi_{qr}\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|\right)\\ &\le\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>c^2\varepsilon-\frac{c^3}{2}\varepsilon\right)+\mathbb{P}\left(\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|>\frac{c^3\varepsilon}{2\pi_{qr}}\right)+\mathbb{P}(\widehat{\gamma}_q^n\widehat{\gamma}_r^n<c^2).\end{aligned}\tag{60}$$

Since $\lim_{n\to+\infty}\widehat{\gamma}_q^n=\gamma_q>0$ in probability, for fixed $\varepsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|<\frac{c^3\varepsilon}{2\pi_{qr}}\ \text{ and }\ \widehat{\gamma}_q^n\widehat{\gamma}_r^n>c^2\right)=1$$

Thus the second and the third terms on the right hand side of (60) tend to zero as n tends to infinity. It remains the first term to be treated. When one edge is changed, the value of f is changed by most $1/n^2$. Applying McDiarmid's concentration [23] for function f , we obtain:

$$\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>c^2\varepsilon-\frac{c^3}{2}\varepsilon\right)\le 2\exp\big(-\frac{2(c^2-\frac{c^3}{2})\varepsilon}{\frac{(n-1)(n-2)}{2}\frac{1}{n^4}}\big)\le 2e^{-4n^2c^2(1-c/2)\varepsilon}.$$

Note that $0<c<1$ then $c^2(1-c/2)>0$. We use Borel-Cantelli's Theorem to conclude that $\lim_{n\to+\infty}\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>c^2\varepsilon-\frac{c^3}{2}\varepsilon\right)=0$ and hence,

$$\big|\frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\frac{\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_r}}{\widehat{\gamma}_q^n\widehat{\gamma}_r^n}-\pi_{qr}\big|\longrightarrow 0$$

in probability as $n\to\infty$. This finishes the proof for Case 1.

Case 2, $q=r$**:** The proof follows by similar arguments, with notice that there are a few modifications because the expression of $N_n^{q\leftrightarrow q}$ is slightly different:

$$N_n^{q\leftrightarrow q}=\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q,X_{i+1}\in I_q}+\frac{1}{2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_q}.$$

Then,

$$\widehat{\rho}^n_{qq} = \frac{1}{\widehat{\gamma}^n_q\big(n\widehat{\gamma}^n_q - 1\big)}\left(\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q, X_{i+1}\in I_q}\right) + \frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\frac{\mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_q}}{\widehat{\gamma}^n_q\big(\widehat{\gamma}^n_q - 1/n\big)} \tag{61}$$

We have that the first term on r.h.s. of (61) converges in probability to 0 as in case 1. For the second term on r.h.s. of (61), we define the function f as in Case 1 by

$$f(G_n) = \frac{1}{2n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_q},$$

For a fixed $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\frac{\mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_q}}{\widehat{\gamma}^n_q\big(\widehat{\gamma}^n_q - 1/n\big)} - \pi_{qq}\Big| > \varepsilon\right)\\
&\qquad\le \mathbb{P}\left(\big|f(G_n) - \mathbb{E}[f(G_n)]\big| > \varepsilon\widehat{\gamma}^n_q\big(\widehat{\gamma}^n_q - 1/n\big) - \pi_{qq}\left|\frac{(n-1)(n-2)}{n^2}(\gamma_q)^2 - \widehat{\gamma}^n_q\big(\widehat{\gamma}^n_q - 1/n\big)\right|\right)\\
&\qquad\le \mathbb{P}\left(\big|f(G_n) - \mathbb{E}[f(G_n)]\big| > c\big(c - \frac{1}{n}\big)\varepsilon - \frac{c^3}{2}\varepsilon\right) + \mathbb{P}(\widehat{\gamma}^n_q < c)\\
&\qquad\qquad + \mathbb{P}\left(\left|\frac{(n-1)(n-2)}{n^2}(\gamma_q)^2 - \widehat{\gamma}^n_q\big(\widehat{\gamma}^n_q - \frac{1}{n}\big)\right| > \frac{c^3\varepsilon}{2\pi_{qq}}\right).
\end{aligned}$$

As in Case 1, the second and the third term on r.h.s. of above inequality are negligible. Applying McDiarmid's concentration for f with notice that when changing 1 edge in G_n, the value of f changes at most $1/n^2$,

$$\mathbb{P}\left(\big|f(G_n) - \mathbb{E}[f(G_n)]\big| > c(c-1/n)\varepsilon - \frac{c^3}{2}\varepsilon\right) \le 2\exp\big(-\frac{2(c^2 - c/n - \frac{c^3}{2})\varepsilon}{\frac{(n-1)(n-2)}{2}\frac{1}{n^4}}\big) \le 2e^{-2(n^2c^2(1-c/2)-nc)\varepsilon}.$$

Finally, using Borel-Cantelli's Theorem, $|f(G_n) - \mathbb{E}[f(G_n)]| \to 0$ almost surely as n tends to infinity. Thus, the point (i) is proved. □

4.1.2 Proof of Proposition 10

From now on, for the sake of simplicity, we assume for the that there are two classes of vertices in the graph, i.e. $Q = 2$. The proof can be generalized to general Q by following the same steps. Our parameters' notations are simplified as $\gamma^1_n =: \gamma_n$ and $\gamma^1_\infty =: \gamma_\infty = \Gamma(\alpha)$.

Our purpose is to prove a convergence of graphons for the distance d_{sub} introduced in (7) using the densities (5). If F is an edge (meaning that $F = K_2$, the complete graph of 2 vertices), then the density of F in $G_n := G(X_n, H_n, \kappa)$ is the proportion of edges,

$$t(F, G_n) = \frac{1}{n(n-1)}\sum_{\ell,\ell'\in[\![1,n]\!]}\mathbf{1}_{\ell\sim_{G_n}\ell'}$$

$$\text{and}\quad t(F,\chi_n) = \int_{[0,1]^2}\widehat{\chi}_n(x_1,x_2)dx_1dx_2 = \sum_{q,r=1}^{Q}\widehat{\gamma}^n_q\widehat{\gamma}^n_r\widehat{\rho}^n_{qr}.$$

In general case, if F is a graph of k vertices,

$$t(F,G_n)=\frac{1}{(n)_k}\sum_{(i_1,\cdots i_k)\in[\![1,n]\!]}\prod_{\{\ell,\ell'\}\in E(F)}\mathbf{1}_{i_\ell\sim_G i_{\ell'}} \tag{62}$$

$$t(F,\chi_n)=\int_{[0,1]^k}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_n^{qr}\mathbf{1}_{J_q^n\times J_r^n}(x_\ell,x_{\ell'})\right)dx_1\cdots dx_k \tag{63}$$

Let us first consider the case where F is an edge.

$$\begin{aligned}
|t(F,G_n)-t(F,\chi_n)| &= \left|\frac{1}{(n)_2}\sum_{(i,j)\in[\![1,n]\!]}\mathbf{1}_{i\sim_{G_n}j}-\int_{[0,1]^2}\widehat{\chi}_n(x_1,x_2)\;dx_1dx_2\right| \\
&\le \left|\frac{1}{(n)_2}\sum_{(i,j)\in[\![1,n]\!]}\left(\mathbf{1}_{i\sim_{G_n}j}-\widehat{\rho}_{Z_i,Z_j}\right)\right| \\
&\quad+\left|\frac{1}{(n)_2}\sum_{(i,j)\in[\![1,n]\!]}\widehat{\rho}_{Z_i,Z_j}-(\widehat{\gamma}_1^n)^2\widehat{\rho}_{11}^n-2\widehat{\gamma}_1^n(1-\widehat{\gamma}_1^n)\widehat{\rho}_{12}^n-(1-\widehat{\gamma}_1^n)^2\widehat{\rho}_{22}^n\right| \\
&\le \left|\frac{1}{(n)_2}\sum_{(i,j)\in[\![1,n]\!]}\left(\mathbf{1}_{i\sim_{G_n}j}-\widehat{\rho}_{Z_i,Z_j}\right)\right|+\left|\widehat{\rho}_{11}^n\left(\sum_{(i,j)\;|\;(Z_i,Z_j)=(1,1)}\frac{1}{(n)_2}-(\widehat{\gamma}_1^n)^2\right)\right| \\
&\quad+\left|\widehat{\rho}_{22}^n\left(\sum_{(i,j)\;|\;(Z_i,Z_j)=(2,2)}\frac{1}{(n)_2}-(1-\widehat{\gamma}_1^n)^2\right)\right|+\left|\widehat{\rho}_{12}^n\left(\sum_{\substack{(i,j)\;|\;(Z_i,Z_j)=(1,2)\\ \text{or}(Z_i,Z_j)=(2,1)}}\frac{1}{(n)_2}-2\widehat{\gamma}_1^n(1-\widehat{\gamma}_1^n)\right)\right|.
\end{aligned}$$

By the law of large numbers and using (57) whose proof does not depend on the Proposition 10, the four terms converge to zero.

In the general case, proceeding in a similar way leads to:

$$\begin{aligned}
&|t(F,G_n)-t(F,\chi_n)| \\
\le&\left|\frac{1}{(n)_k}\sum_{(i_1,\cdots i_k)\in[\![1,n]\!]}\prod_{\{\ell,\ell'\}\in E(F)}\mathbf{1}_{i_\ell\sim_G i_{\ell'}}-\frac{1}{(n)_k}\sum_{(i_1,\cdots,i_k)}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{Z_{i_\ell}=q,Z_{i_{\ell'}}=r}\right)\right| \\
&+\left|\frac{1}{(n)_k}\sum_{(i_1,\cdots,i_k)}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{Z_{i_\ell}=q,Z_{i_{\ell'}}=r}\right)-\frac{1}{n^k}\sum_{1\le i_1,\cdots,i_k\le n}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{Z_{i_\ell}=q,Z_{i_{\ell'}}=r}\right)\right| \\
&+\left|\frac{1}{n^k}\sum_{1\le i_1,\cdots,i_k\le n}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{Z_{i_\ell}=q,Z_{i_{\ell'}}=r}\right)-\int_{[0,1]^k}\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{J_q^n\times J_r^n}(x_\ell,x_{\ell'})\right)dx_1\cdots dx_k\right|
\end{aligned}$$

As $\prod_{\{\ell,\ell'\}\in E(F)}\mathbf{1}_{i_\ell\sim_G i_{\ell'}}$ and $\prod_{\{\ell,\ell'\}\in E(F)}\left(\sum_{q,r=1}^{Q}\widehat{\rho}_{qr}^n\mathbf{1}_{Z_{i_\ell}=q,Z_{i_{\ell'}}=r}\right)$ are bounded by 1, there exist $c(k)$ such that the first term and the second term in the right hand side are bounded by $c(k)/n$. For the third term, it is equal to

$$\left|\sum_{1\le q_1,\ldots,q_k\le Q}\prod_{\{\ell,\ell'\}\in E(F)}\widehat{\rho}_{q_\ell,q_{\ell'}}^n\left(\frac{1}{n^k}\sum_{1\le i_1,\cdots,i_k\le n}\mathbf{1}_{Z_{i_1}=q_{i_1},\cdots,Z_{i_k}=q_{i_k}}-\int_{[0,1]^k}\prod_{h=1}^{k}\mathbf{1}_{J_{q_h}^n}(x_h)dx_1\cdots dx_k\right)\right|$$

Since $0 \leq \prod_{\{\ell,\ell'\}\in E(F)} \widehat{\rho}^n_{q_\ell,q_{\ell'}} \leq 1$ and $\{Z_{i_1} = q_{i_1}, \cdots, Z_{i_k} = q_{i_k}\} = \{\Gamma(X_{i_1}) \in J_{q_1}, \cdots, \Gamma(X_{i_k}) \in J_{q_k}\}$, the third term is thus bounded by

$$\begin{aligned}
&\sum_{1\leq q_1,\ldots,q_k\leq Q} \left| \frac{1}{n^k} \sum_{1\leq i_1,\cdots,i_k\leq n} \mathbf{1}_{\Gamma(X_{i_1})\in J_{q_1},\cdots,\Gamma(X_{i_k})\in J_{q_k}} - \int_{[0,1]^k} \prod_{h=1}^k \mathbf{1}_{J^n_{q_h}}(x_h) dx_1 \cdots dx_k \right| \\
&\qquad = \sum_{1\leq q_1,\ldots,q_k\leq Q} \left| \frac{1}{n^k} \sum_{1\leq i_1,\cdots,i_k\leq n} \prod_{\ell=1}^k \mathbf{1}_{\Gamma(X_{i_\ell})\in J_{i_\ell}} - \prod_{\ell=1}^k \int_{[0,1]} \mathbf{1}_{J^n_{i_\ell}} dx_\ell \right| \\
&\qquad = \sum_{1\leq q_1,\ldots,q_k\leq Q} \left| \frac{\prod_{\ell=1}^k \sum_{i_\ell=1}^n \mathbf{1}_{\Gamma(X_{i_\ell})\in J_{q_l}}}{n^k} - \prod_{\ell=1}^k \int_{J^n_{q_\ell}} dx_\ell \right| \\
&\qquad = \sum_{1\leq q_1,\ldots,q_k\leq Q} \left| \prod_{\ell=1}^k \frac{N^{q_\ell}_n}{n} - \prod_{\ell=1}^k \widehat{\gamma}^n_{q_\ell} \right| = 0.
\end{aligned}$$

Hence $\lim_{n\to+\infty} |t(F,G_n) - t(F,\chi_n)| = 0$. Because $t(F,G_n)$ and $t(F,\chi_n)$ are bounded independently from n, this provides the announced result.

4.2 Incomplete observations and graphon de-biasing

In Proposition 11, it is shown that the 'classical' SBM estimator (20) obtained by neglecting the bias coming from the sampling scheme can be corrected by using the inverse of the cumulative distribution function Γ of m. When the types are unobserved, we proceed in the same way. We assume here that the types Z_i are unobserved, but we need the observation of the marks X_i, otherwise no de-biasing is permitted since the cumulative distribution function Γ can not be estimated. We detail this estimation procedure in the case $Q = 2$ for the sake of simplicity, but generalization is straightforward.

Step 1: First, we perform an estimation of the SBM neglecting the sampling biases. This amounts to computing the estimator proposed in [9]:

- We follow the algorithm described in Section 3.2.1, but with the likelihood $\mathcal{L}^{\text{class}}(Z_i, Y_{ij}; \theta)$ given in (19). We denote the parameter here by $\theta = (\gamma_1, 1-\gamma_1, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$.

- For the proposal distribution of the types Z^c, it is simpler since we assume that the X_i's are known. Assume that we are at step k and that we dispose of the parameters $\theta^{(k-1)}$. We initialize the types by attributing the types 1 to the $X_i \leq \gamma^{(0)}$ and 2 to the others. At each step, the threshold is modified from $\gamma_1^{(k-1)}$ to $\gamma_1^{(k)}$ by following a random walk: a gaussian increment (mean 0 and variance s^2) is added. All the X_i smaller than this increment are given the type $Z_i = 1$ and the others the type $Z_i = 2$.

Step 2: We estimate the cumulative distribution function Γ_n (see (12)) and deduce the graphon estimator $\widehat{\alpha}^n_1$ of α_1 using (58). This provides the estimator of κ:

$$\widehat{\kappa}_n(x,y) := \sum_{q=1}^Q \sum_{r=1}^Q \widehat{\rho}^n_{qr} \mathbf{1}_{[\sum_{k=1}^{q-1}\widehat{\alpha}^n_k, \sum_{k=1}^{q}\widehat{\alpha}^n_k)}(x) \mathbf{1}_{[\sum_{k=1}^{r-1}\widehat{\alpha}^n_k, \sum_{k=1}^{r}\widehat{\alpha}^n_k)}(y). \tag{64}$$

5 Numerical results

For the simulation, we consider RDS graphs obtained from the exploration of SBM graphons with $Q = 2$ classes, of respective proportions $\alpha_1 = 2/3$ and $\alpha_2 = 1/3$. The connection probabilities are:

$$\pi = \begin{pmatrix} 0.7 & 0.4 \\ 0.4 & 0.8 \end{pmatrix}.$$

The RDS graphs consist of $n = 50$ vertices.

We proceed to the four estimations presented in this paper:

- the algorithm of Section 3.1 for complete observations by assuming that the types $Z_i \in \{1, 2\}$ are observed. In the estimation, the system of equations (21)-(26) is solved. For this, we look numerically for the zeros of (31) and choose the solution corresponding to the highest likelihood. For the bisection method (REF), we use a grid of step 10^{-2}.
- the SAEM algorithm of Section 3.2.1 when the types Z_i are unobserved. The SAEM is based on an iteration on k and we perform $K = 200$ iterations.
- the computation of the estimators given in Proposition 11 assuming complete observations,
- the debiasing of the SAEM algorithm of Daudin et al. presented in Section 4.2. Again, we use $K = 200$ iterations for the SAEM iterations.

We proceed to a Monte-Carlo study of the estimators' distributions. We simulate 200 RDS graphs, and for each of them, apply the four estimation strategies. The empirical distribution of the estimators are represented in Fig. 3, and this allows us to estimate the associated mean squares errors (MSE) for each method, see Table 1.

Parameters	Complete likelihood	SAEM	De-biased graphon	De-biased graphon with SAEM
π_{11}	$3.74\ 10^{-4}$	$9.69\ 10^{-3}$	$4.45\ 10^{-4}$	$4.43\ 10^{-4}$
π_{12}	$4.88\ 10^{-4}$	$1.32\ 10^{-2}$	$6.63\ 10^{-4}$	$8.92\ 10^{-4}$
π_{22}	$1.30\ 10^{-3}$	$2.70\ 10^{-2}$	$1.45\ 10^{-3}$	$1.36\ 10^{-3}$
α	$1.04\ 10^{-2}$	$3.77\ 10^{-2}$	$9.35\ 10^{-4}$	$7.60\ 10^{-4}$

Table 1: *Mean square errors.*

Without surprise, the estimation is better when we have complete observations (columns 1 and 3). The estimation of α based on the estimator (58) is better than the MLE obtained in column 1 from an MSE point of view.
To understand the difficulty in estimating α, recall that for the MLE estimators based on the true likelihood, $\widehat{\alpha}$ is estimated from $\widehat{\beta}$ (see (29)). The shape of function $\beta = \frac{\alpha}{1-\alpha}$ (see figure 5) indicates that values of α smaller than $1/2$ give similar values of β and thus, when $\alpha \in (0, 1/2)$, its estimation from β is more difficult. For that reason, when $\alpha < 1/2$, we can not obtain a good estimation, even though π might be well-estimated. Nevertheless, in the case $\alpha \in (1/2, 1)$, β varies sufficiently to allow an estimation of α with better precision. So our recommendation is that when there are 2 classes of vertices, to choose as type 1 the majority type so that $\alpha > 1/2$. However, it seems that estimating α from γ (see (58)) rather than from β is much more precise.

When the types Z_i are not observed, we achieve better MSEs with the debiasing of the classical SAEM method of Daudin et al. (column 4 of Table 1). Notice first that the columns 2 and 4 of Table 1 are not completely equivalent, since the debiasing methods of Section 4 necessitate the knowledge of the positions X_i of the Markov chain, when the likelihood (13) necessitates only the connections Y_{ij} and the types Z_i's. Second, the updating of the types in the SAEM algorithm is easier in Section 4.2 when the X_i's are known since it amounts to choosing the threshold that separates the types 1 and 2. Finally, the SAEM algorithm on the classical likelihood (19) seems to converge more easily than for the likelihood (13).

5.1 Conclusion

Four statistical methods are studied in this paper, for estimating SBM parameters using a subgraph obtained from the exploration of the graphon by a Markov chain. This is a toy model for estimating

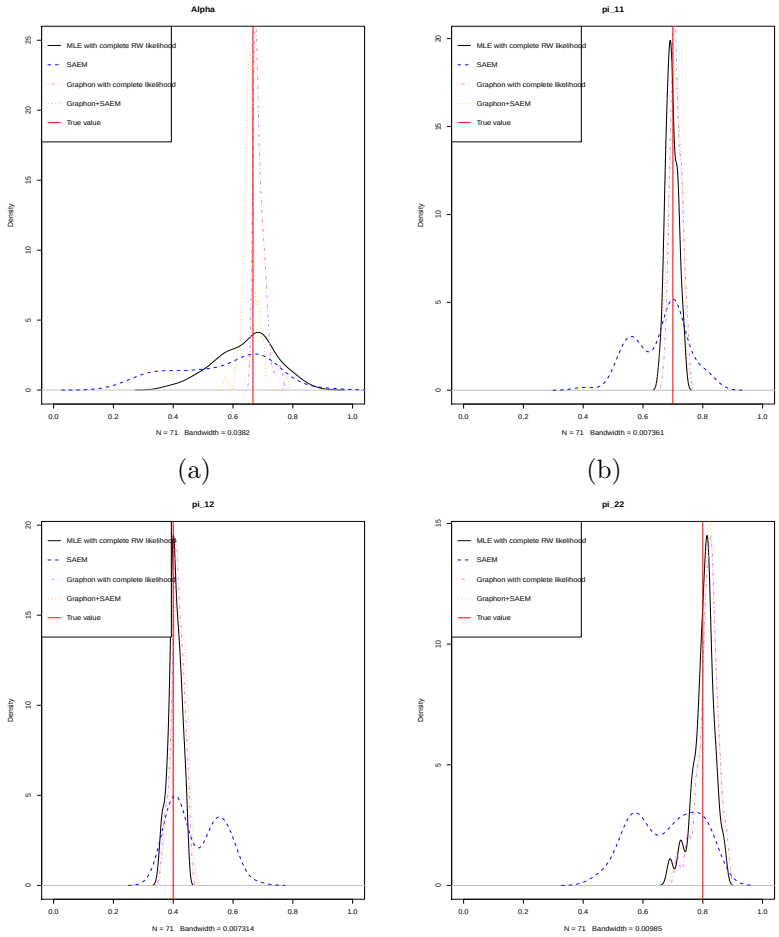

Figure 3: *E*stimation on complete data for a graph of $n = 50$ vertices with $Q = 2$ classes and parameters $\alpha_1 = 2/3$, $\pi_{11} = 0.7$, $\pi_{12} = \pi_{21} = 0.4$ and $\pi_{22} = 0.8$. 200 such graphs are simulated and the empirical distributions of the estimators are represented here with the true parameters in red line. (a): estimator of α, (b): estimator of $\pi_1 1$, (c): estimator of π_{12}, (d) estimator of π_{22}.

random networks from chain-referral sampling techniques and there exist sampling biases. The two first methods compute the maximum likelihood estimator when the types of the nodes are known or unknown. On simulations, it appears that the SAEM algorithm used when the types are unobserved is not very robust and provides relatively large MSEs. An alternative approach is proposed by taking advantage of recent results by Athreya and Roellin [3]: this allows to correct the classical SBM estimators that would be proposed if one ignores the sampling biases. These methods provide good estimators but rely on the precise knowledge of the Markov chain exploring the SBM graphon (in particular the positions X_i's), which is not always available.

References

[1] E. Abbe. Community detection and stochastic block models: recent development. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[2] E. Allman, C. Matias, and J. Rhodes. Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736, 2011.

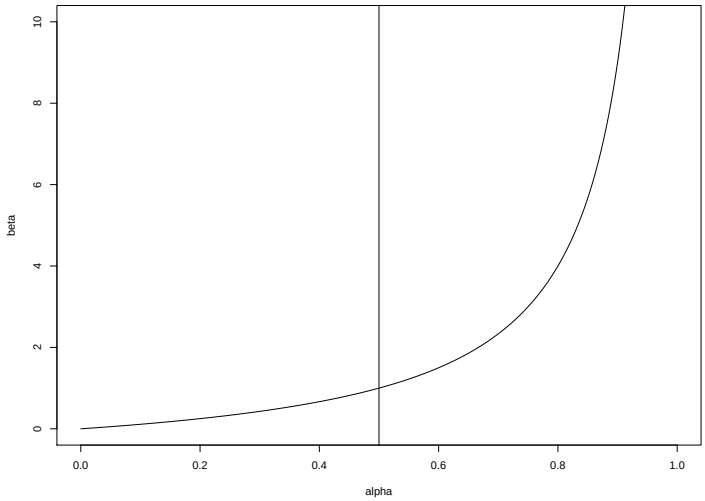

Figure 4: The correlation of β and α.

[3] S. Athreya and A. Röllin. Dense graph limits under respondent-driven sampling. *Annals of Applied Probability*, 44:2193–2210, 2016.

[4] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

[5] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Annals of Mathematics*, pages 151–219, 2012.

[6] G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314, 1996.

[7] G. Celeux and J. Diebolt. The sem algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

[8] F. Crawford, J. Wu, and R. Heimer. Hidden population size estimation from respondent-driven sampling: a network approach. *Journal of the American Statistical Association*, 113:755–766, 2018.

[9] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, 2008.

[10] K. Gile. Improved inference for Respondent-Driven Sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146, 2011.

[11] K. Gile and M. Handcock. Respondent-driven sampling: an assessment of current methodology. *Sociol. Methodol.*, 40:285–327, 2010.

[12] K. Gile, L. Johnston, and M. Salganik. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society A*, 178:241–269, 2015.

[13] L. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961.

[14] D. Heckathorn. Respondent-driven Sampling: a new approach to the study of hidden populations. *Social Problems*, 44(1):74–99, 1997.

[15] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: some first steps. *Social Networks*, 5:109–137, 1983.

[16] T. Jaakkola. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*, Cambridge, 2000. MIT Press.

[17] M. Jordana, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[18] M. Khabbazian, B. Hanlon, Z. Russek, and K. Rohe. Novel sampling design for respondent-driven sampling. *Electronic Journal of Statistics*, 11(2):4769–4812, 2017.

[19] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: PS*, 8:115–131, 2004.

[20] X. Li and K. Rohe. Central limit theorems for network driven sampling. *Electronic Journal of Statistics*, 11(2):4871–4895, 2017.

[21] L. Lovász. *Large networks and graph limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, Rhode Island, 2012.

[22] M. Mariadassou and T. Tabouy. Consistency and asymptotic normality of stochastic block models estimators from sampled data. arXiv:1903.12488, 2019.

[23] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188, Cambridge, 1989. Cambridge University Press.

[24] T. Mouw and A. Verdery. Network sampling with memory: a proposal for more efficient sampling from social networks. *Sociological Methodology*, 42:206–256, 2012.

[25] O. Riordan. The phase transition in the configuration model. *Combinatorics, Probability and Computing*, 21(1-2):265–299, 2012.

[26] K. Rohe. A critical threshold for design effects in network sampling. *Annals of Statistics*, 47(1):556–582, 2019.

[27] D. Rolls, P. Wang, R. Jenkinson, P. Pattison, G. Robins, R. Sacks-Davis, G. Daraganova, M. Hellard, and E. McBryde. Modelling a disease-relevant contact network of people who inject drugs. *Social Networks*, 35(4):699–710, 2013.

[28] T. Tabouy, P. Barbillon, and J. Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 2019.

[29] V. Tran, C. Jangal, P. Feuillet, A. Bardot, C. Dumont, I. Condamine-Ducreux, and M. Jauffret-Roustide. Respondent-driven sampling survey among people who inject drugs in paris. in progress, 2020.

[30] T. Vo. *Exploration d'un graphe aléatoire par des méthodes Respondent Driven Sampling*. PhD thesis, Université Sorbonne Paris Nord, Paris, France, 2020.

[31] E. Volz and D. Heckathorn. Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, 24:79–97, 2008.