Exploration of random graphs by the Respondent Driven Sampling (RDS) method.

Thi Phuong Thuy Vo

Exploration of random graphs by the Respondent Driven Sampling method.

PhD Thesis

VO THI PHUONG THUY

École Doctorale 146

THÈSE

pour obtenir le grade de docteur délivré par

Université Sorbonne Paris Nord

Spécialité doctorale "*Mathématiques appliquées*"

Exploration d'un graphe aléatoire par des méthodes Respondent Driven Sampling (RDS)

présentée publiquement par **VO THI PHUONG THUY**

JURY

Directeur:		
Jean-Stéphane	DHERSIN	Université Sorbonne Paris Nord
Co-encadrant:		
Viet Chi	TRAN	Université Gustave Eiffel
Rapporteurs:		
Stéphane	ROBIN	INRAe
Adrian	RÖLLIN	National University of Singapore
Examinateurs:		
Hélène	GUERIN	Université du Québec à Montréal
Bénédicte	HAAS	Université Sorbonne Paris Nord
Laurent	MÉNARD	Université Paris Nanterre
Amandine	VEBER	Université de Paris
Pierre-André	ZITT	Université Gustave Eiffel

19 Novembre 2020
Villetaneuse, France

Résumé

L'échantillonage en fonction des répondants ("Respondent Driven Sampling", RDS) peut être utilisé pour découvrir des réseaux sociaux dans des populations cachées. Ceci peut conduire à l'étude d'une chaîne de Markov sur un graphe aléatoire dont les sommets représentent les individus et dont les arêtes décrivent les relations entre les deux personnes qu'elles relient. Les personnes interrogées sont invitées à indiquer leurs partenaires et un certain nombre de coupons sont remis à certaines de ces personnes. Par chaînage on peut ainsi retrouver les noeuds cachés dans la population en suivant au hasard les arêtes du réseau social sous-jacent.

Nous considérons un processus renormalisé de la chaîne de référence sur le modèle Erdös-Rényi, puis sur le modèle à blocs stochastiques ("Stochastic Block Model", SBM), qui en est une extension lorsque les populations sont partitionnées en communautés. La difficulté réside dans la gestion de l'hétérogénéité du graphe. Dans notre étude, le graphe et la marche aléatoire sont construits simultanément. Nous démontrons que lorsque la taille de la population est grande et les graphes sparses, le processus aléatoire représentant la fraction du graphe découverte, correctement normalisé, se comporte comme une courbe déterministe qui est la solution unique d'un système d'ODE.

Par ailleurs, nous nous intéressons également au problème de récupérer des informations statistiques sur un modèle à bloc stochastique à partir du sous-graphe découvert par une marche aléatoire (correspondant à un RDS à un coupon). Nous considérons ici le cas dense où le réseau aléatoire peut être approché par un graphon. Tout d'abord, nous écrivons la vraisemblance du sous-graphe découvert par la marche aléatoire: des biais émergent car les "hubs" et les types majoritaires sont plus susceptibles d'être échantillonnés. Même dans le cas où les types sont observés, l'estimateur du maximum de vraisemblance n'est plus explicite. Lorsque les types de sommets ne

sont pas observés, nous utilisons un algorithme SAEM ("Stochastic Approximation version of Expectation-Maximization algorithm") pour maximiser la vraisemblance. Deuxièmement, nous proposons une stratégie d'estimation différente en utilisant les nouveaux résultats d'Athreya et Röllin. Elle consiste à dé-biaiser l'estimateur EM variationnel proposé par Daudin et al. et qui ignore les biais.

Mots clés: *graphe aléatoire; Erdös-Rényi graphe; stochastic block model; graphon; processus stochastique; chaîne de Markov; théorème centrale limite; exploration du marche aléatoire; sondage biaisé; EM estimation; EM approximation stochastique; vraisemblance incomplète; respondent driven sampling*

Abstract

The study of Respondent Driven Sampling (RDS) is invested for the discovery of a social network of hidden populations. It leads to the study of a Markov chain on a random graph whose vertices represent individuals and whose edges describe the relationships between the people connected. Respondents are asked to list their partners and a certain number of coupons are given to some of the latter. The RDS survey searches for hidden nodes in the population by randomly following the edges of the underlying social network, which allows us to trace the sampled individuals.

We consider the normalized process of the reference chain on the Erdös-Rényi model, then on its generalization, the Stochastic Block Model (SBM) when populations are partitioned into communities. We prove that when the population size is large and the graph is sparse, the normalized stochastic process describing the fraction of the graph discovered behaves like a deterministic curve which is the unique solution of a system of ODEs. In our model, the graph and the random walk are constructed simultaneously. The difficulty lies in handling the heterogeneity of the graph.

Furthermore, we are also interested in the problem of recovering statistical information on a SBM from the subgraph discovered by an exploring random walk (RDS with 1 coupon per interviewee). We consider here the dense case where the random network can be approximated by a graphon. First, we write the probability of the subgraph discovered by the random walk: biases emerge because the hubs and the 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 are unobserved, we use an SAEM (Stochastic approximation of Expectation-Maximization) algorithm to maximize the likelihood. Second, we propose a different estimation strategy using new results by Athreya and Röllin. It consists in de-biasing the variational EM estimator proposed in Daudin et al. and

that ignores the biases.

Keywords: *random graph; Erdös-Rényi graph; stochastic block model; graphon; stochastic processes; Markov chain; central limit theorem; random walk exploration; sampling bias; EM estimation; stochastic approximation expectation-maximization; incomplete likelihood; respondent driven sampling; chain-referral survey*

Acknowledgements

First and foremost, I wish to express the deepest gratitude and appreciation to my supervisors, Professor Jean-Stéphane Dhersin and Professor Viet Chi Tran, who had brought me a great opportunity and supported me to pursue my study in France. They always encouraged me and gave me the best guidance to keep going through all the challenges from the start to the end of this thesis. I am thankful and feel extremely fortunate to be a student of such wonderful teachers. Without the their patience and kindness, this work could not have been accomplished.

I would wish to express my deep gratitude to Professor Stéphane Robin and Professor Adrian Röllin for having accepted to be reviewers of my thesis. I am very grateful that they took the time to read the thesis carefully and have given the insightful comments and suggestions.

I would wish to extend my deep gratitude to the rest of my committee: Hélène Guerin, Bénédicte Haas, Laurent Ménard, Amandine Veber and Pierre-André Zitt for having accepted to be members of the jury, especially in this difficult time.

I would also like to acknowledge the supports of Laboratory of Analysis, Geometry and Application (LAGA), University Sorbonne Paris Nord; the ANR Econet (ANR-18-CE02-0010) and the Chair “Modélisation Mathématique et Biodiversité” (MMB) of Veolia Environnement-Ecole Polytechnique-Museum National d’Histoire Naturelle-Fondation X. Without their funding and supports, this thesis could not have achieved its goals.

I am also grateful to thank the staffs, my colleagues, my friends in LAGA for all their helps and supports, especially to the Probability and Statistics team for the

interesting and helpful discussions during the time I worked on my thesis.

I would also like to thank Professor Phung Ho Hai, Professor Nguyen Viet Dung and all the professors from the Institute of Mathematics Hanoi, from University of Hue in Vietnam for having supported and given me the opportunities to pursue the study abroad.

Lastly, I would love to say thanks to my Vietnamese friends and family:

Em xin gửi lời cảm ơn chân thành và sâu sắc đến những người anh, người chị, người bạn thân thiết trường Paris 13 vì đã quan tâm, chia sẻ, giúp đỡ em như những người thân thật sự trong gia đình suốt quãng thời gian ở Pháp.

Cảm ơn thành viên lớp Cao Học Chơi vì những tình cảm thật đẹp, thật ấm áp dành cho em. Đặc biệt cảm ơn cô Hạnh, chị Thuỷ, Nga, Phương vì đã ở bên động viên và ủng hộ em mỗi lúc em khó khăn, chới với nhất.

Và cuối cùng, con muốn cảm ơn ba mẹ và gia đình vì tình yêu thương vô bờ bến dành cho con.

Paris, November 2020
Võ Thị Phương Thuỷ

Contents

Glossaries, notations and symbols

In this section, I list all of the acronyms, operators and notations I will use throughout this thesis. There may be ones that are only defined here and used later without recall the definition at the place they appear.

Acronyms

Initials	Meaning
AIDS	Acquired Immune Deficiency Syndrome
BP	branching Process
càdlàg	(in French) continue à droite, limite à gauche
CRS	Chain referral sampling
ER	Erdös-Rényi
GWP	Galton-Waston process
HCV	Hepatitis C virus
i.i.d.	indentically and independently distributed
MLE	maximum likelihood estimator
MSM	men who have sex with men

Initials	Meaning
PWID	people who inject drugs
RDS	Respondent Driven Sampling
resp.	respectly
RW	random walk
SAEM	Stochastic Approxiamtion of Expectation-Maximization
SBM	Stochastic Block Model
SSBM	symmetric Stochastic Block Model
w.r.t.	with respect to

Notations

Notations	Meaning	Page
$ER(N,p)$	Erdös-Rényi graph with N vertices, each pair of vertices is connected with probability p.	xi, 11
$X \overset{(d)}{=} Y$	the two random variables X and Y have the same law.	xi, 42
$X_n \xrightarrow{(d)} X$	the sequence $(X_n)_n$ converges in distribution to X.	xi, 66
$\#$	the cardinal of a set.	xi, 102
$\mathbb{N}^*$	the set of strictly positive natural numbers $\{1,2,\ldots\}$.	xi, 4
$\mathbb{N}$	the set of natural numbers $\{0,1,2,\ldots\}$.	xi, 4
$\binom{n}{k}$	n chooses k.	xi, 43
$[\![n]\!]$	the set of $\{1,\cdots,n\}$ for every $n \in \mathbb{N}^*$.	xi, 4
$\mathcal{B}in(n,p)$	the binomial distribution of parameters n and p.	xi, 42
$\mathcal{C}(E,F)$	the space of all continuous functions defined in E, taking value in F.	xi, 25
$\mathcal{C}_b(E,F)$	the space of bounded functions defined in E and taking values in F.	xi, 66

Notations	Meaning	Page
$\mathcal{C}_{(2)}$	the second largest component.	xi, 12
$\mathcal{C}_{max}$	the largest component.	xi, 12
$\mathcal{D}(E, F)$	the Skorokhod space, where each element is a right continuous with left limits function, defined in E and taking value in F.	xi, 25
$\mathcal{W}$	the space of all graphons.	xi, 17
$a \wedge b$	the minimum of two numbers of a and b.	xi, 43
$f(n) \asymp g(n)$	two quantities $f(n)$ and $g(n)$ have the same order as n tends to infinity.	xi, 16

Operators

Operators	Meaning	Page
$\delta_{\square}(\kappa_1, \kappa_2)$	The cut-metric of two graphons κ_1 and κ_2.	xi
$\langle M \rangle$	quadratic variation of the martingale M.	xi
$d_{sub}(G, G')$	The subgraph distance of two graphs G and G'.	xi

Introduction

0.1 Motivations

A random graph is used to describe a discrete structure composed of nodes or vertices linked together by edges in some random ways. Graph or network structure is encountered in various situations and several different scales, from the modeling behaviors of human society to the microscopic particles in our body. It attracts attention of researchers in many fields of science and has increasing importance in applications: **General applications** Newman et al. gave some general studies of random graphs applied in internet, epidemics, cellular networks and genetic networks, food webs, traffic networks (see [99] for details); **Public health** Modelling of hepatitis C virus transmission among people who inject drugs [24]; **Social networks** Modelling the relationships of people in Facebook, or the interactions in twitter,... [71] and many more.

In many applications, exploration of random graphs has been used to gather data, to model and to generate classes of networks that evolve in time like: the flow of information in the Internet, the transmission of disease, the biological evolution, ... This procedure is not only applied in describing the mechanism of a system but is also used to reveal networks which is difficult to observe. One of the main applications in this thesis comes from public health and deals with the propagation of diseases associated with sexual or drug exchanges and sociology: how to explore a population, in which each individual contacts to the others in some way but all the information about this group is hidden due to the illegal behaviors such as people who inject drugs (PWID), men having sex with men (MSM),... Discovering the topology of these social networks may be of primary importance for modeling of the spread of diseases such as Acquired Immune Deficiency Syndrome (AIDS) or hepatitis C virus (HCV) in view of public health issues. We refer to [83, 23] for AIDS or to [24, 25, 53] for HCV, for example. Once the random graph or graphon - a continuous version of the graphs that will be presented in the sequel - are estimated, they can serve in modeling applications (see e.g. the SIS model of [103]).

The exploration process is based on a “peer-to-peer" networking, meaning that from a source of items chosen, the network is explored step by step through the connections between relating nodes. Several methods have been proposed to make use of this feature on the exploration such as: snowball sampling, targeting sampling, chain referral sampling, *etc.*, where respondents recruit their peers [47, 49, 70]. Inherited from the idea of referral chain, Respondent Driven Sampling (RDS, see [49, 50, 51]) was developed as an efficient method of sampling. During the survey, at every wave of respondent, all the information of who recruited whom is kept, which is combined later on with the knowledge of each individual’s connections to reweigh the sample. Henceforth, from a group of initial individuals, the hidden graph is explored step by step by propagating the walkers along its edges.

The networks in real world are very complex: visualizing them, modeling them, understanding them, using them raise new challenges. Here, we are interested in the process of discovering them. Typically, a graph is not known in detail and when we may have only a partial information about it or even it can be totally hidden. For this reason, we are encouraged to look for the suitable exploration algorithm with

the hope that from the initial data, we can capture some features of the underlying network. Under this circumstance, a random graph model is helpful and can serve as benchmark. In this thesis, we restrict the mathematics analysis to two classes of random graphs defined precisely in the next section: the Erdös-Rényi (ER) graph [87, 88, 92, 36, 37, 94] where pairs of vertices are linked independently with a probability $p \in (0, 1)$ and the Stochastic Block Model (SBM) [2, 3, 54] allowing to account for covariates and cluster features in the graph.

In the following of this section, we present the random graphs, some basic models, their important properties which are used in our work. The principle of the RDS is explained in details with the basic notations of following chapters. The main results of this thesis are presented. These results are objects of three papers:

- Respondent Driving Sampling on sparse Erdös-Rényi graph. (in progress)
- Chain referral sampling on Stochastic Block Model. (to be published)
- Estimation of dense stochastic block models visited by a random walk. (submitted)

0.2 Random graph and some basic models

With the network structure, we are interested in modeling objects with pair interactions between them. Each object is represented by a node (or a vertex) and the connections between pairs of nodes are indicated by edges.

0.2.1 Basis of graph theory

All the basic definitions in this section are referred from classic books of *Graph theory*: Bollobás [87, 88], Diestel [91], Van der Hofstad [94]. We briefly recall here some notions being used in this thesis.

In our settings, we work with graphs which are *simple*, *undirected* and without *self-loops*, *i.e.* there is at most one edge between two nodes, there is no order in the pair of vertices describing edges, and there is no edge from one node to itself. The formal definition of a *graph* is given as follows.

Definition 0.1 — Graph. Let V be a countable set and E be a subset of distinct pairs of elements in V. The set $G = (V, E)$ is called a graph of vertices in V and edges in E.

We mainly deal with the *finite* graphs, which means that the set of vertices is finite. We will use the notation

$$[\![1, N]\!] := \{1, \dots, N\}, \quad \forall N \in \mathbb{N}^*, \tag{1}$$

and we enumerate the vertices set as $V = [\![1, N]\!]$, $N \in \mathbb{N}^*$. In a graph G, the vertex set is denoted by $V(G)$ and the edge set is $E(G)$.

Remark 0.1 The maximum number of edges in a simple, undirected and without self-loops graph of N vertices is $\frac{N(N-1)}{2}$.

Adjacency Consider a graph G. If there is an edge connecting a pair of vertices u and $v \in V(G)$, we say u and v are *adjacent* or *neighbors* in G. We denote by $u \sim_G v$ or simply $u \sim v$ if u and v are adjacent, otherwise, we write $u \nsim_G v$ or $u \nsim v$. Thus the set of edges in G is: $E(G) = \{\{u, v\} \in V \times V : u \sim_G v\}$.

A graph G is called *complete* if all the vertices in $V(G)$ are pairwise adjacent. A complete graph of N vertices is denoted by K_N.

When studying a graph, we are interested in the relations between vertices, concerning the appearance of edges. It can be represented mathematically by a squared matrix of size $N \times N$, called *adjacency matrix*.

Definition 0.2 Let $G = (V, E)$ be a graph of size N with $V(G) = [\![1, N]\!]$. The squared matrix $A = (a_{ij})_{N\times N}$ defined by:

$$a_{ij} := \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

is called the *adjacency matrix* of G.

Remark 0.2 In our settings, the adjacency matrix A is necessarily symmetric and all the elements on the diagonal are zeros.

Vertex's degree An important information needed for the study of networks is the number of neighbors to a node.

Definition 0.3 Let $G = (V, E)$ be a graph, and v be a vertex of G. We define the *degree* of v by the number of neighbors of v. It is denoted by $d_G(v)$ or $d(v)$.

Remark 0.3 For every vertex $i \in [\![1, N]\!]$, the degree of i is $d(i) = \sum_{j\in[\![1,N]\!]} a_{ij}$.

Connectivity One of the essential properties that receives the most attention in graph theory is connectivity. It is an important measure for the networks recovering problems.

Definition 0.4 — Path. A *path* of length k is a non-empty graph $P = (V, E)$ of the form: $V = \{v_0, \ldots, v_k\}$. We set $E(P) = \big\{\{v_0, v_1\}, \ldots, \{v_i, v_{i+1}\}, \ldots, \{v_{k-1}, v_k\}\big\}$.

Definition 0.5 — Connected graph. A graph G is called *connected* if any pair of its vertices are linked by a path in G, *i.e.* $\forall u, v \in V(G), \exists u_0, \ldots, u_k \in V(G)$ such that $\big\{\{u, u_0\}, \{u_0, u_1\} \ldots, \{u_i, u_{i+1}\}, \ldots, \{u_k, v\}\big\} \subset E(G)$.

Clearly, the connectivity of vertices in a graph is an equivalence relation, where if $(V_k)_k$ denote the associated equivalence classes and $E_k = \big\{\{u, v\} \in E(G) : (u, v) \in V_k \times V_{k'}\big\}$, then the $G_k = (V_k, E_k)$ are connected graphs. $(G_k)_k$ can be seen as partition of G. We want to give a name to the "sub-part" of a graph.

Definition 0.6 Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. If $V' \subset V$ and $E' \subset E$, we say G' is a *subgraph* of G, written $G' \subset G$. And if G' is a subgraph of G and if $\forall u, v \in V', u \sim_G v \Leftrightarrow u \sim_{G'} v$, then G' is called a *induced subgraph* of G.

A maximal connected subgraph in G is called a *component* of G.

Clearly see that a path in G is a connected subgraph of G. Every node in a component must be adjacent to at least one other node and thus have degree at least 1. Hence evidently, every connected graph having n nodes has at least $n - 1$ edges.

Homomorphism, isomorphism Let $G = (V, E)$ and $G' = (V', E')$ be two finite simple graphs.

1. A function ϕ from $G = (V, E)$ to $G' = (V', E')$, written $\phi : G \to G'$, is a *graph homomorphism* if it is an adjacency preserving map matching every node in V to some node in V', *i.e.* $(u, v) \in E \Rightarrow (\phi(u), \phi(v)) \in E'$.
2. When ϕ is a bijection $\phi : G \to G'$ preserving adjacency, we call ϕ a *graph isomorphism* and G and G' are called *isomorphic* , written $G \cong G'$.

From now on, we have enough basis notions in graph theory field for the next parts of this thesis. Let us move to the probability approach.

0.2.2 Random graphs

A random graph is a graph in which properties such as the number of vertices, graph edges and the connections between them are determined by some random procedure.

Definition 0.7 — Random graph. A random graph $\mathcal{G}$ is a random variable valued in the quotient set of all graphs modulo isomorphism.

Some examples of random graphs

Erdös-Rényi graphs: (see Figure 1) The Erdös-Rényi graph is a simple model of random graphs, which was introduced in the earliest works of Erdös and Rényi [36, 37]: the graph $\mathcal{G}(N, p)$ is generated by linking any pair of N nodes with the same probability p, independently from the other pairs (see Definition 0.9). We give more detailed discussions about this random graph later in Section 0.2.4.

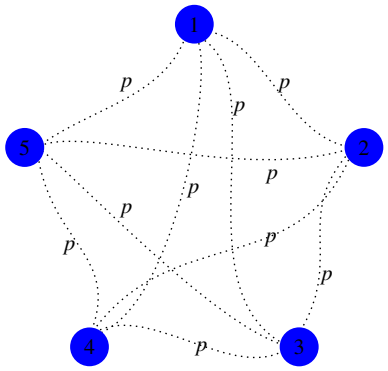


Figure 1. Erdös-Rényi graph

Stochastic Block Models (SBM) : (see Figure 2) An SBM is a generalization of Erdös-Rényi graph, in which (see Definition 0.10)

- the vertices in $[\![1, N]\!]$ are partitioned into a finite number of classes;
- the probability of connecting vertices is no longer equal for every node but depends on the class of each vertex.

This model presents the community structure of a network by the pattern of connections. More precisely, the set of N vertices is partitioned into Q blocks with proportions $\alpha = (\alpha_1, \cdots, \alpha_Q)$ and the probability of find an edge joining a vertex from group ℓ with a vertex of group k is $\pi_{\ell k}$ (see Definition 0.10).

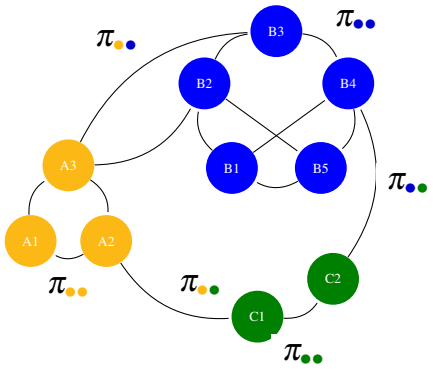


Figure 2. Stochastic Block Model

Configuration model (CM): The third example is

Definition 0.8 Let $N \in \mathbb{N}^*$ and $d = (d_1, d_2, ..., d_N)$ be a sequence of i.i.d. random variables, whose values are non-negative integers such that $\sum_{i=1}^{N} d_i$ is even. Then the Configuration Model (CM) with degree sequence d is a random multigraph with vertices set $V = [\![1, N]\!]$, constructed as follows: (see Figure 3)

- To each vertex i, assign d_i *half-edges*.
- Take a uniform matching of these half-edges.
- For each pair of half-edges in the matching, replace the two half-edges by an edge to obtain the a *multigraph* $CM_N(d)$, in which each vertex i has degree d_i.

Note that the construction as in the definition of $CM_N(d)$ can produce multiple edges and self-loops, meaning that $CM_N(d)$ is a multigraph. Nevertheless, if we assume that the distribution of degree d has finite second moment, then by Durrett [92, Theorem 3.1.2], the number of self-loops and multiple edges are asymptotically independent Poisson random variables as N tends to infinity, which are negligible with respect to the number of edges.

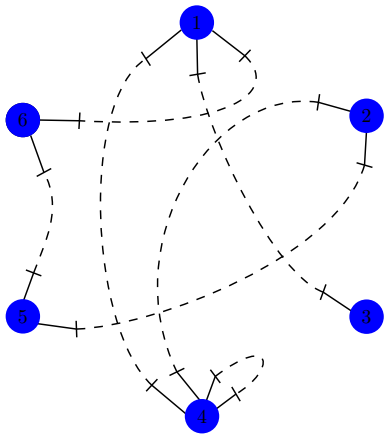


Figure 3. Configuration model with number of vertices $N = 6$ and the degree sequence $d = (3, 2, 1, 4, 2, 2)$.

Apart from those, there are also other models studied for the different purpose of research: the various and related versions of CM and Preferential attachment models (PAM),... are described and discussed in [94, 71, 101].
In this thesis, we focus on the simplest model, Erdös-Rényi graph, and its generalization, Stochastic Block Models.
A usual approach for the connectivity of random graphs is branching process. Let us now review some basis notions and standard results about branching processes, especially Poisson branching processes.

0.2.3 Branching processes

Branching processes serve as a mathematical model for a population evolving in time, where each individual at the n^{th} generation produces a number of individuals for the $n + 1^{th}$ generation and the offspring distribution is the same for every individual in

the population. A common formulation of branching processes is the *Galton-Waston process* (GWP) which is defined by its *offspring distribution* ξ:

- ξ is a random variable taking values in $\mathbb{N}$ with probability distribution $p = (p_k)_{k\in\mathbb{N}}$: $p_k := \mathbb{P}(\xi = k)$ and mean $\mu := \mathbb{E}[\xi] < \infty$;
- $(\xi_{n,i})_{n,i\in N}$, the i.i.d. random variables with the same law as ξ, represent the number of children produced by the member i, $i \in \{1, \dots, Z_n\}$ of generation n, (here, each individual of generation n is indexed by a number $i \in \{1, \dots, Z_n\}$ and the order of indexes is not taken into account);
- Z_n is the number of individuals at generation n, by convention, $Z_0 = 1$. Thus, Z_n satisfies the equality:

$$Z_n = \sum_{i=1}^{Z_{n-1}} \xi_{n,i}. \tag{2}$$

Extinction versus survival probabilities: A population is said to be extinct if at some certain unit of time, there is no more children produced, which means $\exists n_0 \in \mathbb{N}$ such that $Z_n = 0, \forall n \geq n_0$. The fact is that a population may either become extinct or survive forever. We are interested in the question: with what probabilities and under which conditions these events occur. Denote η the *extinction probability*.

$$\eta := \mathbb{P}(\exists n \in \mathbb{N} : Z_n = 0). \tag{3}$$

There is well-known result for the extinction probability of branching processes announced by the following theorem:

Theorem 0.1 — Theorem 3.1 [94]. For a branching process with i.i.d. offspring distribution ξ: $\mathbb{P}(\xi = k) = p_k, k \in \mathbb{N}$ and mean $\mu := \mathbb{E}[\xi] \in [0, +\infty)$, then the extinction probability η is the smallest solution in $[0, 1]$ of equation

$$\eta = G_\xi(\eta),$$

where $G_\xi(s) := \mathbb{E}[s^\xi]$ is the generating function of ξ. Further,

$$\eta = \begin{cases} 1 & \text{if either } (\mu < 1) \text{ or } (\mu = 1 \text{ and } p_1 < 1) \\ \eta_0 < 1 & \text{if } \mu > 1 \\ 0 & \text{if } p_1 = 1 \end{cases}.$$

Depending on the expectation μ of offspring distribution, the branching processes are classified into three regimes:

- subcritical case: $\mu < 1$, the branching process is extinct almost-surely;
- critical case: $\mu = 1$, the branching process is extinct almost-surely if $p_1 < 1$;
- the supercritical case: $\mu > 1$, the branching process survives with probability $1 - \eta$.

Total progeny size: From the equation (2), we see that if a branching process starts with $Z_0 = 1$, the average number of individuals at the n^{th} generation is $\mathbb{E}[Z_n | Z_0 =$

$1] = \mathbb{E}\left[\sum_{i=1}^{Z_{n-1}} \xi_{n,i} | Z_0 = 1\right] = \mathbb{E}[\xi]\mathbb{E}[Z_{n-1}|Z_0 = 1] = \mu\mathbb{E}[Z_{n-1}|Z_0 = 1] < \infty$. By recurrence, we obtain:

$$\mathbb{E}[Z_n|Z_0 = 1] = \mu^n.$$

If $\mu < 1$, $\mathbb{P}(Z_n > 0) \leq \mu^n$. Consequently, when the expected offspring μ satisfies $\mu < 1$, the probability that the population survives up to time n is exponentially small in n.

Conditionally on $Z_0 = 1$, denote

$$T := \sum_{n=1}^{\infty} Z_n$$

the *total progeny* of a branching process (Z_n, ξ). When $\mu < 1$,

$$\mathbb{E}[T] = \mathbb{E}\big[\sum_{n=0}^{\infty} Z_n\big] = \frac{1}{1-\mu} < \infty.$$

Poisson branching processes

When studying the connectivity of the Erdös-Rényi graphs, a specific branching process is utilized for modeling the exploration of graph's component: Poisson branching process, whose offspring distributions is a Poisson random variable with parameter λ. The generating function of the offspring distribution in this case is equal to

$$G_\lambda(s) = \sum_{k=0}^{\infty} s^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{\lambda(s-1)}.$$

Followed by Theorem 0.1, the extinction probability is the smallest solution of equation

$$\eta_\lambda = e^{\lambda(\eta_\lambda - 1)}. \tag{4}$$

Then the survival probability is given by

$$\zeta_\lambda := 1 - \eta_\lambda. \tag{5}$$

For $\lambda < 1$, equation (4) has unique solution $\eta = 1$, which says that the Poisson branching process is almost surely extinct. For $\lambda > 1$, equation (4) has two solutions, of which the smaller is $\eta_\lambda \in (0, 1)$. It means that in this case, both extinction and survival are possibly occurring with non-zero probabilities. Let us look at the branching process conditioned on extinction.

Theorem 0.2 [94, Theorem 3.15] Let $\mu < 1 < \lambda$ be the two real values satisfying equation $\mu e^{-\mu} = \lambda e^{-\lambda}$. The Poisson branching process with mean λ, conditioned on extinction, has the same distribution as a Poisson branching process with mean μ.

Define the large deviation rate function for Poisson random variables with mean λ by

$$I_\lambda = \lambda - 1 - \log(\lambda). \tag{6}$$

The law of the total progeny of a Poisson branching process and its asymptotic behavior is described by the following theorem.

Theorem 0.3 For a Poisson branching process with mean λ, then the total progeny is distributed as:

$$\mathbb{P}(T = n) = \frac{(\lambda n)^{n-1}}{n!} e^{-\lambda n}.$$

Further, as $n \to \infty$,

$$\mathbb{P}(T = n) = \frac{1}{\lambda\sqrt{2\pi n^3}} e^{-I_\lambda n}\big(1 + O(1/n)\big),$$

where I_λ is defined in (6).

Poisson and binomial branching processes When coupling a Poisson branching process with offspring mean λ and a Binomial branching process with parameters n and sucessive probability λ/n, the total progeny of those two processes are related in the following way:

$$\mathbb{P}_{n,\lambda}(T \geq k) = \mathbb{P}_\lambda(T \geq k) + e_n(k, \lambda),$$

where $|e_n(k, \lambda)| \leq k\lambda^2/n$.

0.2.4 Erdös-Rényi graphs and their properties

Definition 0.9 Let $p \in (0, 1)$ and $N \in \mathbb{N}^*$. A random graph $\mathcal{G}$ is called an Erdös-Rényi (ER) graph with distribution, denoted by $ER(N, p)$ if it has N vertices and each pair of nodes $\{i, j\}$ is connected with probability p, independently of the others.

Despite the simplicity of this graph, this model has its own beautiful properties to work on. We refer to the book of Van der Hofstad [94, Chapters 4 and 5] for the detailed study. One of the primary properties studied is the emergence of a giant component. We are interested in a specific class of ER graphs when the local structure is normalized by the system's size N, that is $p = \lambda/N$, where $\lambda \in (0, \infty)$ is a constant and $N \geq \lambda$. There is a sharp threshold for the emergence of giant components:

- sub-critical graphs: $\lambda \in (0, 1)$, the biggest component of $ER(N, \lambda/N)$ has size $O(\log(N))$;
- super-critical graphs: $\lambda \in (1, +\infty)$, the giant component of $ER(N, \lambda/N)$ has size $O(N)$.

Here we refer to the Theorem 4.4 and Theorem 4.5 in [94] (the subcritical case: $\lambda < 1$) and Theorem 4.8 (the supercritical case: $\lambda > 1$) in [94] for basic results

concerning the size of the giant component and the second largest component in the $ER(N, \lambda/N)$ graphs. There are also other results on the giant components for $ER(N, p)$, see [72] for example.

Connectivity, giant component

Denote $\mathcal{C}_{max}$ and $\mathcal{C}_{(2)}$ the largest and the second largest components of $ER(N, \lambda/N)$. There is a phase transition for these quantities when λ varies from subcritical to supercritical regimes.

Subcritical case: For $\lambda < 1$, the following theorem[1] justifies the lower and upper bounds for the largest component in an $ER(N, \lambda/N)$.

Theorem 0.4 — Lower and upper bounds of biggest component. [94, Theorem 4.4 and Theorem 4.5, page 123] For I_λ the large deviation rate function for Poisson random variables with mean λ defined in (6) and for every $a, b > 0$ such that $a < \frac{1}{I_\lambda} < b$, there exist $\delta_a = \delta(\lambda, a)$ and $\delta_b = \delta(\lambda, b)$ such that

$$\mathbb{P}_\lambda(|\mathcal{C}_{\max}| \leq a \log N) = O(N^{-\delta_a}); \tag{7}$$

$$\text{and } \mathbb{P}_\lambda(|\mathcal{C}_{\max}| \geq b \log N) = O(N^{-\delta_b}). \tag{8}$$

A consequence of this theorem is the size of largest component is of order $\log N$.

Proposition 0.1 When $\lambda < 1$, for I_λ defined by (6), we have that

$$\frac{|\mathcal{C}_{max}|}{\log N} \to \frac{1}{I_\lambda} \tag{9}$$

in probability as $N \to \infty$.

Supercritical case: For $\lambda > 1$, there is a constant $\zeta_\lambda > 0$ such that the largest connected component has size approximately $\zeta_\lambda N$, and the second largest component has $O(\log(N))$ vertices. A good approximation result, see [94, Theorem 4.18, page 123] for $|\mathcal{C}_{max}|$ and $|\mathcal{C}_{(2)}|$ is illustrated by

Theorem 0.5 — The giant and the second largest components' sizes. For $\lambda > 1$ and every $\nu \in (1/2, 1)$, there exists $\delta = \delta(\lambda, \nu)$ such that

$$\mathbb{P}_\lambda(|\mathcal{C}_{max} - \zeta_\lambda N| \leq N^\nu) = O(N^{-\delta}), \tag{10}$$

[1] this theorem is a combination of two theorems: Theorem 4.4 and 4.5, see Chapter 4, page 123 in [94].

where $\zeta_\lambda = 1 - \eta_\lambda$ is the survival probability of Poisson branching process with mean offspring λ, η_λ is determined by the smallest solution of equation (4). And the second largest component $\mathcal{C}_{(2)}$ satisfies:

$$\frac{|\mathcal{C}_{(2)}|}{\log N} \longrightarrow \frac{1}{I_{\mu_\lambda}}, \tag{11}$$

in probability as $N \to \infty$, where $\mu_\lambda = \lambda\eta_\lambda$.

The central limit theorem for the giant component's size in the super-critical case is also proved, see [94, Theorem 4.16, page 137].

Theorem 0.6 — Central limit theorem for giant component's size. Fix $\lambda > 1$, then when $N \to \infty$

$$\frac{|\mathcal{C}_{max}| - \zeta_\lambda N}{\sqrt{N}} \longrightarrow Z, \tag{12}$$

in distribution, where Z is a normal random variable with mean 0 and variance $\sigma_\lambda^2 = \frac{\zeta_\lambda(1-\zeta_\lambda)}{(1-\lambda+\lambda\zeta_\lambda)^2}$.

0.2.5 Stochastic Block Models and their properties

The block structure is often encountered in social, physical and other phenomena modeled by the complex networks. It gives rise to the idea of partitioning the whole graph into groups of vertices regarding to the similarity of their connection patterns. From this viewpoint, White, Boorman and Breiger [81] designed a blockmodel to interpret the social structure from the patterns of relations among concrete entities. Based on this deterministic model, Fienberg and Wasserman [38] and then Holland et al. [54] generalized it to a probabilistic version, the Stochastic Block Model (SBM), where the variability of data was taken into consideration. It provides a benchmark for some common tasks such as: community detection, or recovering the patterns of connections in the underlying network [1, 2, 3, 42, 46].

We give below the definition of SBM given by Abbe [1].

Definition 0.10 [1] Let

- N be a positive integer (number of vertices);
- Q be a positive integer (number of blocks or types or classes);
- $\alpha = (\alpha_1, ..., \alpha_Q)$ be a probability distribution on $[\![1, Q]\!]$ (the probabilities on the Q blocks, *i.e.* a vector of $[0, 1]^Q$ such that $\sum_{k=1}^{Q} \alpha_k = 1$);
- Z be an $N-$dimensional random vector of i.i.d. components with distribution α;
- $\pi = (\pi_{k\ell})_{(k,\ell)\in[\![1,Q]\!]^2}$ be a symmetric matrix with entries $\pi_{k\ell} \in [0, 1]$

(connectivity probabilities).
- $[\ell] := \{v \in [\![1, N]\!] : Z_v = \ell\}$ be the block (community) ℓ;
- $N_\ell := |[\ell]|$ be the size of block ℓ, $\ell \in [\![1, Q]\!]$.

The pair (Z, G) is drawn under the distribution SBM(N, Q, α, π) if G is an $N-$vertices $\{1, ..., N\}$ such that each pair of vertices i and j are connected independently of other pairs with probability $\pi_{Z_i Z_j}$.

The SBM is characterized by the number of classes Q, the proportions of each class and the probability matrix of connections.

Remark 0.4 — Remark 3, [2]. By the law of large numbers, almost surely we have

$$\frac{N_\ell}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{Z_i = \ell} \overset{N \to \infty}{\longrightarrow} \alpha_\ell. \tag{13}$$

Remark 0.5 When $Q = 1$, we recover the Erdös-Rényi graph.

A special model of SBM is the *symmetric stochastic block model* (SSBM) , where the inner probabilities are the same for all groups: $\pi_{11} = \cdots = \pi_{QQ} = A$ and different with the outer probabilities, which is $\pi_{\ell k} = B, \forall \ell \neq k$.

Example 0.1 The population of size N is partitioned into $Q = 2$ groups with proportions $(\alpha, 1-\alpha)$ and the probability of connections is given by $\pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$ with $\pi_{12} = \pi_{21}$ by the symmetry of matrix π. A visual representation of SSBM with $N = 100$, $\alpha = (0.3, 0.7)$ and $\pi = \begin{bmatrix} 0.6 & 0.05 \\ 0.05 & 0.6 \end{bmatrix}$ is given in Figure 4.

This is an example of a network divided in two groups, where the members within a group is higher connected than any couple coming from different groups.

Connectivity of SBM

SBM is a generalization of Erdös-Rényi and we also have thresholds for the transition phases of its connectivity. The sparse case corresponds to the case where the probabilities of connections grow proportionally to the graph's size, that is $\pi = (\pi_{\ell k}^N)_{\ell,k} = (\frac{\pi_{\ell k}}{N})_{\ell,k}$, the connectivity depends on the average of degrees.
The following topology properties are proved for the SSBM (see details in [1]).

[2]Credit: This figure is plotted by the package *igraph* of Csardi and Nepusz [28].

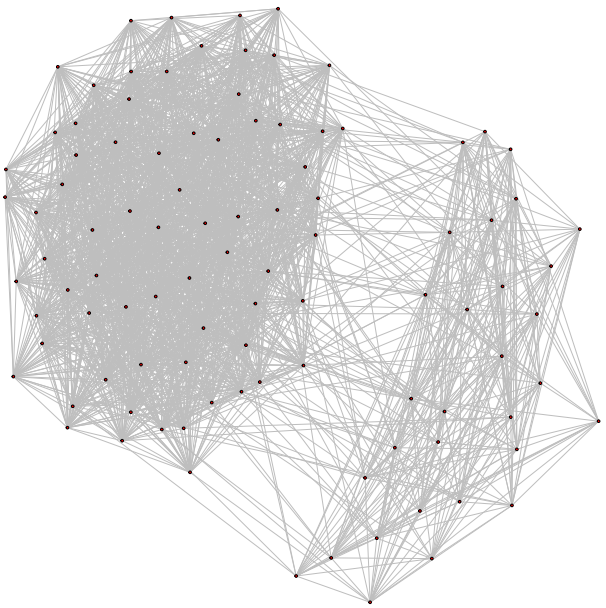


Figure 4. Plot[2]of an SSBM graph of $N = 100$ vertices partitioned into $Q = 2$ classes with proportion $\alpha = (0.3, 0.7)$ and the matrix of connection probabilities $\pi_{11} = \pi_{22} = 0.6, \pi_{12} = 0.05$.

Proposition 0.2 [1, see page 18]

1. For $a, b > 0$, the $SSBM(N, Q, A, B)$ with $A = \frac{a \log N}{N}$, $B = \frac{b \log N}{N}$ is connected with probability $1 - o(N)$ as N tends to infinity if and only if $\frac{a+(Q-1)b}{Q} > 1$ (if a or b is equal to 0, the graph is of course not connected).
2. The $SSBM(N, Q, A, B)$ with $A = \frac{a}{N}, B = \frac{b}{N}$ has a giant component (*i.e.* a component of size linear in N) if and only if $d := \frac{a+(Q-1)b}{Q} > 1$;
3. For $\delta < 1/2$, the neighborhood at depth $r = \delta \log_d N$ of a vertex v in $SSBM(N, Q, A, B)$, where $A = a/N, B = b/N$, tends in total variation to a Galton-Watson branching process of offspring distribution Poisson(d) where $d := \frac{a+(Q-1)b}{Q} > 1$.

If $a = b$, then the SBM collapses with the Erdös-Rényi, and the connectivity properties coincide with the ones we know from the theory of Erdös-Rényi graphs.

0.3 Convergence of subgraphs, graphons

In this thesis, we are interested in large graphs, which means that we aim to study some properties of a graphs family $(G_N)_{N \geq 1}$ when $|V(G_N)|$ tends to infinity as $N \to \infty$. The convergence of graph sequences have been studied for the purpose of understanding the large graphs and their approximations. There have been huge works of Bollobás [87], Janson [56, 101] on the properties of large graphs: degree distributions, the evolution of random graphs for the connectivity, giant components, *etc.*

In a series of papers [14, 15, 16, 39, 67, 68], Lovász and coauthors have developed a beautiful theory of graph limits, which works best for the two extreme cases: *dense graph* where the number of edges in a graph is "close" to its maximum possible edges $\binom{|V(G)|}{2}$, and *sparse graph* where vertices degrees are bounded or at least the average degree is bounded [12, 16].

Definition 0.11 Let $(G_N)_{N \geq 1}$ be a family of graphs such that the size $|V(G_N)|$ tends to infinity as $N \to \infty$.
The family $(G_N)_{N \geq 1}$ is called *dense* if the number of edges is quadratic in their number of vertices. Rigorously, $|E(G_N)| \asymp |V(G_N)|^2$ when $|V(G_N)|$ tends to infinity.
In the other extreme, a family of graphs $(G_N)_{N \geq 1}$ is called *sparse* if the vertex degrees are bounded and $|E(G_N)| \asymp |V(G_N)|$ when $|V(G_N)|$ tends to infinity.

The notions of dense and sparse only make sense for families of graphs whose sizes are sufficiently large, not for a single graph. If a sequence of graphs converges from the left, it means that the graph G_N has more and more similar homomorphic

structure of the every small graph embedded in G_N. For example, the $ER(N,p)$ with p fixed, is a class of dense graphs since the average number of edges is $\frac{N(N-1)}{2}p$. While for $ER(N,\lambda/N)$ for fixed λ, are sparse graph.

Lovás et al. [15, 16, 18] have introduced a notion of convergence using homomorphisms: *convergence from the left*, which is defined in terms of the densities of homomorphisms from small graphs into G_N. Lovás and Szegedy [67] have proved that if a sequence of graphs is convergent "from the left", the limit object is in fact a measurable symmetric function $\kappa : [0,1]^2 \to \mathbb{R}$ that represents the limit density of edges in G_N. This limit object is called *graphon*.

The general graphon was introduced by Lovás and Szegedy, see [67]. Here, we restrict ourselves to non-negative normalized graphons taking value in $[0,1]$.

Definition 0.12 — Graphon. Let $\mathcal{W}$ be the space of all bounded measurable functions $\kappa : [0,1]^2 \to [0,1]$ that are symmetric and integrable. We call the functions in $\mathcal{W}$ *graphons*.

It is quite natural to represent a finite graph G_N in terms of graphon as follows: assume that $V(G_N) = [\![1,N]\!]$, we divide the interval $[0,1]$ into N disjoint intervals $I_1^N, \ldots, I_N^N$, where $I_i^N = \left[\frac{i-1}{N}, \frac{i}{N}\right)$ for every $i \in [\![1, N-1]\!]$ and $I_N^N = \left[\frac{N-1}{N}, 1\right]$,. Define the function κ_{G_N} as

$$\kappa_{G_N}(x,y) = \sum_{i,j \in [\![1,N]\!]} \mathbf{1}_{I_i^N \times I_j^N}(x,y) \mathbf{1}_{i \sim_{G_N} j}.$$

The graph G_N is then associated to the graphon κ_{G_N}.

In fact, the associated graphon κ_{G_N}, however, is not unique for isomorphic unlabeled graphs: if we re-enumerate the vertices in graph G_N, then the associated graphon κ_{G_N} will not be the same. To make the associated graphons "unique", we define the equivalence relation of two graphons as follows: $\kappa, \kappa' \in \mathcal{W}$ are *isomorphic up to a null set* if there is an invertible measure preserving map $\varphi : [0,1] \to [0,1]$ such that $\kappa^\varphi(x,y) := \kappa(\varphi(x), \varphi(y)) = \kappa'(x,y)$ almost everywhere. Then the isomorphism up to a null set is an equivalence relation and the two graphons κ^φ and κ are not "essentially" different.

Then the set of all finite graphs can be embedded in the space of graphons (modulo isomorphism up to a null set) equipped with a suitable topology. As N tends to infinity, the limit of the sequence $(G_N)_{N \geq 1}$ can be interpreted as limit of the associated graphons $(\kappa_{G_N})_{N \geq 1}$. In the left-convergence sense, we consider the space of all graphons (including finite graphs) equipped with the *subgraph distance*, which is given in the sequels.

0.3.1 Subgraph distance, left-convergence of the dense graphs sequence

The convergence from the left was first studied for the dense graphs by Lovás et al. [14, 15, 16, 67] and then extended for the sparse case by Borgs et al. [16, 17, 18]. When saying a graph sequence is convergent from the left, we want to look at the homomorphic structures of every small subgraphs into the sequence of large graphs. The number of "copies" of the "small" graph F into the large graph G is determined by counting the number of *injective homomorphisms* of F into G, denoted by $|\text{inj}(F, G)|$. When we normalize this quantity, it yields the proportion of F found in G. The *injective homomorphism density* of F in G with $|E(F)| = k \leq |V(G)|$ and $V(G) = [\![1, N]\!]$ is defined as

$$t(F, G) := \frac{|\text{inj}(F, G)|}{(N)_k}, \tag{14}$$

where $(N)_k := \frac{N!}{k!} = N(N-1)\dots(N-k+1)$ is the number of injective homomorphisms of F into the complete graph K_N ; and with the convention that if $k > N$, $t(F, G) = 0$.

Now let us classify the set of all finite graphs to isomorphic graphs and enumerate these classes as $(F_i)_{i\geq 1}$, where F_i is the representative of an isomorphism class. We introduce the distance between two graph finite graphs G and G' by

$$d_{sub}(G, G') := \sum_{i\geq 1} \frac{1}{2^i} |t(F_i, G) - t(F_i, G')|.$$

The distance d_{sub} is often called the *subgraph distance*.

It is natural to think of two graphs G and G' seems to be similar if they have similar homomorphism densities. And the notion left-convergence is in fact the convergence of the quantities $t(F, G)$ when $|V(G)|$ tends to infinity.

Definition 0.13 — Lovász et al. [15]. Let $(G_N)_{N\geq 1}$ be a sequence of finite graphs such that $|V(G_N)| \to \infty$ as N tends to infinity. We say that $(G_N)_{N\geq 1}$ is *convergent from the left*, or simply *convergent* if $t(F, G_N)$ converges for any simple graph F.

It turns out (Lovász et al. [15]) that the limiting object of a convergent graphs sequence is a standard kernel represented explicitly by a graphon $\kappa \in \mathcal{W}$. Thus the homomorphism density $t(F, G)$ is naturally extended as the density of a graph F of k vertices on graphon κ as follows:

$$t(F, \kappa) := \int_{[0,1]^k} \prod_{(i,j)\in E(F)} \kappa(x_i, x_j) dx_1 \dots dx_k. \tag{15}$$

And the distance of a graph G to a graphon κ:

$$d_{sub}(G,\kappa) = \sum_{i\geq 1}\frac{1}{2^i}|t(F_i,G) - t(F_i,\kappa)|$$
$$= \sum_{i\geq 1}\frac{1}{2^i}\left|\frac{|\text{inj}(F,G)|}{(|V(G|)_{|E(F)|}} - \int_{[0,1]^{|E(F_i)|}}\prod_{(\ell,\ell')\in E(F)}\kappa(x_\ell,x_{\ell'})dx_1\dots dx_{|E(F)|}\right|.$$

We can say that the graph G is "close" to the graphon κ if for any finite graph F, the proportion of copies of F into G is "close" to the density $t(F,\kappa)$. The following theorem claims that the graphons is a completion of $((F_i)_{i\geq 1}, d_{sub})$, see [67].

Theorem 0.7 — Theorem 3.1 in [15]. Let $(G_N)_{N\geq 1}$ be a dense graph sequence which is Cauchy with respect to d_{sub}. Then there exists a graphon κ such that

$$d_{sub}(G_N,\kappa) \to 0 \tag{16}$$

as N tends to infinity.

Proof. The rigorous proof of the theorem above can be found in [67] using the Szemerédi partitions and the martingale convergence theory. ■

Theorem 0.7 is for the convergent sequence of deterministic graphs. For the random graph, we want to build a model of random graph on N vertices from a graphon κ and see the distribution of these "type" of models.

Given κ a graphon in $\mathcal{W}$ and $X^{(N)} = (X_n)_{n\in[\![1,N]\!]}$ be a sequence of N random variables taking values in $[0,1]$, let us denote $G_N = G(X^{(N)},\kappa)$ the random graph constructed by the fashion as follows: connect i and j in $[\![1,N]\!]$ with probability $\kappa(X_i,X_j)$ independently with other edges.

For the sequence $X^{(N)} = (U_1,\dots,U_N)$ where $(U_n)_{n\in[\![1,N]\!]}$ are i.i.d random variables with uniform distribution in $[0,1]$, then by the law of large number, we have that

$$\lim_{N\to\infty} d_{sub}(G_N,\kappa) = 0 \tag{17}$$

almost surely.

Remark 0.6 For any graphon $\kappa \in \mathcal{W}$, there is a left-convergent sequence of graphs $(G_N)_{N\geq 1}$, for example $G_N = G((U_1,\dots,U_N),\kappa)$, such that $\lim_{N\to\infty} d_{sub}(G_N,\kappa) = 0$.

0.3.2 Cut-distance, convergence in cut-metric of the sparse graphs sequence

For the convergent dense graphs sequence, we have an explicit expression for the limiting object which is a graphon as in the previous theorems. However, most of

the large networks of interest are sparse. And by the notions of left-convergence, all the sparse graphs sequences converge to a zero-graphon, which no longer characterizes the limiting behavior of sparse graphs. For example, the Erdös-Rényi graphs $ER(N, \lambda/N)$ converges to a deterministic graphon $\kappa = 0$. In order to have more interesting limit objects, we think of normalizing the sequence of associated graphons $(\kappa_{G_N})_{N\geq 1}$.

In the work of Bollobás and Riordan [10], they studied different types of metric for sparse graphs: cut distance $\delta_\square$, subgraph distance d_{sub} and partition distance d_{part} (we will not introduce the third metric in this thesis). They built a bridge for the gaps between the two extremes: under a bounded density assumption, graphons remain the appropriate limit objects for the sequences of sparse graphs after rescaling. Their assumption (see [10, Assumption 4.1]) restricts the model setting to the class of sparse graphs G_N that have the edge densities in every subgraphs are all of the same scale. At this extreme, Borgs et al. [12, 13, 17, 18, 19] extended the theory of graphons to handle the sequence of sparse graphs with N vertices and $O(N)$ edges, *i.e.* it covers the case of graphs containing dense spots.

The most important metric used to study sparse graphs is the *cut-metric*. We firstly give the definition of *cut-distance* introduced by Frieze and Kannan [40]:

Definition 0.14 — Cut-distance. The cut-norm of a graphon κ is defined as:

$$\|\kappa\|_\square = \sup_{S,T\subset[0,1]} \left| \int_{S\times T} \kappa(x,y)dxdy \right|$$

Given κ_1 and κ_2 two graphons, let

$$d_\square(\kappa_1, \kappa_2) := \|\kappa_1 - \kappa_2\|_\square.$$

The *cut-metric* of κ_1 and κ_2 is

$$\delta_\square(\kappa_1, \kappa_2) := \inf_\sigma d_\square(\kappa_1^\sigma, \kappa_2), \tag{18}$$

where the infimum ranges over all measure-preserving bijections $\sigma : [0,1] \to [0,1]$ and $\kappa^\sigma(x,y) = \kappa(\sigma(x), \sigma(y))$.

We then have the cut-distance of two graphons κ_1 and κ_2 as:

$$\delta_\square(\kappa_1, \kappa_2) = \inf_\sigma \sup_{S,T\in[0,1]} \left| \int_{S\times T} \big(\kappa_1^\sigma(x,y) - \kappa_2(x,y)\big)dxdy \right|, \tag{19}$$

where the infimum ranges over all the measure-preserving bijections from $[0,1]$ to $[0,1]$.

Since the set of all finite graphs is embedded in the space of graphons, we can define the cut-distance of two graphs G and G' through their associated graphons:

$$\delta_\square(G, G') := \delta_\square(\kappa_G, \kappa_{G'}). \tag{20}$$

We also can define the cut-distance between a graph G and a graphon κ as:

$$\delta_\square(G, \kappa) := \delta_\square(\kappa_G, \kappa). \tag{21}$$

Left-convergence vs convergence in cut-metric It is proved that convergence from the left is equivalent to convergence in the metric $\delta_\square$. Indeed, this assertion is claimed by the fact that the metric space $(\mathcal{W}, \delta_\square)$ is compact (see [15, Proposition 3.6]) and the following theorem

Theorem 0.8 [15, Theorem 2.6] The sequence of finite simple graphs $(G_N)_{N\geq 1}$ is left-convergent if and only if it is a Cauchy sequence in the metric $\delta_\square$.

0.4 Exploration of random graph by the RDS method

The Respondent Driven Sampling (RDS) was first introduced by Heckathorn [49] in a program of prevention of the spread of HIV. The aim of RDS is to detect the identities of hidden individual and study the large-scale structure of a target population. The idea of exploring a (random) graph by random walks is natural and has been investigated in a large literature (e.g. [11, 35]).

Let us first explain the principle of RDS methodology.

0.4.1 RDS description

The sampling process is conducted as follows: from a group of initial recruited individuals, we ask for their contacts in the social network, whom they know can offer some more information. The new contacts collected are invited to participate in the survey and investigators ask them for new referrals. Keep tracing the connections between subjects, we can recruit the subsequent participants. Intuitively, the wave of respondents moves from node to node along the edges connecting them. The explored part of the network, *i.e.* the vertices discovered and the edges used for propagating the RDS, induce a subgraph of the underlying real graph. The information coming from the interviews gives knowledge on other non-interviewed individuals and edges, providing a larger subgraph (which may not be a tree). We aim at understanding this recruitment process from properties of the explored subgraph.

To handle the two sources of randomness, the graph and the exploring process on it are constructed simultaneously. In the graph, each vertex is at either one of the three following states:

- *inactive*: if it has not been contacted for interviews;
- *active*: constituting the next interviewees;

- *off-mode*: if it has been interviewed already.

At the beginning of the survey, all the nodes of the underlying graph are hidden and marked as inactive nodes. We then choose some individuals as seeds of the investigation to do the interview and switch them on active mode. The interviewee is asked to name his/her contacts and then we will choose among the new contacts maximum c people to deliver to each of them a coupon. Every carrier of a coupon is turn into the active state and can come to a private interview to be asked in turn to give the names of his/her peers. Whenever a new person is named, one edge connecting the interviewee and his/her contact is added but they remain inactive until they receive a coupon. After finishing the interview, a maximum number of c new contacts receive one coupon each and are activated. So if the interviewee names more than c people, a number of them are not given any coupon and can be still explored later provided another interviewee mentions them. After that, the node associated to the person who has just been interviewed is switched to off-mode and is no longer recruited again, see Figure 5.

We repeat the procedure of interviewing, referring, distributing coupons until there is no more active vertex in the graph (no more coupon is returned). Each person returning a coupon receives some money as a reward for their participation, and an extra bonus depending on the number contacts that will later return the coupons. Notice that each individual in the population is interviewed just once and we assume here that there is no restriction on the total number of coupons.

In the design of RDS, researcher keep track of the degree of each respondent, meaning that whenever a person is recruited, we know the number of theirs contacts and who they are. When interviewed, respondent reports his/her number of neighbors and select uniformly at random from theirs personal network maximum c new people to be recruited. The respondent unit is chosen in a random way without replacement.

The advantages of RDS have been discussed in several papers of statistics and sociology research, for example in [49, 63, 70, 80]. The key is restricting the maximum number of people to be explored in each wave of respondent will reduce the bias of sampling towards high-degree nodes. This clever idea have helped reduce the dependence of final sample on the initial one after several waves of respondent, which allows for inferring the resulting samples without using an ordinary sampling frame.

Our approach: The quantities we aim at keeping in track are: the degree of each respondent; the number of candidates keeping coupons and the amount of individuals explored by RDS. On the other hand, the random network considered here is structured preferably as not a tree due to the methodology of sampling, which make difficulties in handling the randomness of detection process while the subsequent referrals are chosen without-replacement. In our work, we describe the RDS as a Markov process for two models of random network: Erdös-Rényi and the Stochastic Block Model on the supercritical case of sparse graph. Under the sparsity assumptions, the normalized process of the Markov chain converges in distribution to a deterministic continuous function, which is quite classic. However, our model is considered for the general c and other factors impacting the choice of recruitment

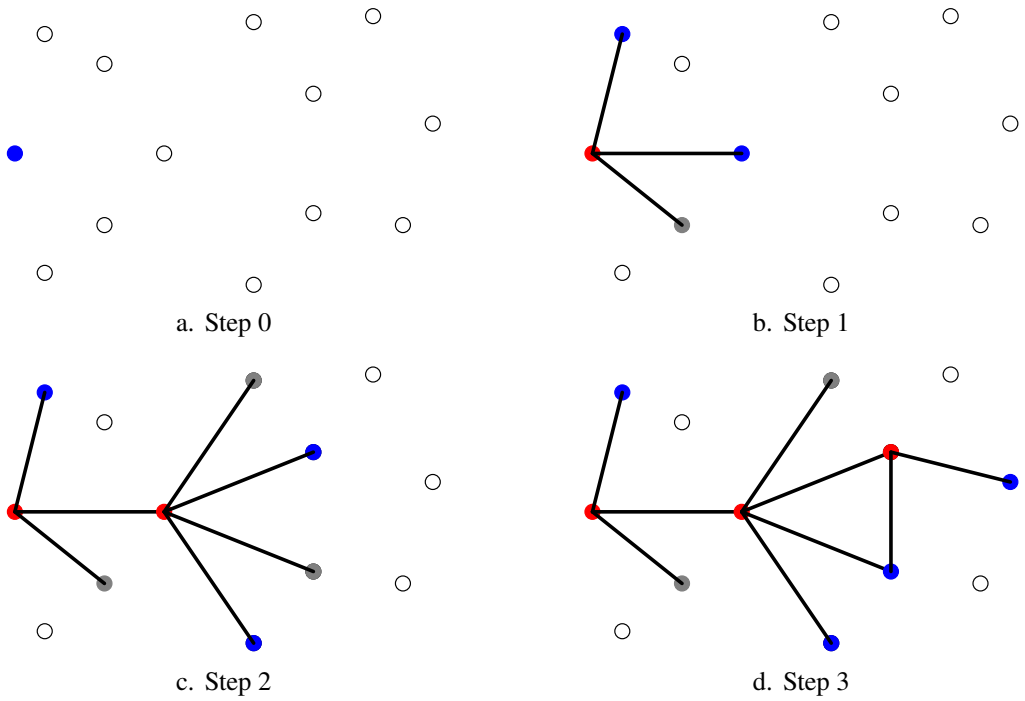


Figure 5. Description of how the RDS works in the case $c = 2$. In our model, the random network and the RDS are constructed simultaneously. For example at step 3, an edge between two vertices who are already known at step 2 is revealed.

units, such as the explored but not treated individuals, are also taken into account. The details of these studies are found in Chapter 1 and Chapter 2 of this thesis.

Mathematical framework

Consider a population of size $N, N \in \mathbb{N}^*$, partitioned in Q classes with proportions $\alpha = (\alpha_1, ..., \alpha_Q)$. Each individual in class k is connected with every node in class ℓ independently of each others with probability $\pi_{k\ell}$. Note that for $Q = 1$, the model induces an ER graph, which has only one type of connections and is studied in Chapter 1. And when $Q > 1$, the model is then an SBM, which is considered in Chapters 2 and 3. Here, we introduce the notations for the general value of Q.

The process of interest counts the number of coupons present in the population. We also want to know how many people are detected, which means that the number of people explored (but without coupons) are also kept track in. Denote by

- $n \in [\![0, N]\!]$ the number of interviews completed;
- $A_n = (A_n^{(1)}, \dots, A_n^{(Q)}) \in \mathbb{N}^Q$ the vector of Q elements, where $A_n^{(\ell)}, \ell \in [\![1, Q]\!]$, is the number of individuals of type ℓ that have received coupons but that have not been interviewed yet (number of active vertices);
- $B_n = (B_n^{(1)}, \dots, B_n^{(Q)}) \in \mathbb{N}^Q$ the vector of Q elements, where $B_n^{(\ell)}, \ell \in [\![1, Q]\!]$, is the number of individuals of type ℓ cited in the interviews but who have not been given any coupon (number of found but still inactive vertices);
- $U_n = (U_n^{(1)}, \dots, U_n^{(Q)}) \in \mathbb{N}^Q$ the vector of Q elements, where $U_n^{(\ell)}, \ell \in [\![1, Q]\!]$, indicates the total number of individuals of type ℓ having been interviewed (number of off-mode nodes).

We define the RDS as the following stochastic process $X_n := (A_n, B_n, U_n), n \in [\![0, N]\!]$:

$$X_n := \begin{pmatrix} A_n \\ B_n \\ U_n \end{pmatrix} = \begin{pmatrix} A_n^{(1)} & \cdots & A_n^{(Q)} \\ B_n^{(1)} & \cdots & B_n^{(Q)} \\ U_n^{(1)} & \cdots & U_n^{(Q)} \end{pmatrix}, \quad n \in [\![0, N]\!].$$

For the more detailed description of $A_n^{(\ell)}, B_n^{(\ell)}, U_n^{(\ell)}$, $\ell \in [\![1, Q]\!]$ is found in the subsequent chapters (Section 1.1, Chapter 1 for $Q = 1$ and Section 2.2, Chapter 2 for $Q \in \mathbb{N}*$ in general). Note that X_n depends on N, but for the sake of simplicity, we do not put the N in the notation.

The main objective of this thesis is to establish an approximation result when the size of the random graph tends to infinity. In this case, the RDS process is correctly renormalized,

$$X_t^N := \frac{1}{N} X_{\lfloor Nt \rfloor} = \left(\frac{A_{\lfloor Nt \rfloor}}{N}, \frac{B_{\lfloor Nt \rfloor}}{N}, \frac{U_{\lfloor Nt \rfloor}}{N} \right) \in [0, 1]^3, \quad t \in [0, 1]. \tag{22}$$

For all N, the process $X_\cdot^N$ lives in the space of càdlàg processes $\mathcal{D}([0, 1], [0, 1]^{3\times Q})$ equipped with Skorokhod topology (see [93, 55, 59]).

In the chapters 1 and 2 of the thesis, we consider spaces $\mathbb{R}^d$ equipped with the L^1-norm defined for $x = (x^1, ..., x^d)$ as $\|x\| = \sum_{k=1}^{d} |x^k|$.

Choice of seeds for the RDS: The RDS is constructed by the similar principle of an epidemic spread and starts with a single individual. There are two main phases of evolution (see [8]): the initial phase is well approximated by a branching process (which we are neglecting here) and the second phase is when the stochastic process is approximated by an deterministic curve. In the chapters 1 and 2, we focus on the second phase: the RDS survey begins with a positive fraction of individuals in the population, *i.e.* the RDS process is conditioned on $\{\lim_{N\to\infty} \|X_0^N\| > 0\}$.

0.4.2 RDS in supercritical Erdös-Rényi graphs

We consider the RDS process on a supercritical ER model $ER(N, \lambda/N)$ ($\lambda > 1$). In this case the model has only one type of vertices, *i.e.* $Q = 1$. Hence, A_n and B_n take values in $\mathbb{N}$, and $U_n = n$ in fact counts the number of steps. Henceforth, in the ER case, it is sufficient that we only consider the process $X_n = (A_n, B_n) \in \mathbb{N}^2$, and the information of U_n is deduced directly.

The normalized process $(X^N)_N$ (defined in (22)) is now written in a simpler form:

$$X_t^N = \frac{1}{N}(A_{\lfloor Nt \rfloor}, B_{\lfloor Nt \rfloor}) = (A_t^N, B_t^N), t \in [0, 1],$$

which is a process in the Skorokhod space $\mathcal{D}([0, 1], [0, 1]^2)$.

For the supercritical ER graphs, Barbour and Reinert [8], the early phase of an RDS can be approximated by a supercritical branching process. Hence if N tends to infinity, when we start with a single individual, after a finite number of steps, we can reach $O(N)$ individuals with a positive probability. Here, we study the behavior of the RDS process under the assumption:

Assumption 0.1 Set $a_0, b_0 \in [0, 1]$ with $a_0 > 0$ and $b_0 = 0$. We assume that the sequence $X_0^N = \frac{1}{N} X_0$ converges in probability to the vector $x_0 = (a_0, b_0)$ as N tends to infinity.

Theorem 0.9 — The case of ER graph. Under the assumption 0.1, when N tends to infinity, the sequence of processes $X^N = (A^N, B^N)$ converges in distribution in $\mathcal{D}([0, 1], [0, 1]^2)$ to a deterministic path $x = (a, b) \in \mathcal{C}([0, 1], [0, 1]^2)$, which is the unique solution of the following system of ordinary differential equations

$$x_t = x_0 + \int_0^t f(s, x_s) ds, \tag{23}$$

where $f(t, x_t) = (f_1(t, x_t), f_2(t, x_t))$ has the explicit formula:

$$f_1(t, x_t) = c - \sum_{k=0}^{c-1}(c-k)p_k(t+a_t) - \mathbf{1}_{a_t>0} \tag{24}$$

$$f_2(t, x_t) = (1-t-a_t-b_t)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t+a_t) - c, \tag{25}$$

with

$$p_k(a_t) := \frac{\lambda^k(1-a_t)^k}{k!}e^{-\lambda(1-a_t)}, \quad k \in \{0, ..., c\}, \tag{26}$$

and c is the maximum value of coupons distributed at each time step.

The main idea of the proof is using limit theory of càdlàg semi-martingale vector processes embedded with Skorokhod topology (see [93]) and Poisson approximations (see [84]). It follows four steps: write the Doob's decomposition of $(X^N)_{N\geq 1}$; study the tightness of the martingale and the finite variation in the decomposition; find the limiting values and finally prove the uniqueness of ODEs' solution.

Proposition 0.3 Let us denote

$$t_0 := \inf\{t \in [0,1] : |a_t| = 0\}. \tag{27}$$

Then $a_t = 0, \forall t \in [t_0, 1]$.

It means that once the a touches 0, a stays at 0. Hence, $t_0 + b_{t_0}$ represents the proportion of explored people in the population.

We also studied the speed of convergence with a central limit theorem for the RDS process on the giant component of ER graph. When we consider the sequence of càdlàg processes $(Y^N)_{N\geq 1}$,

$$W_t^N := \frac{X_{\lfloor Nt \rfloor} - Nx_t}{\sqrt{N}} = \frac{1}{\sqrt{N}}\left((A_{\lfloor Nt \rfloor}, B_{\lfloor Nt \rfloor}) - (Na_t, Nb_t)\right), \quad t \in [0, t_0].$$

Theorem 0.10 — Central limit theorem. The sequence of processes $(W^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}([0,t_0], \mathbb{R}^2)$ to the process $W = (W^1, W^2)$, which satisfies

$$W_t = W_0 + \int_0^t G(s, a_s, b_s, W_s)ds + M(t, a_t, b_t), \quad t \in [0, t_0], \tag{28}$$

where

$$G(t, a, b, W) := \begin{pmatrix} \phi'(t+x)W_t^1 \\ -\lambda(W_t^1 + W_t^2) - \phi'(t+x)W_t^1 \end{pmatrix}; \tag{29}$$

$\phi(z) = c - \sum_{k=0}^{c-1}(c-k)\frac{[\lambda(1-z)]^k}{k!}e^{-\lambda(1-z)}$, $\phi'(z)$ is the derivative of ϕ at z; and M is a zero-mean martingale with the quadratic variation

$$\langle M(\cdot, a_\cdot, b_\cdot)\rangle_t := \left(\int_0^t m_{ij}(s, a_s, b_s)ds\right)_{i,j\in\{1,2\}}, \tag{30}$$

in which

$$m_{11}(t,a,b) := \sum_{k=0}^{c}(c-k)^2 p_k(t+a) - \left(\sum_{k=0}^{c}(c-k)p_k(t+a)\right)^2; \tag{31}$$

$$m_{22}(t,a,b) := \lambda(1-t-a-b) + 2\lambda(1-t-a-b)\left(c(\lambda-1) + \sum_{k=0}^{c} p_k(t+a)\right) + m_{11}(t,a,b); \tag{32}$$

$$m_{12}(t,a,b) := \lambda(1-t-a-b)\left(c(\lambda-1) + \sum_{k=0}^{c} p_k(t+a)\right) - m_{11}(t,a,b). \tag{33}$$

0.4.3 RDS process for the Stochastic Block Model

For the more general model, SBM, see Chapter 2, a convergence theorem similar to Theorem 0.9 for the process $(X^N)_{N\geq 1} \subset \mathcal{D}([0,1],[0,1]^{3\times Q})$ is also proved, but the function f is a more complicated one, and depends on c, Q, α, π.

Assumption 0.2 For each $\ell, k \in [\![1, Q]\!]$, denote $\mu_{\ell k} = \lambda_{\ell k}\pi_k$. We assume that the matrix $\mu = (\mu_{\ell k})_{\ell,k\in[\![1,Q]\!]}$ is *irreducible* and the largest eigenvalue of μ is larger than 1.

Remark 0.7 Under the Assumption 0.2, from the proof of Theorem 3.2 of Barbour and Reinert [8], the early stages of the RDS can now be approximated by a multitype branching process with the offspring distributions determined by the matrix μ. Thanks to the Assumption 0.2 the multitype branching process associated with the offspring matrix μ is supercritical. The analogous results for the extinction probability and for the number of offspring at the n^{th} generation as in the single branching process have been proved in Chapter 5 of [82]: the mean matrix of the population size at time n is proportional to μ^n. And follow the claim (3.11) of Barbour and Reinert [8], we can deduce that if N tends to infinity, when we start with a single individual, then after a finite number of

steps, we can reach a positive fraction of explored individuals in the population with a positive probability.

Assumption 0.3 Let $a_0, b_0, u_0 \in [0,1]^Q$, $a_0 = (a_0^{(1)}, \ldots, a_0^{(Q)})$ such that $\sum_{i=1}^Q a_0^{(i)} = \|a_0\| \in [0,1]$, and set $b_0, u_0 \in [0,1]^Q$, with $b_0 = (0, \ldots, 0)$ and $u_0 = (0, \ldots, 0)$. We assume that the sequence $X_0^N = \frac{1}{N} X_0$ converges in probability to the vector (a_0, b_0, u_0), as $N \to +\infty$.

It means that the initial number of individuals with type ℓ at the beginning of the survey is approximately $\lfloor a_0^{(\ell)} N \rfloor$. A possible way to initializing the process is to draw A_0 from a multinomial distribution $\mathcal{M}(\lfloor \|a_0\| N \rfloor; \pi_1, \ldots, \pi_Q)$.

Theorem 0.11 Under the assumptions 0.2 and 0.3, we have: when N tends to infinity, the process $(X_\cdot^N)_N$ converges in distribution in $\mathcal{D}([0,1],[0,1]^{3\times Q})$ to a deterministic vectorial function $x = (x_\cdot^{(\ell)})_{1\le\ell\le m} = (a_\cdot^{(\ell)}, b_\cdot^{(\ell)}, u_\cdot^{(\ell)})_{1\le\ell\le Q}$ in $\mathcal{C}([0,1],[0,1]^{3\times Q})$, which is the unique solution of the system of differential equations

$$x_t = x_0 + \int_0^t f(x_s) ds, \tag{34}$$

where $f(x_s) := (f_{i\ell}(x_s))_{\substack{1\le i\le 3 \\ 1\le\ell\le Q}}$ has an explicit formula described as follows. Denote

$$t_0 := \inf\{t \in [0,1] : a_t^{(1)} + \ldots + a_t^{(Q)} = 0\}. \tag{35}$$

For $s \in [0, t_0]$,

$$f_{1\ell}(x_s) = \sum_{k=1}^Q \frac{a_s^{(k)}}{\|a_s\|} \frac{\lambda_s^{k,\ell}}{\Lambda_s^k} \left(c - \sum_{h=0}^c (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right) - \frac{a_s^{(\ell)}}{\|a_s\|}; \tag{36}$$

$$f_{2\ell}(x_s) = \sum_{k=1}^Q \frac{a_s^{(k)}}{\|a_s\|} \mu_s^{k,\ell} - \sum_{k=1}^Q \frac{a_s^{(k)}}{\|a_s\|} \frac{\lambda_s^{k,\ell}}{\Lambda_s^k} \left(c - \sum_{h=0}^c (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right); \tag{37}$$

$$f_{3\ell}(x_s) = \frac{a_s^{(\ell)}}{\|a_s\|}; \tag{38}$$

with

$$\lambda_s^{k,\ell} := \lambda_{k\ell} \left(\pi_\ell - a_s^{(\ell)} - u_s^{(\ell)} \right); \quad \Lambda_s^k := \sum_{\ell=1}^m \lambda_s^{k,\ell} \tag{39}$$

and $\mu_s^{k,\ell} := \lambda_{k\ell}(\pi_\ell - a_s^{(l\ell)} - b_s^{(\ell)} - u_s^{(\ell)})$.

For $s \in [t_0, 1]$, $f(x_s) = f(x_{t_0})$.

Notice that in this model, the time corresponds to the fraction of the population interviewed. The time t_0 is the first time at which $\|a_t\|$ reaches 0 and can be seen

as the proportion of the population interviewed when there is no more coupon to keep the CRS going. Necessarily, $t_0 \leq 1$. We see that $\|a_t\| = 0$ only if $a_t^{(1)} = \ldots = a_t^{(Q)} = 0$. It implies that $f(x_t) = 0, \forall t \in [t_0, 1]$. Then, the solution of the system of ODEs (0.11) becomes constant over the interval $[t_0, 1]$.

0.4.4 RDS and Statistics

Existing RDS estimation

RDS was first used as a sampling method for estimating the size of a hidden population. The constitution of a sample S is the number of sampled individuals n_S, the rest of the population being unknown. There are several estimators being proposed from the RDS data such as Salganik and Heckathorn [77], Volz and Heckathorn [80]. In their works, they assume that at each interview stage, respondent chooses only one person to distribute coupon (*i.e.* $c = 1$ in the description above). And the replacement is allowed, *i.e.* one subject might be recruited many times. Also, the underlying network is supposed to be connected, which means that everybody in the population can be reached by a finite path. When we distribute only one coupon at each time we interview someone, *i.e.* $c = 1$, the RDS process can be modeled as a random walk on a graph and the time scale is counting by the number of interviews taking place.

Denote Z_i some real-valued variable of interest measured on the i^{th} individual of the sample S (for example the degree of i). A general estimator form of Horvitz-Thompson for estimating the average of Z is:

$$\mathbb{E}[Z] \approx \frac{\sum_{i \in S} p_i^{-1} z_i}{\sum_{i \in S} p_i^{-1}}, \tag{40}$$

where $p_i = \mathbb{P}(i \in S)$ is the inclusion probability of individual i and $(z_i)_{i \in S}$ is a realization of Z. Based on this general result, Volz and Heckathorn [80] have given an estimator when the data S is sampled by an RDS survey. This estimator relies heavily on the estimation of inclusion probability, which is estimated through the subjects' degrees:

$$p_i \approx \frac{d_i}{\sum_{j \in S} d_j}. \tag{41}$$

When $(Z_i)_{i \in S}$ is the sequence of vertex's degree in the sample S, then the estimation (40) becomes

$$\mathbb{E}[Z] \approx \frac{n_S}{\sum_{i \in S} d_i^{-1}}, \tag{42}$$

which recovers the estimation result for the average degree of Salganik and Heckathorn [52].

In practice, the inclusion probability is complicated to compute because of the fact that the sampling is without replacement and the dependencies between subjects are difficult to control. Some numerical calculations have followed to handle these difficulties, see *e.g.* [44, 45, 70]. Gile [43] proposed an improved estimator for

population means taking into account the without replacement sampling, and Rohe established critical threshold for the design effects [65]. Because of the privacy restrictions, a lot of information is missing in the RDS data. A variant of estimation methods have been developed to infer the graph's structure, see *e.g.* [26, 27].

These works have focused on estimating quantity such as the size of hidden population for the general graph's structure. Here, we are interested in studying the topology of hidden graphs. We desire to give a rigorous study of the of the RDS process: the important quantities collected by RDS such as the number of explored individuals in a hidden population (studied in chapters 1 and 2); the topology of a particular random graph model (in chapter 3). The advantage of my work in this thesis is that: the RDS process is described for the general value c of maximum number of coupons distributed at each wave of respondent; the with-replacement is taken into account; the explored sub-graphs' structure is also kept in track. However, also due to the lack of information, we have to restrain ourselves to some more particular structures, in this thesis, the Erdös-Rényi and more general the Stochastic Block Model.

Topology of sub-graphs explored by the RDS on an SBM graphon

The topology of the sub-graphs induced by the RDS has attracted attention for studies of random network. In the case of graphons, we can consider that the topology is given by the knowledge of the function κ. For SBMs, this function is described by a vectors of parameters and we will focus on this case in what follows. In the works of Athreya and Röllin [4, 5], they have established the convergence for induced subgraphs constructed by an RDS process. The underlying network considered is a random graph defined through a deterministic object: graphon. Athreya and Röllin have built rigorous theory of RDS on the two extremes of random graphs: the dense graphs [4] and the case where degrees grow, but sublinearly [5].

For the sparse graphs, as we see in the limit Theorem 0.9: with $\lambda > 1$, the giant component of the hidden graph is explored by the RDS process. In chapter 3 of this thesis, we focus on the dense case, in particular, a dense SBM. The aim is to develop estimator for the the parameter $\theta = (\alpha, \pi)$. Let us have an overview on the results of Athreya and Röllin [4] for general dense graphs.

Let $X^{(n)} = (X_1, \ldots, X_n), n \in \mathbb{N}^*$, be a vector of random variables taking values in $[0, 1]$; κ be a graphon in $\mathcal{W}$. Assume that there is a probability measure m on $[0, 1]$ such that the following two conditions hold:

Assumption 0.4 (i) for all bounded and measurable functions f, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int_0^1 f(x) m(dx) \tag{43}$$

almost surely;

(ii) $\kappa \in \mathcal{W}$ is continuous $m \times m-$almost everywhere.

Let H_n be the path whose nodes are the n vertices visited by the random walk $X^{(n)} = (X_1, \ldots, X_n)$ as in the Assumption 0.4. Then H_n is a graph with the set of

vertex if $V(H_n) = [\![1, n]\!]$ and the set of edges is $E(H_n) = \cup_{i=1}^{n-1}\{i, i+1\}$, where each vertex $i \in [\![1, n]\!]$ is associated with the value $X_i \in X^{(n)}$.

Note that we do not know the size of underlying graph (N is unknown), and n indicates the n^{th} step of the random walk. We denote $G_n = G(X^{(n)}, H_n, \kappa)$ the random graph, which is completed from H_n w.r.t. graphon κ by the following manner:

Definition 0.15 Define the completion of H_n by graph $G_n = G(X^{(n)}, H_n, \kappa)$ as follows:

- $V(G_n)$ is the same set of vertices as $V(H_n) = [\![1, n]\!]$;
- if i and j are connected in H_n, then connect i and j in G_n;
- if i and j are not connected in H_n, an edge between i and j is included in G_n with probability $\kappa(X_i, X_j)$.

It turns out that the limit object is not the original graphon κ but its transformation by the generalized inverse of the cumulative distribution of m.

Theorem 0.12 [4, Theorem 2.1 and Corollary 2.2] Let $X^{(n)} = (X_1, \ldots, X_n)$ be a sequence taking values in $[0, 1]$, $\kappa \in \mathcal{W}$ be a graphon, m be a probability measure and let G_n be the graph defined in Definition 0.15. Under the Assumption 0.4 and when the number of edges in G_n is proportional to n^2,

$$\lim_{n\to\infty} d_{sub}(G_n, \kappa_{\Gamma^{-1}}) = 0 \tag{44}$$

almost surely, where $\Gamma(x) := m([0, x])$ is the cumulative distribution of m, the generalized inverse of Γ is given by

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

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

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

This result confirms that when sample data are collected from an RDS on a graphon, there is some bias on the resulting data towards high degree items. And they have shown how the RDS as above procedures bias the network. The limiting object gives a framework for the estimation of interesting quantities.

Example 0.2 Let κ be a graphon taking constant value $p \in [0, 1]$ and $(U_1, \ldots, U_n)$ be a sequence of i.i.d uniform random variables in the interval $[0, 1]$. The sequence of graphs $G_n = ((U_1, ..., U_n), \kappa)$ is constructed by connecting i and j with probability $\kappa(U_i, U_j) = p$. Then G_n is an Erdös-Rényi graph: $G_n = ER(n, p)$. By Theorem 0.12, G_n converges in d_{sub} to the graphon $\kappa(x, y) = p$ almost surely.

Statistic results

Using the convergence result (Theorem 0.12) for the dense graph of Athreya and Röllin [4], we develop an empirical estimator for the SBM graphs explored by RDS process, sampled from an ergodic process $X^{(n)}$ under the similar conditions in [4, Theorem 2.1]. Of course, there are biases but we can recover the true parameter since the estimation is transformed from the original ones by a measure that can be estimated as well.

Let $\alpha = (\alpha_1, \ldots, \alpha_Q) \in (0,1)^Q$ be a probability vector, *i.e.* $\sum_{q=1}^{Q} \alpha_q = 1$; $\pi = (\pi_{qr})_{q,r\in[\![1,Q]\!]} \in [0,1]^{Q\times Q}$ be a symmetric $Q \times Q-$matrix whose entries take values in the interval $[0,1]$. In all the sequels, we denote by θ the vector of parameters:

$$\theta := (\pi, \alpha) = (\pi_{qr}, \alpha_q)_{q,r\in\{1,\cdots,Q\}}.$$

We define $I = (I_1, \ldots, I_Q)$ a partition of $[0,1]$, where

$$I_q = \left[\sum_{k=1}^{q-1} \alpha_k, \sum_{k=1}^{q} \alpha_k\right), \qquad q \in [\![1, Q]\!]. \tag{45}$$

Let κ be a graphon given by

$$\kappa(x,y) = \sum_{q=1}^{Q}\sum_{r=1}^{Q} \pi_{qr}\, \mathbf{1}_{I_q}(x)\mathbf{1}_{I_r}(y). \tag{46}$$

Consider the SBM graphon having the form as in equation (46). Denote $Z = (Z_1, \ldots, Z_n)$ the types of each node in the Markov chain: $Z_i \in [\![1, Q]\!], 1 \le i \le n$ and $Y = (Y_{ij})_{1\le i,j\le n}$ the adjacency matrix of the graph $G_n = G(X^{(n)}, H_n, \kappa)$. We want to estimate the SBM parameters α and π in two cases: the types Z of the nodes are observed and unobserved.

Assumption 0.5 We assume that κ is *connected*, *i.e.* for all measurable subset $A \subset [0,1]$ such that $|A| \in (0,1)$,

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

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:

$$\begin{aligned} N_n^{q\leftrightarrow r} &= \#\big\{\{i,j\} \quad | \quad X_i, X_j \in X^{(n)},\ Z_i = q,\ Z_j = r,\ Y_{i,j} = 1\big\}; \\ (\textit{resp. } N_n^{q\nleftrightarrow r} &= \#\big\{\{i,j\} \quad | \quad X_i, X_j \in X^{(n)},\ Z_i = q,\ Z_j = r,\ Y_{i,j} = 0\big\}) \end{aligned}$$

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

When the types are observed: When $(Z_i, 1 \le i \le n)$ are observed, the likelihood of complete observations has the form:

$$\mathcal{L}(Z, Y, X, \theta) = \prod_{q=1}^{Q} \left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}} (1-\pi_{qq})^{N_n^q(N_n^q-1)/2}$$

$$\times \prod_{q\neq r}\left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}} (1-\pi_{qr})^{N_n^q N_n^r} \times \prod_{q=1}^{Q} \frac{\alpha_q^{N_n^q}}{(\sum_{q'=1}^{Q}\pi_{qq'}\alpha_{q'})^{N_n^q - \mathbf{1}_{Z_n=q}}}. \tag{47}$$

Proposition 0.4 The maximum likelihood estimator (MLE) $\widehat{\theta} = (\widehat{\pi}, \widehat{\alpha})$ is the solution of the system:

$$\frac{N_n^q}{\widehat{\alpha}_q} - \sum_{p=1}^{Q} \frac{N_n^p \widehat{\pi}_{pq}}{\sum_{q'=1}^{Q}\widehat{\pi}_{pq'}\widehat{\alpha}_{q'}} = 0; \tag{48}$$

$$\frac{N_n^{q\leftrightarrow r}}{\widehat{\pi}_{qr}} - \frac{N_n^{q\nleftrightarrow r}}{1-\widehat{\pi}_{qr}} - N_n^q \frac{\widehat{\alpha}_r}{\sum_{q'=1}^{Q}\widehat{\pi}_{qq'}\widehat{\alpha}_{q'}} = 0. \tag{49}$$

We want to compare this MLE with the new estimator developed from the convergence result of Athreya and Roellin obtained by Theorem 0.12.

Suppose that the limit object of the sequence $G_n = G(X^{(n)}, H_n, \kappa)$ is an SBM graphon of blocks proportions $\gamma = (\gamma_1, ..., \gamma_Q)$ and the connection probabilities $\rho = (\rho_{qr})_{q,r\in\{1,...,Q\}}$. It means that the SBM graphon is:

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

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

$$J_q = \left[\sum_{k=1}^{q-1}\gamma_k, \sum_{k=1}^{q}\gamma_k\right), \qquad q \in [\![1, Q]\!]. \tag{50}$$

The empirical estimator for χ_∞ is given by:

Definition 0.16 Denote by

$$\widehat{\gamma}_q^n := \frac{N_n^q}{n}; \quad \widehat{\rho}_{qr}^n := \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r} \quad \text{for} \quad q \neq r \quad \text{and} \quad \widehat{\rho}_{qq}^n := \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q - 1)}. \tag{51}$$

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

$$\widehat{\chi}_n(x,y) := \sum_{q=1}^{Q}\sum_{r=1}^{Q} \widehat{\rho}_{qr}^n \mathbf{1}_{J_q^n}(x)\mathbf{1}_{J_r^n}(y), \tag{52}$$

with $J_q^n = \left[\sum_{k=1}^{q-1}\widehat{\gamma}_k^n, \sum_{k=1}^{q}\widehat{\gamma}_k^n\right)$, $q \in [\![1, Q]\!]$.

We define the empirical cumulative distribution of m:

$$\Gamma_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \le x} \qquad \text{and} \qquad \Gamma_n^{-1}(y) = \inf\left\{x \in [0,1] \ : \ \Gamma_n(x) \ge y\right\}. \tag{53}$$

The consistency of $\widehat{\rho}$ is claimed by the following proposition.

Proposition 0.5 Under Assumption 0.5,
(i) $\widehat{\rho}$ is a consistent estimator of π, and for $q, r \in [\![1, Q]\!]$,

$$\lim_{n\to+\infty} \widehat{\rho}^n_{qr} = \pi_{qr}; \qquad \lim_{n\to+\infty} \widehat{\gamma}^n_q = \Gamma\left(\sum_{\ell=1}^{q} \alpha_\ell\right) - \Gamma\left(\sum_{\ell=1}^{q-1} \alpha_\ell\right) =: \gamma_q.$$

It follows that a consistent estimator of α_q is

$$\widehat{\alpha}^n_q = \Gamma_n^{-1}\left(\sum_{\ell=1}^{q} \widehat{\gamma}^n_\ell\right).$$

(ii) In the special case of $Q = 2$, let us denote $(\alpha_1, \alpha_2) = (\alpha, 1-\alpha)$, $\widehat{\alpha}^n = \widehat{\alpha}^n_1$ and $\widehat{\gamma}^n = \widehat{\gamma}^n_1$. An estimator of α is $\widehat{\alpha}^n = \Gamma_n^{-1}(\widehat{\gamma}^n)$.

When the types are unobserved: When $(Z_i, 1 \le i \le n)$ are unobserved, the quantities $N^q_n, N^{q\leftrightarrow r}_n$ are intractable. Then $\mathcal{L}(X, Y, Z, \theta)$ can not be used for estimating θ. In this situation, we can use the stochastic approximation by EM algorithm (SAEM, see [31, 32]) to generate the types Z and the estimation of θ is now the value at which the conditional likelihood $\mathcal{L}(\theta|X, Z, Y)$ attains its maximum.

The SAEM alternates the E-step and M-step as follows:

- Initialization: set the initial values $\theta^{(0)}$ and define the quantity $\mathcal{Q}^{(0)}(\theta) := \mathbb{E}[\log \mathcal{L}(Z, Y, \theta^{(0)})]$.
- At iteration k of algorithm:
  - **E-step:**
    - Simulation: draw the non-observed data $Z^{(k)}$ with the conditional distribution $q(\cdot\,|Y, \theta^{(k-1)})$;
    - Stochastic approximation: update the quantity

$$\mathcal{Q}^{(k)}(\theta) = \mathcal{Q}^{(k-1)}(\theta) + s_k\left(\log \mathcal{L}(Z^{(k)}_i, Y_{ij}, \theta) - \mathcal{Q}^{(k-1)}(\theta)\right), \tag{54}$$

    where $(s_k)_{k\in\mathbb{N}}$ is a positive decreasing step sizes sequence: $\sum_{k=1}^{\infty} s_k = \infty$ and $\sum_{k=1}^{\infty} s_k^2 < \infty$.
  - **M-step:** Choose $\theta^{(k)}$ to be the value of θ that maximizes $\mathcal{Q}^{(k)}$

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{Q}^{(k)}(\theta). \tag{55}$$

Choosing the appropriate types Z is also a complicated work. We use the variational

approach (see Latouche et al. [64]) to generate the data for Z at each iteration k. Suppose that the distribution of Z_i is $\mathcal{M}\big(Q;(\tau_{i1},\ldots,\tau_{iQ})\big)$.

The likelihood $\mathcal{L}(Y,\theta)$ of the incomplete data is

$$\mathcal{L}(Y_{ij};i,j\in[\![1,n]\!];\theta)=\sum_{q_1,\cdots q_n=1}^{Q}\Big[\prod_{i=1}^{n}\mathbf{1}_{Z_i=q_i}\frac{\prod_{i=1}^{n}\alpha_{q_i}}{\prod_{i=1}^{n-1}\sum_{q=1}^{Q}\pi_{q_iq}\alpha_q}\times\prod_{i,j:X_i,X_j\in(X_n)}b(Y_{ij},\pi_{q_iq_j})\Big], \tag{56}$$

where $b(Y_{ij},\pi_{q_iq_j})=\pi_{q_iq_j}^{Y_{ij}}(1-\pi_{q_iq_j})^{1-Y_{ij}}$. A lower bound $\mathcal{J}(R_{Y,\theta})$ of $\mathcal{L}(Y,\theta)$ is:

$$\mathcal{J}(R_{Y,\theta}):=\mathcal{L}(Y,\theta)-\mathrm{KL}(R_{Y,\theta}(Z),\mathcal{L}(Z|Y,\theta)), \tag{57}$$

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 the maximum value of $\mathcal{L}(Y,\theta)$.

Then we have

Proposition 0.6 Given α,π, the optimal parameter

$$\hat{\tau}:=\arg\max_{\tau}\mathcal{J}(R_{Y,\theta}), \tag{58}$$

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

$$\tau_{iq}\propto\frac{\alpha_q}{\sum_{\ell=1}^{Q}\pi_{q\ell}\alpha_\ell}\prod_{i\neq j}\prod_{\ell=1}^{Q}b(Y_{ij},\pi_{q\ell})^{\tau_{j\ell}}. \tag{59}$$

0.5 Presentation of the main results

The main results in this thesis are given in 3 Chapters:

Chapter 1: The RDS process on Erdös-Rényi graph In this chapter, we give the expression of quantities of interests concerned the RDS: the degree of each vertex and the coupons distributed at each stage of interview. The model considered is an Erdös-Rényi graph in the super-critical case: the probability of connecting each edge is λ/N, where N is the size of population, $\lambda>1$. In this sparse case, the size of the giant component is close to N with probability tends to 1 as N tends to infinity and thus we can derive the RDS in wide scope with very few times of disruptions due to the disconnections within population. The main result in this chapter is the convergence of the normalized Markov process to a deterministic continuous function, proved in Theorem 0.9. We also study the central limit theorem for the RDS process given by Theorem 0.10. This work is in progress with Anthony

Cousien[3], Jean-Stéphane Dhersin[4] and Viet Chi Tran[5].

Chapter 2: The RDS process on SBM In this chapter, we generalized the result in chapter 1 for the Markov chain obtained from deriving RDS process on an SBM(N, Q, α, π), where N is the size of population, Q is the number of blocks, $\alpha = (\alpha_1, \ldots, \alpha_Q)$ is the proportion of each block and $\pi = (\pi_{k\ell})_{k,\ell \in [\![1,Q]\!]}$ is the matrix representing the probabilities of connecting nodes across blocks. The generalized result is deduced and extended from the convergence theorem in Chapter 1. We also conclude that the normalized process converges in distribution to a deterministic continuous vectorial-function. The proof follows the similar strategy as in the previous chapter, but with the more carefulness and complexity in treatment. This work has been submitted to ESAIM.

Chapter 3: Estimation of SBM by RDS process in the associated graphon In this chapter, we estimate the parameter of an SBM from the RDS data constructed by a graphon. We consider two situations: the types $(Z_1, \ldots, Z_n)$ are observed and unobserved. For each case, we give some statistic results to compare the "classic" estimator by the maximum likelihood with the new estimators developed from the result of Athreya and Röllin [5]. This work is submitted in collaboration with Viet Chi Tran.

[3]French Institute of Health and Medical Research
[4]University Sorbonne Paris Nord
[5]University Paris Gustave Eiffel

II Main results

1. The RDS process on supercritical Erdös-Rényi graphs

Contents

The work in this chapter is in progress in collaboration with Anthony Cousien[1], Jean-Stéphane Dhersin[2] and Viet Chi Tran[3].

1.1 Introduction

Discovering the topology of social networks for hard to reach populations like people who inject drugs (PWID) or men who have sex with men (MSM) may be of primary importance for modeling the spread of diseases such as AIDS or HCV in view of

[1]French Institute of Health and Medical Research
[2]University Sorbonne Paris Nord
[3]University Paris Gustave Eiffel

public health issues for instance. We refer to [83, 23, 41, 73, 74] for AIDS or to [24, 25, 53] for HCV, for example. To achieve this in cases where the populations are hidden, it is possible to use chain-referral sampling methods, where respondents recruit their peers [47, 49, 70]. These methods are commonly used in epidemiological or sociological survey to recruit hard to reach populations: the interviewees (or ego) are asked about their contacts (alters), where the term "contact" depends on the study population (injection partners for PWID, sexual partners for MSM ...) and some among the latter are recruited for further interviews. In one of the variant, Respondent Driven Sampling (RDS, see [27, 43, 48, 49, 65, 80]), an initial set of individuals are recruited in the population (with possible rules) and each of them is given a certain number of coupons. The coupons are distributed by recruited individuals to their contacts. Either the contacts receiving coupons are chosen at random (which is the case considered in this thesis) or the choice can be guided by information brought by the interviewees (people who are more likely to respond or who are known to have more contacts). The coupon holders come to take an interview and receive in turn coupons to distribute etc. The information of who recruited whom is kept, which, in combination with the knowledge of the degree of each individual, allows to re-weight the obtained sample to compensate for the fact that the sample was not collected in a completely random way. A tree connecting egos and their alters can be produced from the coupons. Additionally, it is also possible to investigate for the contacts between alters - which is a less reliable information since obtained from the ego and not the alters themselves. This provides a network that is not necessarily a tree, with cycles, triangles etc. For PWID populations in Melbourne, Rolls et al. [75, 76] have carried such studies to describe the network of PWID who inject together. The results and the impacts from a health care point of view on Hepatitis C transmission and treatment as prevention are then studied. A similar study on French data is currently in progress [58].

In this chapter, we consider a population of fixed size N that is structured by a social static random network $G = (V, E)$, where the set V of vertices represents the individuals in the population and $E \subset V^2$ is the set of non-oriented edges *i.e.* the set of couple of vertices that are in contact. Although the graph is non-oriented, the two vertices of an edge play different roles as the RDS process spreads on the graph. At the beginning, there is one individual chosen and interviewed. He or she names their contacts and then receives a maximum of c coupons, depending on the number of their contacts and the number of the remaining coupons to be distributed. If the degree D of the individual is larger than c, c coupons are distributed uniformly at random to c people among these D contacts. But when $D < c$, only D coupons are distributed. We assume here that there is no restriction on the total number of coupons. In the classical RDS, the interviewee chooses among his/her contacts c people (who have not yet participated to the study) to whom the coupons are distributed. When the latter come with the coupons, they are in turn interviewed. Each person returning a coupon receives some money, as well as the person who distributed the coupons and depending on how many of the coupons he or she distributed were returned.
To the RDS we can associate a random graph where we attach to each vertex the

contacts to whom he/she has distributed coupons. This tree is embedded into the graph that we would like to explore and which is unknown. Additionally, we have some edges obtained from the direct exploration of the interviewees' neighborhood. This enrich the tree defined by the coupon into a subgraph (not necessarily a tree any more) of the graph of interest. Here we do not consider the information obtained from an interviewee between their alters.

RDS exploration process

We would like first to investigate the proportion of the whole graph discovered by the RDS process. Thus, let us first define the RDS process describing the exploration of the graph. We sum up the exploration process by considering only sizes of –partially– explored components. We thus introduce the process:

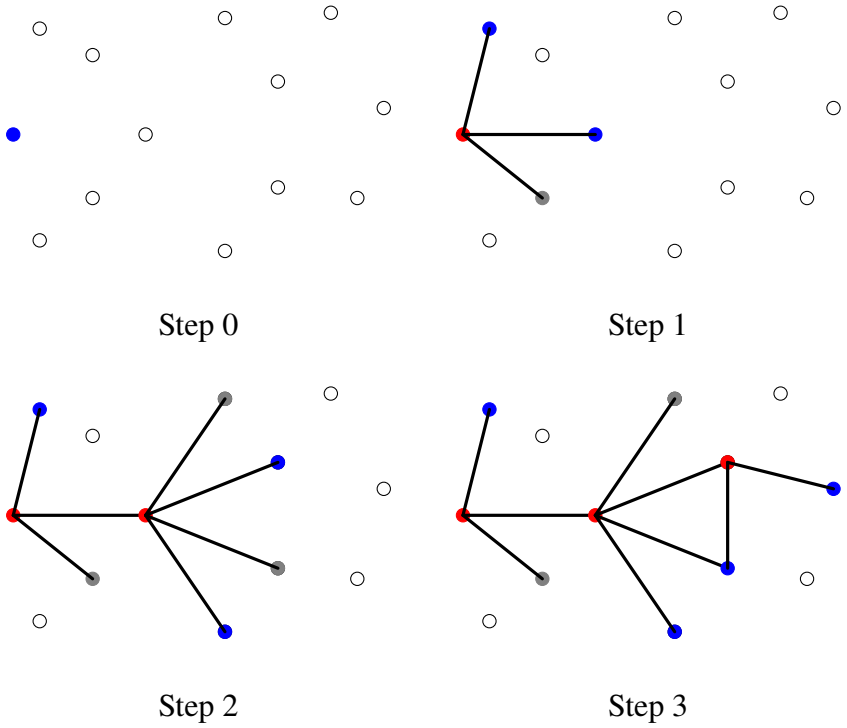
$$X_n = (A_n, B_n) \in \{0, \dots N\}^2, \quad n \in \mathbb{N}. \tag{1.1}$$

The discrete time n is the number of interviews completed, A_n corresponds to the number of individuals that have received coupons but that have not been interviewed yet, B_n to the number of individuals cited in interviews but who have not been given any coupon. We set $X_0 = (A_0, B_0)$: $A_0 > 1$ individual is recruited randomly in the population and we assume that the random graph is unknown at the beginning of the study. The random network is progressively discovered when the RDS process explores it. At time $n \in \mathbb{N}$, the number of unexplored vertices is $N - (n + A_n + B_n)$.

Let us describe the dynamics of $X = (X_n)_{n\in\mathbb{N}}$. At the time $n + 1$, if $A_n > 0$, one individual among these A_n people with coupons is interviewed and is given a maximum of c coupons that he/she would distributed to his/her contacts. If $A_n = 0$, a new individual chosen from the unexplored population (including the individuals mentioned before) is recruited, no coupon is distributed, and we continue the survey. For the sake of simplicity, we assume that the new seeds are chosen uniformly among the unexplored individuals. The process stops at $n = N$, when all vertices in the population have been explored. Thus,

$$\begin{aligned} A_{n+1} =& A_n - \mathbf{1}_{\{A_n \geq 1\}} + Y_{n+1} \wedge c, \\ B_{n+1} =& B_n + H_{n+1} - (H_{n+1} + K_{n+1}) \wedge c \end{aligned} \tag{1.2}$$

where Y_{n+1} is the number of new neighbors, who have not received any coupon before, of the $(n+1)^{\text{th}}$-individual interviewed; H_{n+1} is the number of the $(n+1)^{\text{th}}$-interviewee's new neighbors, who were not mentioned before, and K_{n+1} is the number of the $(n+1)^{\text{th}}$-interviewee's new neighbors, who are chosen amongst the individuals that we knew but do not have any coupon. Of course, $Y_{n+1} = H_{n+1} + K_{n+1}$. At this point, we can see that the transitions of the process $(X_n)_{n\in\mathbb{N}}$ depend heavily on the graph structure: this will determine the distributions of the random variables Y_{n+1}, H_{n+1} and K_{n+1} and their dependencies with the variables corresponding to past interviews (indices n, $n-1$..., 0).



- off-mode node (who has been interviewed)
- active node (who has coupon but has not been interviewed yet)
- explored but still inactive node (who has been named but did not receive coupons)

Figure 1.1. Description of how the chain-referral sampling works. In our model, the random network and the CRS are constructed simultaneously. For example at step 3, an edge between two vertices who are already known at step 2 is revealed.

Case of Erdös-Rényi graphs

If the graph that we explore is an Erdös-Rényi graph [89, 94], then the process $(X_n)_{n\in\mathbb{N}}$ become a Markov process. In this first chapter, we carefully study this simple case and consider an Erdös-Rényi graph in the supercritical regime, where each pair of vertices is connected independently from the other with a given probability λ/N, with $\lambda > 1$.

In this case, we have, conditionally to A_{n-1} and B_{n-1} at step n, that

$$Y_n \overset{(d)}{=} \mathcal{B}in\big(N - n - A_{n-1}, \frac{\lambda}{N}\big) \tag{1.3}$$

$$H_n \overset{(d)}{=} \mathcal{B}in\big(N - n - A_{n-1} - B_{n-1}, \frac{\lambda}{N}\big) \tag{1.4}$$

$$K_n \overset{(d)}{=} \mathcal{B}in\big(B_{n-1}, \frac{\lambda}{N}\big). \tag{1.5}$$

We recall that $Y_n = H_n + K_n$ and conditionally to A_{n-1} and B_{n-1}, H_n and K_n are independent.

Plan of the chapter

In Section 1.2, we show that the process $(X_n)_{n\in\mathbb{N}}$ is a Markov chain and provide some computation for the time at which the number of coupons distributed touches zero, meaning that the RDS process has stopped and should be restarted with another seed. In Section 1.3, the limit of the process $(X_n)_{n\in\mathbb{N}}$, correctly renormalized, is studied. We show that the rescaled process converges to the unique solution on $[0,1]$ of a system of ordinary differential equations. The fluctuations associated with this convergence are established in Section 1.4

Notation: In all the paper, we consider for the sake of simplicity that the space $\mathbb{R}^d$ is equipped with the norm denoted by $\|.\|$: for all $x = (x_1,\dots,x_d) \in \mathbb{R}^d$, $\|x\| = \sum_{i=1}^d |x_i|$.

1.2 Study of the discrete-time process describing the RDS exploration of the graph

1.2.1 Markov property and state space

When the graph underlying the RDS process is an Erdös-Rényi graph, the RDS process $(X_n)_{n\in\mathbb{N}}$ becomes an inhomogeneous Markov process thanks to the identities (1.3). It is then possible to compute the transitions of this process that depend on the time $n \in \{0,\dots N\}$.

Proposition 1.1 Let us consider the Erdös-Rényi random graph on $\{1,\dots N\}$ with probability of connection λ/N between each pair of distinct vertices. Consider the random process $X = (X_n)_{n\in\{0,\dots N\}}$ defined in (1.1)-(1.3). Let $\mathcal{F}_n := \sigma(\{X_i, i \le n\})$ be the canonical filtration associated with the process $(X_n)_{n\in\{0,\dots N\}}$. The process $(X_n)_{n\in\{0,\dots N\}}$ is an inhomogeneous Markov chain with the following transition probabilities: $\mathbb{P}(X_n = (a',b') \mid X_{n-1} = (a,b)) = P_n((a,b),(a',b'))$.

$$P_n((a,b),(a',b')) = \sum_{(h,k)} \binom{b}{k} \binom{N-n-a-b}{h} p^{h+k}(1-p)^{N-n-a-h-k}, \tag{1.6}$$

where the sum is ranging over (h,k) such that $a' = a - \mathbf{1}_{a\ge 1} + (h+k)\wedge c$ and $b' = b + h - (h+k)\wedge c$.

Proof. For $n < N$, we compute $\mathbb{P}(X_{n+1} = (a',b') \mid \mathcal{F}_n)$ using (1.2) and (1.3). The fact that this probability depends only on X_n shows the Markov property and provides the transition probability (1.6). ■

Of course, $A_n, B_n \in \{0,\dots N\}$ but there are more constraints on the components of the process (X_n). First, the number of coupons in the population plus the number of interviewed individuals cannot be greater than the size of the population N, implying

that:

$$A_n + n \leq N \qquad \Leftrightarrow \qquad A_n \leq N - n. \tag{1.7}$$

Also, assume that at time $m \geq 0$, $X_m = (\ell, k)$. Then, the number of coupons distributed in the population can not increase of more than $c-1$ at each step and can not decrease of more than 1. Thus,

$$\ell - (n-m) \leq A_n \leq \ell + (n-m) \times (c-1). \tag{1.8}$$

Thus, the points (n, A_n), for $n \geq m$, belong to the gray area on Fig. 1.2. Let us denote by S this grey region defined by (1.7) and (1.8):

$$S = \Big\{ (n,a) \in \{m, ..., N\} \times \{0, ..., N-\ell\} \;\Big|\; \max\{\ell - (n-m), 0\} \leq a \leq \min\{\ell + (n-m) \times (c-1), N-n\} \Big\}.$$

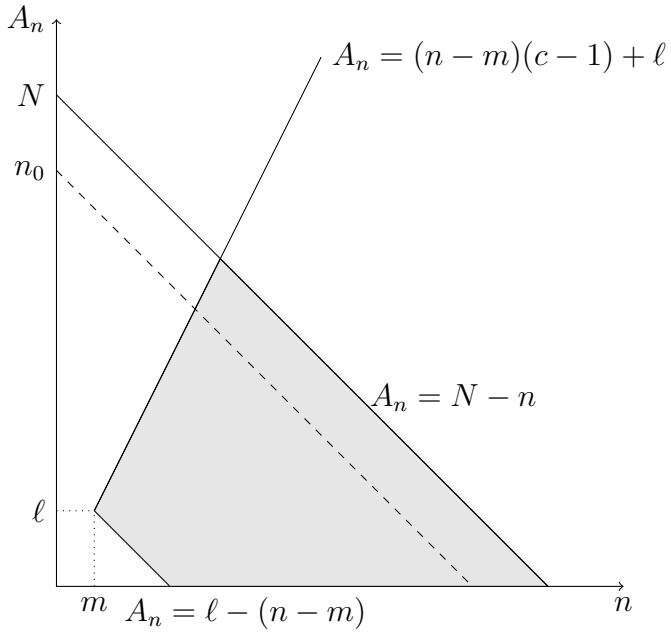


Figure 1.2. *Grey area S: Set of states susceptible to be reach from the process (A_n) started at time m with $A_m = \ell$, as defined by the constraints* (1.7) *and* (1.8)*. The process (A_n) can be stopped upon touching the abscissa axis, which corresponds to the state when the interviews stop because there is no coupons in population any more. The chain conditioned on touching the abscissa axis at $(n_0, 0)$ can not cross the dashed line, which is an additional constraint on the state space.*

1.2.2 Stopping events of the RDS process

We now investigate the first time τ when $A_\tau = 0$, i.e. the time at which the RDS process stops if we do not add another seed because there is no more coupon in the population. Let us define by

$$\tau := \inf\{n \geq 0, A_n = 0\} \tag{1.9}$$

the first time where the RDS process touches the abscissa axis. This stopping time corresponds to the size of the population that we can reach without additional seed other than the initial ones.
Our process evolves in a finite population of size N, and we have seen that the process $A_n \leq N - n$. Thus, $\tau \leq N < +\infty$ almost surely.

For $(n_0, m, \ell) \in \mathbb{N}^3$, let us define the probability that the RDS process without additional seed stops after having seen n vertices and discovered n_0 other existing potential vertices:

$$u_{n_0}(m, \ell) = \mathbb{P}\big(\tau = n_0 \mid A_m = \ell\big). \tag{1.10}$$

By potential theory, $u_{n_0}(.,.) \ : \ S \mapsto [0,1]$ is the smallest solution of the system which, thanks to the previous remarks on the state space of the process, involves only a finite number of equations:

$$u_{n_0}(n_0, 0) = 1, \qquad \forall n \neq n_0,\ u_{n_0}(n, 0) = 0, \tag{1.11}$$

$$u_{n_0}(n, a) = \sum_{a' \ | \ (n+1,a') \in S} P_n\big(a, a'\big) u_{n_0}(n+1, a'), \quad n \leq n_0 - 1,\ a \leq N, \tag{1.12}$$

where $P_n(a, a') = \mathbb{P}\big(A_{n+1} = a' \mid A_n = a\big)$ In fact, the support of u_{n_0} is strictly included in S_{n_0} defined as follows, when $n_0 < N$:

$$S_{n_0} = \Big\{(n, a) \in \{m, ..., N\} \times \{0, ..., N - \ell\} \ \Big| \ \max\{\ell - (n - m), 0\} \leq a \leq \min\{\ell + (n - m) \times (c - 1), n_0 - n\}\Big\} \tag{1.13}$$

since the maximal number of interviewed individuals (and hence of distributed coupons) is n_0 on the event of interest (see dashed line in Fig. 1.2).

For Erdös-Rényi graphs with connection probability λ/N, we have more precisely:

$$P_n(a, a') = \begin{cases} \binom{N - (n+1) - a}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-(n+1)-a-k} & \text{if } (-1 \leq a' - a = k - 1, k < c); \\ 1 - \sum_{k=0}^{c-1} \binom{N - (n+1) - a}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-(n+1)-a-k} & \text{if } a' - a = c - 1; \\ 0 & \text{otherwise .} \end{cases}$$

Let us define for $n \geq 0$:

$$\mathbf{U}_{n_0}^{(n)} := \begin{pmatrix} u_{n_0}(n,1) \\ \vdots \\ u_{n_0}(n,a) \\ \vdots \\ u_{n_0}(n,n_0) \end{pmatrix} \tag{1.14}$$

and $\mathbf{P}_{n_0}^{(n)}$ the $n_0 \times n_0$ matrix with entries $(P_n(a,a'); 1 \leq a, a' \leq n_0)$.

Then, for $n < n_0 - 1$, the system of equations (1.11)-(1.12) becomes

$$\mathbf{U}_{n_0}^{(n)} = \mathbf{P}_{n_0}^{(n)} \mathbf{U}_{n_0}^{(n+1)}.$$

And for $n = n_0 - 1$, the boundary condition gives that

$$\mathbf{U}_{n_0}^{(n_0-1)} := \begin{pmatrix} u_{n_0}(n_0-1,1) \\ \vdots \\ u_{n_0}(n_0-1,a) \\ \vdots \\ u_{n_0}(n_0-1,n_0) \end{pmatrix} = \begin{pmatrix} P_{n_0-1}(1,0) \\ \vdots \\ P_{n_0-1}(a,0) \\ \vdots \\ P_{n_0-1}(n_0,0) \end{pmatrix} = \begin{pmatrix} P_{n_0-1}(1,0) \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}.$$

1.3 Limit of the normalized RDS process

For an integer $N \geq 1$, let us consider the following renormalization $X^N = (A^N, B^N)$ of the process X:

$$X_t^N := \frac{1}{N} X_{\lfloor Nt \rfloor} = \left(\frac{A_{\lfloor Nt \rfloor}}{N}, \frac{B_{\lfloor Nt \rfloor}}{N} \right) \in [0,1]^2, \quad t \in [0,1]. \tag{1.15}$$

Notice that X^N is constant by part and jumps at the times $t_n = n/N$ for $n \in \{1, \ldots, N+1\}$. Thus the process X^N belongs to the space $\mathcal{D}([0,1],[0,1]^2)$ of càdlàg processes from $[0,1]$ to $[0,1]^2$ embedded with the Skorokhod topology [86, 55]. Define the filtration associated to X^N as $(\mathcal{F}_t^N)_{t\in[0,1]} = (\mathcal{F}_{\lfloor Nt \rfloor})_{t\in[0,1]}$. We aim to study the limit of the normalized process $X^N = (A^N, B^N)$ when N tends to infinity.

Assumption 1.1 Let $a_0, b_0 \in [0,1]$ with $a_0 > 0$ and $b_0 = 0$. We assume that the sequence $X_0^N = \frac{1}{N} X_0$ converges in probability to the vector $x_0 = (a_0, b_0)$ as N tends to infinity.

Theorem 1.1 Under the assumption 1.1, when N tends to infinity, the sequence of processes $X^N = (A^N, B^N)$ converges in distribution in $\mathcal{D}([0,1],[0,1]^2)$ to a deterministic path $x = (a,b) \in \mathcal{C}([0,1],[0,1]^2)$, which is the unique solution of the following system of ordinary differential equations

$$x_t = x_0 + \int_0^t f(s, x_s)ds, \tag{1.16}$$

where $f(t, x_t) = (f_1(t, x_t), f_2(t, x_t))$ has the explicit formula:

$$f_1(t, x_t) = c - \sum_{k=0}^{c-1}(c-k)p_k(t + a_t) - \mathbf{1}_{a_t>0} \tag{1.17}$$

$$f_2(t, x_t) = (1 - t - a_t - b_t)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t + a_t) - c, \tag{1.18}$$

with

$$p_k(z) := \frac{\lambda^k(1-z)^k}{k!}e^{-\lambda(1-z)}, \quad k \in \{0, ..., c\}, \tag{1.19}$$

and c is the maximum value of coupons distributed at each time step.

Remark 1.1 Since the limiting process $x \in \mathcal{C}([0, 1], [0, 1]^2)$ is deterministic, the convergence in distribution of Theorem 1.1 is in fact a convergence in probability.

The proof of Theorem 1.1 follows the steps below. First, we enounce a semi-martingale decomposition for $(X^N)_{N\geq 1}$ that allows us to prove the tightness of the sequence $(X^N)_{N\geq 1}$ by using Aldous-Rebolledo criteria. Then, we identify the equation satisfied by the limiting values of $(X^N)_{N\geq 1}$, and show that the latter has a unique solution.

Let us first have some comments on the solution of (1.16).

Proposition 1.2 Let us denote

$$t_0 := \inf\{t \in [0, 1] : |a_t| = 0\}. \tag{1.20}$$

Then $a_t = 0, \forall t \in [t_0, 1]$.

Proof. For $c = 1$, (1.17)-(1.18) gives that

$$\frac{da}{dt} = 1 - p_0(t + a) - \mathbf{1}_{a>0} = \begin{cases} -e^{-\lambda(1-t-a)} < 0 & \text{if } a > 0 \\ 1 - e^{-\lambda(1-t)} > 0 & \text{if } a = 0. \end{cases}$$

Recall also that for all $t \in [0, 1]$, $a_t + t \in [0, 1]$ since it corresponds to the proportion of individuals who have received a coupon (already interviewed or not). The right hand side of (1.17)-(1.18) has a discontinuity on the abscissa axis that implies that the solution stays at 0 after t_0. Notice that this was expected since when $c = 1$, $\{0, 1\}$ is an absorbing state for the Markov process $(A^N)_{N\geq 1}$.

Let us now consider the case $c > 1$. We have then that

$$\frac{da}{dt} = \phi(a+t) - \mathbf{1}_{a>0},$$

where

$$\phi(z) := c - \sum_{k=0}^{c-1}(c-k)p_k(z) = c - \sum_{k=0}^{c-1}(c-k)\frac{\lambda^k(1-z)^k}{k!}e^{-\lambda(1-z)}. \tag{1.21}$$

By Lemma 1.4, the function ϕ is strictly decreasing with $\phi(1) = 0$ and $\phi(1-1/\lambda) > 1$. From this we deduce that ϕ is a positive function on $(0,1)$ and that there exists a unique $z_c \in (1-1/\lambda, 1)$ such that $\phi(z_c) = 1$. For all t such that $0 < t < t_0$, we have

$$\frac{d(a_t+t)}{dt} = \phi(a+t) - 1 + 1 = \phi(a_t+t) > 0.$$

It implies that $t \mapsto t + a_t$ is a strictly increasing function on $[0, t_0]$ and thus

$$a_0 < t + a_t < t_0, \quad \forall t \in (0, t_0).$$

If $z_c > t_0$, then $1 = \phi(z_c) < \phi(t_0) < \phi(t + a_t)$ for all $t \in (0, t_0)$. It follows that $\frac{da_t}{dt} > 0$. Hence, a_t is strictly increasing in the interval $(0, t_0)$. Notice that $t + a_t$ is continuous function on $[0, 1]$, and since $t + a_t$ is strictly increasing, we deduce that $0 < a_0 < a_{t_0} = 0$, which is impossible.
If $z_c < a_0 < t_0$, then $1 = \phi(z_c) > \phi(t + a_t)$ for all t such that $t + a_t > z_c$. And thus $\frac{da_t}{dt} = \phi(t + a_t) - 1 < 0$ whenever $t + a_t > z_c$ and $a_t > 0$.
If $z_c \in [a_0, t_0]$, then there exists a unique $t_c \in [0, t_0]$ such that $t_c + a_{t_c} = z_c$. It follows that there is a value t_c in the interval $[0, t_0]$ such that $\phi(t_c + a_{t_c}) = 1$. Then $\phi(t + a_t) > 1$ for all $t \in (0, t_c)$ and $\phi(t + a_t) > 1$ for $t \in (t_c, 1)$. Thus,

$$\frac{da_t}{dt} > 0 \text{ when } t \in (0, t_c) \quad \text{and} \quad \frac{da_t}{dt} < 0 \text{ when } t \in \{t > t_c : a_t > 0\}.$$

After the time t_0, there is again a discontinuity in the vector field $(t, a) \mapsto \phi(t + a_t) - \mathbf{1}_{a>0}$ which is directed toward negative ordinates when $a > 0$ and positive ordinate when $a < 0$. This implies that the solution of the dynamical system stays at 0 after time t_0.

■

Now, for the first step of the proof of Theorem 1.1, we write the Doob's decomposition of $(X^N)_{N\geq 1}$ as follows.

Lemma 1.1 The process X^N, for $N \in \mathbb{N}^*$, admits the following Doob decom-

position: $X_t^N = X_0^N + \Delta_t^N + M_t^N$, or in the vectorial form

$$\begin{pmatrix} X^{N,1} \\ X^{N,2} \end{pmatrix} = \begin{pmatrix} A_0^N \\ B_0^N \end{pmatrix} + \begin{pmatrix} \Delta^{N,1} \\ \Delta^{N,2} \end{pmatrix} + \begin{pmatrix} M^{N,1} \\ M^{N,2} \end{pmatrix}. \tag{1.22}$$

The predictable process with finite variations Δ^N is:

$$\begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \end{pmatrix} = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \begin{pmatrix} \mathbb{E}[Y_n \wedge c \mid \mathcal{F}_{n-1}] - \mathbf{1}_{A_{n-1} \geq 1} \\ \mathbb{E}[H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}] \end{pmatrix} \tag{1.23}$$

The square integrable centered martingale M^N has quadratic variation process $\langle M^N \rangle$ given as follows:

$$\langle M^N \rangle_t = \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \begin{pmatrix} \mathrm{Var}\big(Y_n \wedge c \mid \mathcal{F}_{n-1}\big) & \mathrm{Cov}\big(Y_n \wedge c, H_n - Y_n \wedge c\big) \\ \mathrm{Cov}\big(Y_n \wedge c, H_n - Y_n \wedge c\big) & \mathrm{Var}\big(H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}\big) \end{pmatrix}. \tag{1.24}$$

Notice that the quantities in (1.23) and (1.24) can be computed as functions of $A^N_{t_{n-1}}$ and $B^N_{t_{n-1}}$ for $n \in \{1, ..., N\}$:

$$\mathbb{E}[Y_n \wedge c \mid \mathcal{F}_{n-1}] = c - \sum_{k=0}^{c-1} (c-k) \mathbb{P}(Y_n = k | \mathcal{F}_{n-1}) \tag{1.25}$$

where

$$\begin{aligned} \mathbb{P}(Y_n = k | \mathcal{F}_{n-1}) =& \frac{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1)!}{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1 - k)! N^k} \frac{\lambda^k}{k!} \\ & \times \left(1 - \frac{\lambda}{N}\right)^{N(1 - t_{n-1} - A^N_{t_{n-1}})} \left(1 - \frac{\lambda}{N}\right)^{-k-1}, \end{aligned} \tag{1.26}$$

and

$$\mathbb{E}[H_n | \mathcal{F}_{n-1}] = \lambda \left(1 - \frac{n}{N} - A^N_{t_{n-1}} - B^N_{t_{n-1}}\right). \tag{1.27}$$

For the bracket in (1.24), the terms can be computed from:

$$\mathbb{E}\left[(Y_n \wedge c)^2 \,\middle|\, \mathcal{F}_{n-1}\right] = c^2 + \sum_{k=0}^{c} (k^2 - c^2) \mathbb{P}(Y_n = k | \mathcal{F}_{n-1}); \tag{1.28}$$

$$\mathbb{E}\left[Y_n \wedge c \middle| \mathcal{F}_{n-1}\right]^2 = \left(c + \sum_{k=0}^{c} (k-c) \mathbb{P}(Y_n = k | \mathcal{F}_{n-1})\right)^2; \tag{1.29}$$

$$\mathrm{Var}(H_n | \mathcal{F}_{n-1}) = \lambda \left(1 - \frac{n}{N} - \frac{A_{n-1}}{N} - \frac{B_{n-1}}{N}\right)\left(1 - \frac{\lambda}{N}\right); \tag{1.30}$$

and

$$\mathbb{E}[H_n (Y_n \wedge c) | \mathcal{F}_{n-1}] = \sum_{k=0}^{N-n-A_{n-1}} (k \wedge c) \mathbb{E}(H_n | Y_n = k) \mathbb{P}(Y_n = k | \mathcal{F}_{n-1})$$

$$= \frac{N-n-A_{n-1}-B_{n-1}}{N-n-A_{n-1}} \left[\sum_{k=0}^{c} k^2 \mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + \sum_{k=c+1}^{N-n-A_{n-1}} ck\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) \right]$$
$$= \left(1 - \frac{B_{n-1}}{N-n-A_{n-1}}\right) \left[\sum_{k=0}^{c} (k^2 - ck)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\mathbb{E}[Y_n|\mathcal{F}_{n-1}] \right]$$
$$= \left(1 - \frac{B_{n-1}}{N-n-A_{n-1}}\right) \left[\sum_{k=0}^{c} (k^2 - ck)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\lambda \left(1 - \frac{n}{N} - \frac{A_{n-1}}{N}\right) \right]. \tag{1.31}$$

Proof. Since the components of X^N take their values in $[0,1]$, the process X^N is clearly square integrable. It is classical to write X_t^N as

$$\begin{aligned} X_t^N &= X_0^N + \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} (X_n - X_{n-1}) \\ &= X_0^N + \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] \\ &\qquad + \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} (X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]). \end{aligned}$$

Let us call Δ_t^N the second term in the right hand side, and M_t^N the third term. We will prove that Δ^N is an $\mathcal{F}_t^N$-predictable finite variation process and that M^N is a square integrable martingale.

Let us first consider $(\Delta_t^N)_{0 \le t \le 1}$. From (1.2), we have that for the first component:

$$A_n - A_{n-1} = Y_n \wedge c - \mathbf{1}_{\{A_{n-1} \ge 1\}}, \quad B_n - B_{n-1} = H_n - Y_n \wedge c.$$

Moreover, for each $n \in \{1, ..., N\}$, $\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]$ is $\mathcal{F}_{n-1}$-measurable. Hence, Δ_t^N is $\mathcal{F}_{\lfloor Nt \rfloor - 1}$-measurable. The total variation of Δ^N is:

$$\begin{aligned} V(\Delta_t^N) &= \sum_{n=1}^{\lfloor Nt \rfloor} \|\Delta_{t_n}^N - \Delta_{t_{n-1}}^N\| \\ &= \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} |\mathbb{E}[A_n - A_{n-1} \mid \mathcal{F}_{n-1}]| + |\mathbb{E}[B_n - B_{n-1} \mid \mathcal{F}_{n-1}]| \\ &\le (2c + \lambda)t < +\infty, \end{aligned}$$

by using (1.2), as $Y_n \wedge c \le c$ and $\mathbb{E}[H_n \mid \mathcal{F}_{n-1}] \le \lambda$.

Furthermore, using (1.2), we can recover the expression (1.23) of Δ^N announced in the lemma as:

$$\begin{aligned}\mathbb{E}[Y_n \wedge c|\mathcal{F}_{n-1}] &= \sum_{k=0}^{c} k\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\mathbb{P}(Y_n > c|\mathcal{F}_{n-1}) \\ &= \sum_{k=0}^{c} k\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\big(1 - \mathbb{P}(Y_n \leq c|\mathcal{F}_{n-1})\big) \\ &= c - \sum_{k=0}^{c}(c-k)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}),\end{aligned}$$

where

$$\begin{aligned}\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) &= \binom{N - Nt_n - NA^N_{t_n} - 1}{k}\Big(\frac{\lambda}{N}\Big)^k\Big(1 - \frac{\lambda}{N}\Big)^{N-Nt_n-NA^N_{t_n}-1-k} \\ &= \frac{(N - Nt_n - NA^N_{t_n} - 1)!}{(N - Nt_n - NA^N_{t_n} - 1 - k)!k!}\Big(\frac{\lambda}{N}\Big)^k\Big(1 - \frac{\lambda}{N}\Big)^{N-Nt_n-NA^N_{t_n}-1-k} \\ &= \frac{(N - Nt_n - NA^N_{t_n} - 1)!}{(N - Nt_n - NA^N_{t_n} - 1 - k)!N^k}\left(1 - \frac{\lambda}{N}\right)^{-k-1} \\ &\qquad \times \left(1 - \frac{\lambda}{N}\right)^{N(1-t_n-A^N_{t_n})}\frac{\lambda^k}{k!}.\end{aligned}$$

Let us now show that $(M^N_t)_{0\leq t\leq 1}$ is a bounded $\mathcal{F}^N_t$-martingale and let us compute its quadratic integration process. For every $t \in [0,1]$, M^N_t is $\mathcal{F}^N_t$-measurable and bounded and hence square integrable:

$$|M^N_t| = \left|X^N_t - X^N_0 - \Delta^N_t\right| \leq 2 + (2c+\lambda)t \leq 2 + 2c + \lambda < +\infty.$$

For all $s < t$,

$$\begin{aligned}\mathbb{E}[M^N_t|\mathcal{F}^N_s] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}(X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}])\Big|\mathcal{F}_{\lfloor Ns\rfloor}\right] \\ &\quad + \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{\lfloor Ns\rfloor}(X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}])\Big|\mathcal{F}_{\lfloor Ns\rfloor}\right] \\ &= \frac{1}{N}\sum_{n=1}^{\lfloor Ns\rfloor}(X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]) = M^N_s.\end{aligned}$$

Then M^N_t is an $(\mathcal{F}^N_t)$-martingale.

Let us denote $X^1_n = A_n$ and $X^2_n = B_n$. The quadratic variation process is defined as:

$$\langle M^N\rangle_t = \begin{bmatrix}\langle M^{N,1}, M^{N,1}\rangle_t & \langle M^{N,1}, M^{N,2}\rangle_t \\ \langle M^{N,2}, M^{N,1}\rangle_t & \langle M^{N,2}, M^{N,2}\rangle_t\end{bmatrix}, \tag{1.32}$$

where for $k, \ell \in \{1,2\}$,

$$\langle M^{N,k}, M^{N,\ell}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\Big\{\mathbb{E}\left[(X^k_n - X^k_{n-1})(X^\ell_n - X^\ell_{n-1})\big|\mathcal{F}_{n-1}\right]$$

$$- \mathbb{E}\left[(X_n^k - X_{n-1}^k)|\mathcal{F}_{n-1}\right]\mathbb{E}\left[(X_n^\ell - X_{n-1}^\ell)|\mathcal{F}_{n-1}\right]\Big\}. \tag{1.33}$$

Using (1.2), we have:

$$\begin{aligned}\langle M^{N,1}\rangle_t &= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\left(A_n - A_{n-1} - \mathbb{E}[A_n - A_{n-1}|\mathcal{F}_{n-1}]\right)^2\Big|\mathcal{F}_{n-1}\right]\\ &= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathrm{Var}(Y_n \wedge c|\mathcal{F}_{n-1}) \le \frac{c^2}{N}.\end{aligned} \tag{1.34}$$

Proceeding similarly for the other terms, we obtain

$$\begin{aligned}\langle M^{N,2}\rangle_t &= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathrm{Var}(H_n - Y_n \wedge c|\mathcal{F}_{n-1}) \le \frac{\lambda}{N},\\ \langle M^{N,1}, M^{N,2}\rangle_t &= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathrm{Cov}(Y_n \wedge c, H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}) \le \frac{c\sqrt{\lambda}}{N}.\end{aligned} \tag{1.35}$$

This finishes the proof of the Lemma. ■

1.3.1 Tightness of the renormalized process

Lemma 1.2 The sequence $(X^N)_{N\ge1}$ is tight in $\mathcal{D}([0,1],[0,1]^2)$.

Proof. The proof of tightness is based on the classical criterion of Aldous-Rebolledo ([98, Theorem 2.3.2] and its Corollary 2.3.3). For this we have to check that finite distributions are tight, and control the modulus of continuity of the sequence of finite variation parts and of quadratic variation of the martingale parts.

For each $t \in [0,1]$, $|A_t^N| + |B_t^N| \le 2$, implying that (A_t^N, B_t^N) is tight for every $t \in [0,1]$.

Let $0 \le s,t \le 1$,

$$\begin{aligned}\|\Delta_t^N - \Delta_s^N\| &= |\Delta_t^{N,1} - \Delta_s^{N,1}| + |\Delta_t^{N,2} - \Delta_s^{N,2}|\\ &\le \frac{1}{N}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\left(\left|\mathbb{E}\left[A_n - A_{n-1}|\mathcal{F}_{n-1}\right]\right| + \left|\mathbb{E}\left[B_n - B_{n-1}|\mathcal{F}_{n-1}\right]\right|\right)\\ &\le (2c+\lambda)|t-s|.\end{aligned}$$

Thus, for each positive ε and η, there exists $\delta_0 = \dfrac{\varepsilon\eta}{2c+\lambda}$ such that for all $0 < \delta < \delta_0$,

$$\mathbb{P}\left(\sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}}\left|\Delta_t^N-\Delta_s^N\right|>\eta\right)\le\frac{1}{\eta}\mathbb{E}\left[\sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}}\left|\Delta_t^N-\Delta_s^N\right|\right]\le\frac{(2c+\lambda)\delta}{\eta}\le\varepsilon,\quad\forall N\ge 1. \tag{1.36}$$

By Aldous criterion, this provides the tightness of $(\Delta^N)_{N\in\mathbb{N}}$.

Similarly, for the quadratic variations of the martingale parts, using (1.34) and (1.35), we have for all $0\le s<t\le 1$,

$$\begin{aligned}
\left|\langle M^{N,1}\rangle_t-\langle M^{N,1}\rangle_s\right| &= \frac{1}{N^2}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\mathrm{Var}\left(Y_n\wedge c\middle|\mathcal{F}_{n-1}\right)\le\frac{c^2}{N}|t-s|;\\
\left|\langle M^{N,2}\rangle_t-\langle M^{N,2}\rangle_s\right| &= \frac{1}{N^2}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\mathrm{Var}\left(H_n-Y_n\wedge c\middle|\mathcal{F}_{n-1}\right)\\
&\le\frac{2(\lambda+c^2)}{N}|t-s|;\\
\left|\langle M^{N,1},M^{N,2}\rangle_t-\langle M^{N,1},M^{N,2}\rangle_s\right| &\le \frac{1}{N^2}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\left(\mathrm{Var}(Y_n\wedge c|\mathcal{F}_{n-1})\right)^{1/2}\\
&\qquad\times\left(\mathrm{Var}(H_n-Y_n\wedge c|\mathcal{F}_{n-1})\right)^{1/2}\\
&\le\frac{c(\sqrt{\lambda}+c)}{N}|t-s|.
\end{aligned}$$

Thus, using the matrix norm on $\mathcal{M}_{2\times 2}(\mathbb{R})$ associated with $\|.\|_1$ on $\mathbb{R}^2$,

$$\begin{aligned}
\sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}}\|\langle M^N\rangle_t-\langle M^N\rangle_s\| &\le \sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}}\Big(\left|\langle M^{N,1}\rangle_t-\langle M^{N,1}\rangle_s\right|+\left|\langle M^{N,2}\rangle_t-\langle M^{N,2}\rangle_s\right|\\
&\qquad+2\left|\langle M^{N,1},M^{N,2}\rangle_t-\langle M^{N,1},M^{N,2}\rangle_s\right|\Big)\\
&\le\frac{c^2+4(\lambda+c^2)+c(\sqrt{\lambda}+c)}{N}\delta.
\end{aligned} \tag{1.37}$$

Consequently, for any $\varepsilon>0,\eta>0$, choose δ such that $\dfrac{c^2+4(\lambda+c^2)+c(\sqrt{\lambda}+c)}{\eta N}\delta<\varepsilon$, we have

$$\mathbb{P}\left(\sup_{\substack{|t-s|<\delta\\0\le s,t\le 1}}|\langle M^N\rangle_t-\langle M^N\rangle_s|>\eta\right)<\varepsilon,\quad\forall N\ge 1,$$

which implies that $\langle M^N\rangle$ is also tight. This achieves the proof of the Lemma. ■

1.3.2 Identification of the limiting values

Since $(X^N)_{N\ge1}$ is tight, there exists a subsequence $(\ell_N)_{N\ge1}$ in $\mathbb{N}$ such that $(X^{\ell_N})_{N\ge1}=(A^{\ell_N},B^{\ell_N})_{N\ge1}$ converges in distribution in $\mathcal{D}([0,1],[0,1]^2)$ to a limiting value $(\bar{a},\bar{b})\in\mathcal{D}([0,1],[0,1]^2)$ (*e.g.* [85]). We now want to identify that limiting value.

Proposition 1.3 The sequence of martingales $(M^N)_{N\geq 1}$ converges uniformly to 0 in probability when $N \to \infty$.

Proof. With a computation similar the one leading to (1.37), we get

$$\|\langle M\rangle_t\| \leq |\langle M^{N,1}\rangle_t| + |\langle M^{N,2}\rangle_t| + 2|\langle M^{N,1}\rangle_t|^{1/2}|\langle M^{N,2}\rangle_t|^{1/2} \leq \frac{(6c^2+4\lambda)t}{N} \tag{1.38}$$

By Doob's inequality,

$$\mathbb{E}[\sup_{t\in[0,1]} \|M_t^N\|^2] \leq 4\mathbb{E}[\|\langle M\rangle_1\|] \leq 4\frac{6c^2+4\lambda}{N}.$$

For every $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,1]} \|M_t^N\|^2 > \varepsilon\right) \leq \lim_{N\to\infty} \frac{1}{\varepsilon}\mathbb{E}[\sup_{t\in[0,1]} \|M_t^N\|^2] \leq \lim_{N\to\infty} \frac{4(6c^2+4\lambda)}{\varepsilon N} = 0.$$

■

The remaining work is figuring out the limit of finite variation part Δ^N. Let us recall that

$$f_1(t,a) := c - \sum_{k=0}^{c-1}(c-k)p_k(t+a) - \mathbf{1}_{a_t>0}$$

$$f_2(t,a,b) := (1-t-a-b)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t+a) - c.$$

and

$$f(t,a,b) := \begin{pmatrix} f_1(t,a) \\ f_2(t,a,b) \end{pmatrix} \tag{1.39}$$

the r.h.s. of (1.17)-(1.18), where $p_k(x)$ is the function defined in (1.19).

Proposition 1.4 There exists a constant $C = C(\lambda, c) > 0$ such that for all $N \geq 1$,

$$\sup_{t\in[0,1]} \left\|\Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\left(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N}\right)\right\| \leq \frac{C}{N} \tag{1.40}$$

Proof. Recall the equations for Δ^N in (1.23) and (1.26). Using (1.27), we have that:

$$\left\|\Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\left(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N}\right)\right\|$$

$$
\begin{aligned}
\leq & \Big| \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \Big(c - \sum_{k=0}^{c} (c-k) \mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - \mathbf{1}_{A_{n-1} \geq 1} \Big) \\
& \qquad - \Big(c - \sum_{k=0}^{c} (c-k) p_k\big(\frac{n-1}{N} + \frac{A_{n-1}}{N}\big) - \mathbf{1}_{\frac{A_{n-1}}{N} > 0} \Big) \Big| \\
& + \Big| \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \Big(\mathbb{E}[H_n \mid \mathcal{F}_{n-1}] + \sum_{k=0}^{c} (c-k) \mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - c \Big) \\
& \qquad - \Big(\lambda\big(1 - \frac{n-1}{N} - \frac{A_{n-1}}{N} - \frac{B_{n-1}}{N}\big) - \sum_{k=0}^{c} (c-k) p_k\big(\frac{n-1}{N} + \frac{A_{n-1}}{N}\big) \Big) \Big| \\
\leq & \frac{2}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \sum_{k=0}^{c} (c-k) \left| \mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - p_k \left(\frac{n-1}{N} + \frac{A_{n-1}}{N} \right) \right|. \qquad (1.41)
\end{aligned}
$$

We are thus led to consider more carefully the difference between $\mathbb{P}(Y_n = k \mid \mathcal{F}_{n-1})$ and $p_k(t_{n-1} + A^N_{t_{n-1}})$. We have

$$
\begin{aligned}
& \frac{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1)!}{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1 - k)! N^k} \\
& = (1 - t_{n-1} - A^N_{t_{n-1}} - \frac{1}{N})(1 - t_{n-1} - A^N_{t_{n-1}} - \frac{2}{N}) \cdots (1 - t_{n-1} - A^N_{t_{n-1}} - \frac{k}{N}) \\
& = Q_k(1 - t_{n-1} - A^N_{t_{n-1}}),
\end{aligned}
$$

where for $k \leq c$,

$$
Q_k(x) = \prod_{n=1}^{k} (x - x_n) = \sum_{j=0}^{k} (-1)^{k-j} e_{k-j} x^j
$$

is a polynomial of degree k, with the notation $x_n = n/N$, $e_0 = 1, e_j = \sum_{1 \leq i_1 < \ldots < i_j \leq k} x_{i_1} \ldots x_{i_j}$,$1 \leq j \leq k$. Since

$$
|Q_k(x) - x^k| = \left| \sum_{j=0}^{k-1} (-1)^{k-j} e_{k-j} x^j \right| \leq \sum_{j=0}^{k-1} |e_{k-j}||x^j| \leq \sum_{j=0}^{k-1} \Big(\frac{(k-1)}{N} \Big)^{k-j} |x^j|,
$$

this yields:

$$
\begin{aligned}
& \left| \frac{(N - Nt_i - NA^N_{t_i} - 1)!}{(N - Nt_i - NA^N_{t_i} - k - 1)! N^k} - (1 - t_i - A^N_{t_i})^k \right| \\
& \leq \sum_{j=0}^{k-1} \left(\frac{k-1}{N} \right)^{k-j} \leq \frac{\sum_{\ell=1}^{k} (k-1)^\ell}{N}. \qquad (1.42)
\end{aligned}
$$

Secondly, we upper bound the difference between $(1 - \lambda/N)^{N(1 - t_{n-1} - A^N_{t_{n-1}})}$ and $\exp(-\lambda(1 - t_{n-1} - A^N_{t_{n-1}}))$. Using a Taylor expansion, we obtain that:

$$
\begin{aligned}
\Big(1 - \frac{\lambda}{N}\Big)^{N(1 - t_{n-1} - A^N_{t_{n-1}})} & = \exp\Big(N(1 - t_{n-1} - A^N_{t_{n-1}}) \log\Big(1 - \frac{\lambda}{N}\Big) \Big) \\
& = \exp\Big(N(1 - t_{n-1} - A^N_{t_{n-1}}) \log\Big(1 - \frac{\lambda}{N}\Big) \Big)
\end{aligned}
$$

$$= e^{-\lambda(1-t_{n-1}-A^N_{t_{n-1}})} \exp\Big(-(\frac{\lambda^2}{2N} + r_N)(1 - t_{n-1} - A^N_{t_{n-1}})\Big)$$

where there exists some constant $C = C(\lambda) > 0$ such that $0 \leq r_N < C/N^3$. Using that for $x > 0$, $1-x < e^{-x} < 1$, we obtain that for some constant $C_0 = C_0(\lambda)$,

$$0 \leq e^{-\lambda(1-t_n-A^N_{t_n})} - \Big(1 - \frac{\lambda}{N}\Big)^{N(1-t_n-A^N_{t_n})} \leq \frac{C_0}{N}. \tag{1.43}$$

Lastly, there exists a constant $C_1 = C_1(c, \lambda) \geq 0$ such that

$$1 \leq \Big(1 - \frac{\lambda}{N}\Big)^{-(k+1)} \leq 1 + \frac{C_1}{N}. \tag{1.44}$$

Gathering (1.26), (1.42), (1.43) and (1.44), there thus exists a constant $C_2 = C_2(c, \lambda)$ such that

$$\Big|\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) - p_k(t_{n-1} + A^N_{t_{n-1}})\Big| \leq \frac{C_2(\lambda, c)}{N}. \tag{1.45}$$

As a result, from (1.41) and (1.45) we have for some constant $C = C(\lambda, c) \geq 0$

$$\Big\|\Delta^N_t - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N})\Big\| \leq \frac{C(\lambda, c)}{N}.$$

This proves the proposition. ■

Corollary 1.1 The limiting values of $(X^N)_{N\geq 1}$ are solutions of (1.17)-(1.18).

Proof. Let us consider a limiting value $(\bar{a}, \bar{b}) \in \mathcal{D}([0,1], [0,1]^2)$ of $(X^N)_{N\geq 1}$. With an abuse of notation, we denote by $(X^N)_{N\geq 1}$ the subsequence converging to $(\bar{a}, \bar{b})$. From (1.22), Propositions 1.3 and 1.4, we obtain that the process

$$\Big(X_t - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\left(\frac{n-1}{N}, A^N_{\frac{n-1}{N}}, B^N_{\frac{n-1}{N}}\right),\ t \in [0,1]\Big)$$

converges uniformly to zero when $N \to +\infty$. Using Lemma 1.3, the process

$$\Big(\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f(\frac{n-1}{N}, A^N_{\frac{n-1}{N}}, B^N_{\frac{n-1}{N}}),\ t \in [0,1]\Big)$$

converges uniformly to the process

$$\Big(\int_0^t f(s, \bar{a}_s, \bar{b}_s)ds,\ t \in [0,1]\Big).$$

We deduce from this that the limiting value of $(X^N)_{N\geq 1}$ is necessarily solution of (1.17)-(1.18). ■

1.3.3 Uniqueness of the ODE solutions

To prove Theorem 1.1, it remains to prove the uniqueness of the limiting value, *i.e.* that:

Proposition 1.5 The system of differential equations (1.17)-(1.18) admits a unique solution.

Proof. Suppose that (1.17)-(1.18) have two solutions (a^1, b^1) and (a^2, b^2), then for all $t \in [0,1]$,

$$|a_t^1 - a_t^2| \leq \int_0^t |g(s, a_s^1) - g(s, a_s^2)| ds + \int_0^t \left| \mathbf{1}_{\{a_s^1 > 0\}} - \mathbf{1}_{\{b_s^2 > 0\}} \right| ds, \tag{1.46}$$

where

$$g(t, a_t, b_t) := c - \sum_{k=0}^{c-1} (c-k) p_k(t + a_t). \tag{1.47}$$

In the first term of the right hand side of (1.46), we have

$$|g(s, a_s^1) - g(s, a_s^2)| \leq |\partial_a g(s, \xi_s)| |a_s^1 - a_s^2|, \tag{1.48}$$

for some real value ξ_s between a_s^1 and a_s^2, *i.e.* $\min\{a_s^1, a_s^2\} \leq \xi_s \leq \max\{a_s^1, a_s^2\}$. For the second term, we want to prove that for all $t \in [0,1]$,

$$\int_0^t \left| \mathbf{1}_{a_s^1 > 0} - \mathbf{1}_{a_s^2 > 0} \right| ds = 0. \tag{1.49}$$

In order to do so, we first prove that all the solutions of (1.17) touch zero at the same point and that after touching zero, they stay at zero. Consider the equation:

$$\frac{d\bar{a}_t}{dt} = g(t, \bar{a}_t) - 1. \tag{1.17'}$$

Because the function $(t, a) \mapsto f_1(t, a) - 1$ is continuous with respect to t and Lipschitz with respect to a on $[0,1]$, Equation (1.17)' has unique solution $\bar{a}_t$ for t in $[0,1]$. Let us define

$$\bar{t}_0 := \inf\{t > 0 : \bar{a}_t = 0\}$$

and

$$t_0 := \inf\{t > 0 : a_t = 0\}$$

where a_t is a solution of (1.17). Since the two equations (1.17) and (1.17') coincide on $[0, t_0 \wedge \bar{t}_0]$, $a_t = \bar{a}_t$ for all $t \in [0, t_0 \wedge \bar{t}_0]$. Thus, $\bar{t}_0 = t_0$ and $a_t^1 = a_t^2 = a_t$ for all $t \leq t_0$ implying that $\int_0^t \left| \mathbf{1}_{a_s^1 > 0} - \mathbf{1}_{a_s^2 > 0} \right| ds = 0$, for all $t \leq t_0$.

To conclude the proof of (1.49), it remains to show that a^1 and a^2 stay at zero after time t_0. Indeed, this fact is claimed by the Proposition 1.2.

Consequently, from (1.48) and (1.49), we have

$$|a_t^1 - a_t^2| \leq \int_0^t |\partial_a g(s, \xi_s)||a_s^1 - a_s^2|ds. \tag{1.50}$$

And because $f_2(.,.,b)$ is differentiable, we also have

$$|b_t^1 - b_t^2| \leq \int_0^t \max_{a\in[0,1]} |\partial_b f_2(s, a, \zeta_s)||b_s^1 - b_s^2|ds, \tag{1.51}$$

where ζ_s is a value between b_s^1 and b_s^2, that is $\min(b_s^1, b_s^2) \leq \zeta_s \leq \max(b_s^1, b_s^2)$. Applying the Gronwall's inequality, we obtain

$$\begin{aligned}
&|a_t^1 - a_t^2| + |b_t^1 - b_t^2| \\
&\leq (|a_0^1 - a_0^2| + |b_0^1 - b_0^2|) \exp\left(\int_0^t \left[|\partial_a f_1(s, \xi_s)| + \max_{a\in[0,1]} |\partial_b f_2(s, a, \zeta_s)|\right] ds\right) = 0,
\end{aligned}$$

for all t in $[0, 1]$. That means the equations (1.17)-(1.18) have at most one solution. ■

The function $(a(t, x), b(t, x))$ is continuous, then by Lemma 1.3, Proposition 1.3, we conclude that every subsequence $(X^{\ell_N})_{N\geq 1} \subset (X_N)_{N\geq 1}$ converges in distribution to a solution of the differential equations (1.17)-(1.18). And because of the uniqueness of the solution of (1.17)-(1.18), which is proved above, we conclude that the sequence $(X^N)_{N\geq 1} = (A^N, B^N)_{N\geq 1}$ converges in distribution to that unique solution.

1.4 The central limit theorem

For every $N \in \mathbb{N}^*$, let us define:

$$\tau_0^N := \inf\{t > 0, A_t^N = 0\}. \tag{1.52}$$

When the underlying networks are supercritical Erdös-Rényi graphs: $ER(N, \lambda/N)$, $\lambda > 1$, the size of the largest and the second largest components (by Theorem 0.5) is approximated as $|\mathcal{C}_{max}| = O(N)$ and $|\mathcal{C}_{(2)}| = O(\log(N))$ as N tends to infinity.
The probability that one of the initial A_0 individuals belongs to the giant component converges to 1. Indeed, we can consider that our initial condition consists of the first nodes explored until $\lfloor \|x_0\| N \rfloor$ individuals are discovered. Each time there is no more coupon, a new seed is chosen uniformly in the population, of which the giant component represents a proportion ζ_λ (see Theorem 0.5). Hence, the number of seeds S until we first hit the giant component follows roughly a Geometric distribution with parameter ζ_λ. Since for seeds outside the giant component, the associated exploration trees are of size at most $\log(N)$, the number of individuals discovered before finding the giant component is of order $\log(N) < \lfloor \|x_0\| N \rfloor$. Under the assumption 1.1, there is a positive fraction of seeds belonging to the giant component of $ER(N, \lambda/N)$ with a probability converging to 1.
For all $n = \lfloor Nt \rfloor$ with $t \geq \tau_0^N$, the RDS process restarts by choosing new seeds from the next components, whose sizes are at most $O(\log(N))$ (by Theorem 0.5) as N tends to infinity. When we normalize the process $(A_n)_{n\geq \lfloor N\tau_0^N \rfloor}$ by the size of the

population N, the normalized process $(A_t^N; t \geq \tau_0^N)_{N\geq 1}$ converges in probability uniformly to 0.

For the central limit theorem, we are interested in the limit of the RDS process in the giant component of $ER(N, \lambda/N), \lambda > 1$. By the proposition 1.5, we see that the Markov process $(A_t^N)_{N\geq 1}$ absorbs after the time t_0 with probability approximately 1 as N tends to infinity. Thus, in the sequels, we work conditionally on $\{\tau_0^N \geq t_0\}$ and all the processes are treated only in the interval $[0, t_0]$.

We now consider the process

$$W_t^N := \frac{X_{\lfloor Nt \rfloor} - N(a_t, b_t)}{\sqrt{N}} = \sqrt{N}(X_t^N - x_t), t \in [0, t_0], N \in \mathbb{N}^*. \tag{1.53}$$

Assumption 1.2 Let $W_0 = (W_0^1, W_0^2)$ be a Gaussian vector: $W_0 \sim \mathcal{N}(0; \Sigma)$. Assume that $W_0^N = \sqrt{N}(X_t^N - x_0)$ converges in distribution to W_0 as $N \to \infty$.

Theorem 1.2 Under Assumption 1.2, conditionally on $\{\tau_0^N \geq t_0\}$, the process $(W^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}([0, t_0], \mathbb{R}^2)$ to Y, which satisfies

$$W_t = W_0 + \int_0^t G(s, a_s, b_s, W_s)ds + M(t, a_t, b_t) \tag{1.54}$$

where

$$G(t, a, b, w) := \begin{pmatrix} \phi'(t + a)w^1 \\ -\lambda(w^1 + w^2) - \phi'(t + a)w^1 \end{pmatrix}; \tag{1.55}$$

$$\phi(z) := c - \sum_{k=0}^{c-1}(c - k)\frac{\lambda^k(1 - z)^k}{k!}e^{-\lambda(1-z)}, \tag{1.56}$$

and $\phi'(z)$ is the derivative with respect to z of ϕ; M is a zero-mean martingale with the quadratic variation

$$\langle M(\cdot, a., b.)\rangle_t := \left(\int_0^t m_{ij}(s, a_s, b_s)ds\right)_{i,j\in\{1,2\}}, \tag{1.57}$$

in which

$$m_{11}(t, a, b) := \sum_{k=0}^{c}(c - k)^2 p_k(t + a) - \left(\sum_{k=0}^{c}(c - k)p_k(t + a)\right)^2; \tag{1.58}$$

$$m_{22}(t, a, b) := \lambda(1 - t - a - b) + 2\lambda(1 - t - a - b) \times \left(c(\lambda - 1) + \sum_{k=0}^{c} p_k(t + a)\right) + m_{11}(t, a, b); \tag{1.59}$$

$$m_{12}(t,a,b) := \lambda(1-t-a-b)\left(c(\lambda-1)+\sum_{k=0}^{c} p_k(t+a)\right) - m_{11}(t,a,b). \tag{1.60}$$

The proof is divided into several steps: first, we write W^N in the form of a Doob's composition; then we claim the tightness of the sequence $(W^N)_{N\geq 1}$ in $\mathcal{D}([0,t_0],\mathbb{R}^2)$ by proving the tightness of both terms: the finite variation part and the martingale; next, we identify the limiting values of the sequence $(W^N)_{N\geq 1}$; and finally we demonstrate that all the limiting values are the same.

Recall from Lemma 1.1 that:

$$\begin{pmatrix} X_t^{N,1} \\ X_t^{N,2} \end{pmatrix} = \begin{pmatrix} A_0^N \\ B_0^N \end{pmatrix} + \begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \end{pmatrix} + \begin{pmatrix} M_t^{N,1} \\ M_t^{N,2} \end{pmatrix},$$

where

$$\Delta_t^{N,1} = \frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\left\{c - \sum_{k=0}^{c-1}(c-k)\mathbb{P}(Y_i=k|\mathcal{F}_{i-1}) - 1\right\},$$

$$\Delta_t^{N,2} = \frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\left\{\lambda\left(1-\frac{i}{N}-\frac{A_{i-1}}{N}-\frac{B_{i-1}}{N}\right) - \left(c-\sum_{k=0}^{c-1}(c-k)\mathbb{P}(Y_i=k|\mathcal{F}_{i-1})\right)\right\},$$

and where

$$\langle M^N\rangle_t = \begin{bmatrix} \langle M^{N,1},M^{N,1}\rangle_t & \langle M^{N,1},M^{N,2}\rangle_t \\ \langle M^{N,2},M^{N,1}\rangle_t & \langle M^{N,2},M^{N,2}\rangle_t \end{bmatrix}. \tag{1.61}$$

From the proof of Lemma 1.1, we recall the equation (1.40):

$$\left\|\Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\left(\frac{n-1}{N},\frac{A_{n-1}}{N},\frac{B_{n-1}}{N}\right)\right\| \leq \frac{C}{N}, \tag{1.62}$$

where f is defined in (1.39): $f(t,a,b) = (f_1(t,a,b), f_2(t,a,b))$,

$$f_1(t,a) := c - \sum_{k=0}^{c-1}(c-k)p_k(t+a) - 1$$

$$f_2(t,a,b) := (1-t-a-b)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t+a) - c.$$

and recall the components of the quadratic variation $\langle M^N\rangle_t$ given by (1.24):

$$\langle M^{N,1}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathrm{Var}(Y_n\wedge c|\mathcal{F}_{n-1}),$$

$$\langle M^{N,1}, M^{N,2}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor} \mathrm{Cov}(Y_n \wedge c, H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}),$$

$$\langle M^{N,2}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor} \mathrm{Var}(H_n - Y_n \wedge c|\mathcal{F}_{n-1}).$$

Notice that in this section, we work conditionally on $\{\tau_N^0 \geq t_0\}$ and that all processes are defined in the time interval $[0, t_0]$, then all terms $\mathbf{1}_{A_{i-1}\geq 1}$, $1 \leq i \leq \lfloor Nt_0\rfloor$, $\mathbf{1}_{A_t^N>0}$, $\mathbf{1}_{a_t>0}$ are replaced by 1.

For all $N \in \mathbb{N}^*$ and for all $t \in [0, t_0]$, W_t^N is written as:

$$\begin{aligned} W_t^N =& \sqrt{N}\begin{pmatrix} A_0^N - a_0 \\ B_0^N - b_0\end{pmatrix} + \sqrt{N}\begin{pmatrix} \Delta_t^{N,1} - \int_0^t f_1(s, a_s, b_s)ds \\ \Delta_t^{N,2} - \int_0^t f_2(s, a_s, b_s)ds\end{pmatrix} + \sqrt{N}\begin{pmatrix} M_t^{N,1} \\ M_t^{N,2}\end{pmatrix} \\ =& W_0^N + \widetilde{\Delta}_t^N + \widetilde{M}_t^N. \end{aligned}$$

We prove tightness of the process in $\mathcal{D}([0, t_0], \mathbb{R}^2)$ and then identify the limiting values.

1.4.1 Tightness of the process $(W^N)_{N\geq 1}$

Proposition 1.6 The sequence $(W^N)_{N\geq 1}$ is tight in $\mathcal{D}([0, t_0], \mathbb{R}^2)$.

Proof. To prove that the distributions of the semi-martingales $(W^N)_{N\geq 1}$ form a tight family, we use the Aldous-Rebolledo criterion as in Lemma 1.2. To achieve this, we start with establishing some moment estimates that will be useful.

Step 1: moment estimates

From (1.38), we have

$$\mathbb{E}[\|\langle \widetilde{M}^N\rangle_t\|] \leq (6c^2 + 4\lambda)t.$$

For the term $\tilde{\Delta}_t^N$:

$$\begin{aligned} |\widetilde{\Delta}_t^{N,1}| \leq \sqrt{N}\Bigg|\Delta_t^{N,1} - \frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\Bigg\{c - \sum_{k=0}^{c}(c-k)p_k\left(\frac{i-1}{N} + \frac{A_{i-1}}{N}\right) - 1\Bigg\}\Bigg| \\ + \sqrt{N}\Bigg|\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\Bigg\{c - \sum_{k=0}^{c}(c-k)p_k\left(\frac{i-1}{N} + \frac{A_{i-1}}{N}\right) - 1\Bigg\} \\ - \sum_{i=1}^{\lfloor Nt\rfloor}\int_{(i-1)/N}^{i/N}\left(c - \sum_{k=0}^{c}(c-k)p_k\left(s + a_s\right) - 1\right)ds\Bigg| \end{aligned}$$

$$+\sqrt{N}\left|\int_{\lfloor Nt\rfloor/N}^{t}\left(c-\sum_{k=0}^{c}(c-k)p_k\left(s+a_s\right)-1\right)ds\right|. \tag{1.63}$$

Thanks to (1.62), we have that

$$\begin{aligned}&\sqrt{N}\left|\Delta_t^{N,1}-\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\left\{c-\sum_{k=0}^{c}(c-k)p_k\left(\frac{i-1}{N}+\frac{A_{i-1}}{N}\right)-1\right\}\right|\\ \le&\sqrt{N}\left\|\Delta_t^{N}-\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}f\left(\frac{i-1}{N},\frac{A_{i-1}}{N},\frac{B_{i-1}}{N}\right)\right\|\le\frac{C}{\sqrt{N}}.\end{aligned}$$

Because f_1 is continuous and is defined in a compact set $[0,1]^3$, then the third term in the r.h.s. of (1.63) is upper bounded by $\frac{\max_{(t,a,b)\in[0,1]^3}|f_1(t,a,b)|}{\sqrt{N}}$.

For all $s\in\left[\frac{i-1}{N},\frac{i}{N}\right)$,

$$\left|p_k(s+a_s)-p_k\left(\frac{i-1}{N}+\frac{A_{i-1}}{N}\right)\right|\le\left(\left|s-\frac{i-1}{N}\right|+\left|a_s-A^N_{\frac{i-1}{N}}\right|\right)\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right| \tag{1.64}$$

$$\le\left(\frac{1}{N}+\left|\frac{W_s^{N,1}}{\sqrt{N}}\right|\right)\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|. \tag{1.65}$$

The second term in the r.h.s. of (1.63) is bounded by

$$\begin{aligned}&\sqrt{N}\sum_{i=1}^{\lfloor Nt\rfloor}\sum_{k=0}^{c}(c-k)\int_{(i-1)/N}^{i/N}\left|p_k(s+a_s)-p_k\left(\frac{i-1}{N}+\frac{A_{i-1}}{N}\right)\right|ds\\ &\le\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|\frac{c(c-1)}{2}\left(\frac{1}{\sqrt{N}}+\int_0^t|W_s^{N,1}|ds\right).\end{aligned}$$

Thus,

$$\begin{aligned}|\widetilde{\Delta}_t^{N,1}|\le&\frac{C+\max_{(t,a,b)\in[0,1]^3}|f_1(t,a,b)|+\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|\frac{c(c-1)}{2}}{\sqrt{N}}\\ &+\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|\frac{c(c-1)}{2}\int_0^t|W_s^{N,1}|ds.\end{aligned}$$

Using the similar argument, we have that

$$\begin{aligned}|\widetilde{\Delta}_t^{N,2}|\le&\frac{C+\sup_{(t,a,b)\in[0,1]^3}|f_2(t,a,b)|+\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|\frac{c(c-1)}{2}+\lambda}{\sqrt{N}}\\ &+\left(\sup_{z\in[0,1]}\left|\frac{dp_k}{dz}(z)\right|\frac{c(c-1)}{2}+\lambda\right)\int_0^t|W_s^{N,1}|ds+\lambda\int_0^t|W_s^{N,2}|ds.\end{aligned}$$

Hence,

$$\|\widetilde{\Delta}_t^N\| \leq \frac{C'(\lambda,c)}{\sqrt{N}} + C''(\lambda,c)\int_0^t \|W_s^N\|ds \tag{1.66}$$

Then for every $t \in [0,t_0]$,

$$\begin{aligned}\mathbb{E}[\|W_t^N\|] &\leq \mathbb{E}[\|\widetilde{\Delta}_t^N\|] + \mathbb{E}[\|\widetilde{M}_t^N\|] \\ &\leq (6c^2+4\lambda)t + \frac{C'(\lambda,c)}{\sqrt{N}} + C''(\lambda,c)\int_0^t \mathbb{E}[\|W_s^N\|]ds.\end{aligned}$$

And thus by the Grönwall's inequality, we deduce that

$$\sup_{t\in[0,t_0]} \mathbb{E}[\|W_t^N\|] \leq (6c^2+4\lambda+C'(\lambda,c))e^{C''(\lambda,c)} = C''', \quad \forall N \geq 1. \tag{1.67}$$

Let $0 \leq s < t \leq t_0$,

$$\begin{aligned}\mathbb{E}[\|W_t^N - W_s^N\|] &\leq \frac{C'(\lambda,c)(t-s)}{\sqrt{N}} + (6c^2+4\lambda)(t-s) + C''(\lambda,c)\int_s^t \mathbb{E}[\|W_u^N\|]du, \\ &\leq (C'(\lambda,c)+6c^2+4\lambda+C''(\lambda,c)C''')(t-s)\end{aligned}$$

Then for given $\varepsilon > 0, \eta > 0$, choose δ such that $\delta < \eta\varepsilon(C'(\lambda,c)+6c^2+4\lambda+C''(\lambda,c)C''')^{-1}$,

$$\mathbb{P}\left(\sup_{\substack{|t-s|<\delta \\ 0\leq s<t\leq 1}} \|W_t^N - W_s^N\| > \eta\right) \leq \eta^{-1}\mathbb{E}\left[\sup_{\substack{|t-s|<\delta \\ 0\leq s<t\leq 1}} \|W_t^N - W_s^N\|\right] < \varepsilon. \tag{1.68}$$

By (1.67) and (1.68), we can conclude that $(W^N)_{N\geq 1}$ is tight in $\mathcal{D}([0,t_0],\mathbb{R}^2)$. ■

Proposition 1.7 The martingale $(\widetilde{M}^N)_{N\geq 1}$ converges in distribution to a Gaussian process $(M_t)_{0\leq t\leq t_0}$ on $[0,t_0]$.

Proof. Keeping in mind that $A_n - A_{n-1} = Y_n \wedge c - 1$ and $B_n - B_{n-1} = H_n - Y_n \wedge c$ and by (1.33), we have

$$\langle \widetilde{M}^{N,1}\rangle_t = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left\{\mathbb{E}\left[(Y_n\wedge c)^2\,|\mathcal{F}_{n-1}\right] - \left(\mathbb{E}\left[Y_n\wedge c|\mathcal{F}_{n-1}\right]\right)^2\right\}; \tag{1.69}$$

$$\begin{aligned}\langle \widetilde{M}^{N,2}\rangle_t = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\Big\{&\mathrm{Var}(H_n|\mathcal{F}_{n-1}) - 2\Big(\mathbb{E}[H_n(Y_n\wedge c)|\mathcal{F}_{n-1}] \\ &- \mathbb{E}[H_n|\mathcal{F}_{n-1}]\mathbb{E}[Y_n\wedge c|\mathcal{F}_{n-1}]\Big)\Big\} + \langle \widetilde{M}^{N,1}\rangle_t;\end{aligned} \tag{1.70}$$

$$\langle \widetilde{M}^{N,1}, \widetilde{M}^{N,2} \rangle_t = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \Big\{ \mathbb{E}\left[H_n(Y_n \wedge c) \big| \mathcal{F}_{n-1} \right] - \mathbb{E}\left[H_n | \mathcal{F}_{n-1} \right] \mathbb{E}\left[Y_n \wedge c | \mathcal{F}_{n-1} \right] \Big\} - \langle \widetilde{M}^{N,1} \rangle_t \quad (1.71)$$

From (1.69), (1.28), (1.29) and (1.45),

$$\left| \langle \widetilde{M}^{N,1} \rangle_t - \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} m_{11}\left(\frac{i-1}{N}, \frac{A_{i-1}}{N}, \frac{B_{i-1}}{N} \right) \right|$$
$$\leq \sum_{k=0}^{c} (c-k)^2 \frac{C(\lambda,k)}{N} + \sum_{k,\ell=0}^{c} \left(\frac{(c-k)C(\lambda,k)}{N} + \frac{(c-\ell)C(\lambda,\ell)}{N} \right) \leq \frac{D_1(\lambda,c)}{N}.$$

From (1.70), (1.30), (1.31) and (1.45),

$$\left| \langle \widetilde{M}^{N,2} \rangle_t - \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} m_{22}\left(\frac{i-1}{N}, \frac{A_{i-1}}{N}, \frac{B_{i-1}}{N} \right) \right| \leq \frac{D_2(\lambda,c)}{N} + \frac{D_1(\lambda,c)}{N},$$

where $D_2(\lambda,c) = \lambda + 2\sum_{k=0}^{c}(k^2 - ck)C(\lambda,k) + 2c\lambda + 1 + \sum_{k=0}^{c}(c-k)C(\lambda,k)$ and from (1.71), (1.31),

$$\left| \langle \widetilde{M}^{N,1}, \widetilde{M}^{N,2} \rangle_t - \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} m_{12}\left(\frac{i-1}{N}, \frac{A_{i-1}}{N}, \frac{B_{i-1}}{N} \right) \right| \leq \frac{D_3(\lambda,c)}{N} + \frac{D_1(\lambda,c)}{N},$$

where $D_3(\lambda,c) = \sum_{k=0}^{c}(k^2 - ck)C(\lambda,k) + c\lambda$. And since the vectorial function $(m_{k\ell})_{1\leq k,\ell\leq 2}$ are continuous, then by Lemma 1.3, we obtain that $\langle \widetilde{M}^N \rangle_t$ converges uniformly in distribution to $\int_0^t (m_{k,\ell}(s,a_s,b_s))_{k,\ell\in\{1,2\}} ds$. By Theorem 2 in [100], we can conclude that $(M^N)_{N\geq 1}$ converges uniformly in distribution to the Gaussian process $(M_t)_{t\in[0,t_0]}$, which is identified by its quadratic variation $\langle M \rangle_t = \int_0^t (m_{ij}(s,a_s,b_s))_{i,j\in\{1,2\}} ds$. ■

Proposition 1.8 The finite variation $\left(\widetilde{\Delta}_t^N, t \in [0,t_0] \right)_{N\geq 1}$ converges in distribution to the process $\left(\Delta_t, t \in [0,t_0] \right)$, which is the unique solution of the stochastic differential

$$\Delta_t = \int_0^t G(s,a_s,b_s,W_s)dt \quad (1.72)$$

Proof.

$$\widetilde{\Delta}_t^N = \sqrt{N} \left(\Delta_t^N - \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} f\left(\frac{i-1}{N}, A^N_{\frac{i-1}{N}}, B^N_{\frac{i-1}{N}} \right) \right)$$

$$+\left(\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\sqrt{N}f\left(\frac{i-1}{N},A^N_{\frac{i-1}{N}},B^N_{\frac{i-1}{N}}\right)-\int_0^t\sqrt{N}f(s,a_s,b_s)ds\right)$$
$$=D^N_t+E^N_t, \tag{1.73}$$

where

$$f(t,a,b):=\begin{pmatrix}c-\sum_{k=0}^{c-1}(c-k)\frac{\lambda^k}{k!}(1-t-a)^ke^{-\lambda(1-t-a)}-1\\(1-t-a-b)\lambda-c+\sum_{k=0}^{c}(c-k)\frac{\lambda^k}{k!}(1-t-a)^ke^{-\lambda(1-t-a)}\end{pmatrix}$$
$$=\begin{pmatrix}f_1(t,a,b)\\f_2(t,a,b)\end{pmatrix} \tag{1.74}$$

From (1.62), we have

$$\|D^N_t\|=\left\|\sqrt{N}\left(\Delta^N_t-\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}f\left(\frac{i-1}{N},A^N_{\frac{i-1}{N}},B^N_{\frac{i-1}{N}}\right)\right)\right\|\le\frac{C(\lambda,c)}{\sqrt{N}}.$$

We need to find the limit of E^N_t.

$$E^N_t=\sum_{i=1}^{\lfloor Nt\rfloor}\sqrt{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}}\left(f(\frac{i-1}{N},A^N_{\frac{i-1}{N}},B^N_{\frac{i-1}{N}})-f(s,a_s,b_s)\right)ds-\sqrt{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^{t}f(s,a_s,b_s)ds \tag{1.75}$$

Because f is continuous function, defined in the compact set $[0,1]^3$, the second term in the r.h.s. of (1.75) is bounded by $\frac{\max_{(t,a,b)\in[0,1]^3}\|f(t,a,b)\|}{\sqrt{N}}$ and thus converges to 0 as $N\to\infty$.

We write f as

$$f(t,a,b)=\begin{pmatrix}\phi(t+a)\\\psi(t+a+b)-\phi(t+a)\end{pmatrix}$$

where $\phi(z)=c-\sum_{k=0}^{c-1}(c-k)\frac{\left[\lambda(1-z)\right]^k}{k!}e^{-\lambda(1-z)}$ and $\psi(z)=\lambda(1-z)$. Then

$$\begin{aligned}&\phi\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)-\phi(s+a_s)\\=&\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)\left((\frac{i-1}{N}-s)+(A^N_{\frac{i-1}{N}}-a_s)\right)\\&-\phi''(\xi_{i,s})\left((\frac{i-1}{N}-s)+(A^N_{\frac{i-1}{N}}-a_s)\right)^2\\=&(\frac{i-1}{N}-s)\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)+\frac{W^{N,1}_s}{\sqrt{N}}\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)\end{aligned}$$

$$- \left((\frac{i-1}{N} - s) + \frac{W_s^{N,1}}{\sqrt{N}} \right)^2 \phi''(\xi_{i,s}),$$

where $\xi_{i,s}$ takes the value between $\frac{i-1}{N} + A^N_{\frac{i-1}{N}}$ and $s + a_s$; $\phi'(\xi_{i,s})$ (*resp.* $\phi''(\xi_{i,s})$)) is first derivative (*resp.* the second derivative) of ϕ at $\xi_{i,s}$. And

$$\begin{aligned}
&\psi\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}} + B^N_{\frac{i-1}{N}}\right) - \psi(s + a_s + b_s) \\
&= -\lambda\left((\frac{i-1}{N} - s) + (A^N_{\frac{i-1}{N}} - a_s) + (B^N_{\frac{i-1}{N}} - b_s)\right) \\
&= -\lambda\left((\frac{i-1}{N} - s) + \frac{W_s^{N,1}}{\sqrt{N}} + \frac{W_s^{N,2}}{\sqrt{N}}\right).
\end{aligned}$$

So the first term in the right hand side of (1.75) can be written as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor} \begin{pmatrix} W^{N,1}_{\frac{i-1}{N}} \phi'\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}}\right) \\ -\lambda\left(W^{N,1}_{\frac{i-1}{N}} + W^{N,2}_{\frac{i-1}{N}}\right) - \phi'\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}}\right) W^{N,1}_{\frac{i-1}{N}} \end{pmatrix} \\
&\quad + \sum_{i=1}^{\lfloor Nt\rfloor} \begin{pmatrix} \int_{\frac{i-1}{N}}^{\frac{i}{N}} \left\{ \sqrt{N}\big(\frac{i-1}{N} - s\big)\phi'\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}}\right)\right\} ds \\ -\int_{\frac{i-1}{N}}^{\frac{i}{N}} \left\{ \sqrt{N}\big(\frac{i-1}{N} - s\big)\left(1 + \phi'\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}}\right)\right)\right\} ds \end{pmatrix} \\
&\quad + \sum_{i=1}^{\lfloor Nt\rfloor} \begin{pmatrix} -\int_{\frac{i-1}{N}}^{\frac{i}{N}} \sqrt{N}\left\{ \left(\big(\frac{i-1}{N} - s\big) + \frac{W_s^{N,1}}{\sqrt{N}}\right)^2 \phi''(\xi_{i,s})\right\} ds \\ \int_{\frac{i-1}{N}}^{\frac{i}{N}} \sqrt{N}\left\{ \left(\big(\frac{i-1}{N} - s\big) + \frac{W_s^{N,1}}{\sqrt{N}}\right)^2 \phi''(\xi_{i,s})\right\} ds \end{pmatrix}
\end{aligned} \tag{1.76}$$

Because $(W^N)_{N\geq 1}$ is tight, there exists a subsequence of $(W^N)_{N\geq 1}$, denoted again $(W^N)_{N\geq 1}$, which converges in distribution to $W = (W^1, W^2) \in \mathcal{D}([0, t_0], \mathbb{R}^2)$. The second and the third term of (1.76) converge in distribution to 0 since

$$\sum_{i=1}^{\lfloor Nt\rfloor} \int_{\frac{i-1}{N}}^{\frac{i}{N}} \sqrt{N}\left|(\frac{i-1}{N} - s)\phi'\left(\frac{i-1}{N} + A^N_{\frac{i-1}{N}}\right)\right| ds \leq \sup_{z\in[0,1]} |\phi'(z)| N^{-1/2},$$

and with $\widetilde{W}^N \overset{(d)}{=} W^N$ defined as in the Skorokhod's representation Theorem, $\widetilde{W}^N$ converges uniformly almost surely to $\widetilde{W} \overset{(d)}{=} W$, we have $(\widetilde{W}^N)_{N\geq 1}$ is bounded and that

$$\sum_{i=1}^{\lfloor Nt\rfloor}\int_{\frac{i-1}{N}}^{\frac{i}{N}}\sqrt{N}\left|\left((\frac{i-1}{N}-s)+\frac{\widetilde{W}_s^{N,1}}{\sqrt{N}}\right)^2\phi''(\xi_{i,s})\right|ds$$
$$\leq\left(\sup_{z\in[0,1]}|\phi''(z)|+\sup_{N\geq 1}\|\widetilde{W}^{N,1}\|\right)N^{-1/2}.$$

Then by Lemma 1.3, $(\widetilde{\Delta}^N)_{N\geq 1}$ converges in distribution to a process, which satisfies equation

$$\widetilde{\Delta}_t=\int_0^t\begin{pmatrix}\phi'(s+a_s)W_s^1\\ -\lambda(W_s^1+W_s^2)-\phi'(s+a_s)W_s^1\end{pmatrix}ds \tag{1.77}$$

■

1.4.2 The uniqueness of the SDEs

Since the process $(W^N)_{N\geq 1}$ defined in a closed interval: $[0,t_0]$ and tight in $\mathcal{D}([0,1];\mathbb{R}^2)$, so uniqueness of the solution of the SDE (1.54) is proved if the criteria in Theorem 3.1 of [95, page 178] is verified. We need to justify that the functions $G(t,w_t)$ and $\sigma(t,w_t)=\langle M(\cdot,w.)\rangle_t$ are Lipschitz continuous, *i.e.* for every $N\geq 1$, there exists $K_N>0$ such that:

$$\|G(t,u)-G(t,w)\|+\|\sigma(t,u)-\sigma(t,w)\|\leq K_N\|u-w\|,\quad\forall u,w\in\mathcal{B}_N,$$

where $\mathcal{B}_N=\{x:\|x\|\leq N\}$.

Indeed, this condition holds because

$$\|G(t,u)-G(t,w)\|\leq\big(2\max_{z\in[0,1]}|\phi'(z)|+\lambda\big)\|u-w\|,$$

and $\sigma(t,w)$ does not depend on w. Hence, the pathwise uniqueness of solutions holds for the equation(1.54).

1.5 Some lemmas used in the proof

Lemma 1.3 Let f be a function in $\mathcal{C}_b([0,1]^3,\mathbb{R}^2)$, and let $(X^N)_{N\geq 1}$ be a sequence of stochastic processes in $\mathcal{D}([0,1],[0,1]^2)$. If $X^N\xrightarrow{(d)}X\in\mathcal{C}([0,1],[0,1]^2)$ for the Skorokhod topology on $\mathcal{D}([0,1],[0,1]^2)$, then

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}f\left(\frac{n-1}{N},X^N_{\frac{n-1}{N}}\right)\xrightarrow{(d)}\int_0^t f(s,X_s)ds.$$

Proof. Since $X^N \xrightarrow{(d)} X$, by Skorokhod's representation theorem [86, Th.25.6, p.287], there exist $\widetilde{X}^N \in \mathcal{D}([0,1],[0,1]^2)$ and $\widetilde{X} \in \mathcal{C}([0,1],[0,1]^2)$ defined on a common probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ such that $\widetilde{X}^N \overset{(d)}{=} X^N$, $\widetilde{X} \overset{(d)}{=} X$ and $\widetilde{X}^N \longrightarrow \widetilde{X}$ a.s. For any $t \in [0,1]$ and for any $N \in \mathbb{N}^*$,

$$\begin{aligned}
&\left| \frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - \int_0^t f(s,\widetilde{X}_s)ds \right| \\
&\leq \left| \frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - \sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} f(s,\widetilde{X}_s)ds \right| + \left| \int_{\frac{\lfloor Nt \rfloor}{N}}^{t} f(s,\widetilde{X}_s)ds \right| \\
&\leq \sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - f(s,\widetilde{X}_s) \right| ds + \frac{\|f\|_\infty}{N}.
\end{aligned}$$

Let $\varepsilon > 0$. From the uniform continuity of f, there exists a positive constant $\delta = \delta(x) > 0$ such that for all $(t,x),(t',x') \in [0,1]\times[0,1]^2$ satisfying $|t-t'| + \|x-x'\|_\infty < \delta$, $|f(t,x) - f(t',x')| < \varepsilon/2$. Now,

$$\begin{aligned}
&\sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - f(s,\widetilde{X}_s) \right| ds \\
&= \sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - f(s,\widetilde{X}_s) \right| \mathbf{1}_{\frac{1}{N} + \|\widetilde{X}^N_s - \widetilde{X}_s\|_1 \geq \delta} ds \\
&+ \sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - f(s,\widetilde{X}_s) \right| \mathbf{1}_{\frac{1}{N} + \|\widetilde{X}^N_s - \widetilde{X}_s\|_1 < \delta} ds.
\end{aligned}$$

Because $\widetilde{X}^N$ converges uniformly to $\widetilde{X}$ a.s., there exists $N_0(\omega)$ such that $\sup_{s\in[0,1]}(1/N + \|\widetilde{X}^N_s - \widetilde{X}_s\|) < \delta, \quad \forall N \geq N_0$ a.s. For $\mathbb{P}$-almost all $\omega \in \Omega$, when $N \geq \max(N_0(\omega), 2\|f\|_\infty/\varepsilon)$,

$$\begin{aligned}
&\left| \frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - \int_0^t f(s,\widetilde{X}_s)ds \right| \\
&\leq \frac{\sup\|f\|}{N} + \sum_{n=1}^{\lfloor Nt \rfloor} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| f\big(\frac{n-1}{N}, \widetilde{X}^N_{\frac{n-1}{N}}\big) - f(s,\widetilde{X}_s) \right| \mathbf{1}_{\frac{1}{N} + \|\widetilde{X}^N_s - \widetilde{X}_s\|_1 < \delta} ds \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

The upper bound is independent of t and thus we have that for all $\varepsilon > 0$:

$$\lim_{N\to+\infty}\sup_{t\in[0,1]}\left|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}f\Big(\frac{n-1}{N},\widetilde{X}^N_{\frac{n-1}{N}}\Big)-\int_0^t f(s,\widetilde{X}_s)ds\right|\le\varepsilon\quad\text{a.s..}\tag{1.78}$$

This finishes the proof.

■

Lemma 1.4 Denote

$$\phi(z):=c-\sum_{k=0}^{c-1}(c-k)\frac{\big[\lambda(1-z)\big]^k}{k!}e^{-\lambda(1-z)},\quad c\ge 2,\lambda>1.\tag{1.79}$$

Then there exists a unique $z_0\in[0,1]$ such that $\phi(z_0)=1$ and $z_0>1-1/\lambda$.

Proof. For all $z\in[0,1]$,

$$\begin{aligned}\phi'(z)=\;&-c\lambda e^{-\lambda(1-z)}+\lambda\sum_{k=1}^{c-1}(c-k)\frac{(\lambda(1-z))^{k-1}}{(k-1)!}e^{-\lambda(1-z)}\\&-\lambda\textstyle\sum_{k=1}^{c-1}(c-k)\frac{(\lambda(1-z))^k}{k!}e^{-\lambda(1-z)}\\=\;&\lambda e^{-\lambda(1-z)}\left[-c+\sum_{k=0}^{c-2}(c-k-1)\frac{(\lambda(1-z))^k}{k!}-\sum_{k=1}^{c-1}(c-k)\frac{(\lambda(1-z))^k}{k!}\right]\\=\;&\lambda e^{-\lambda(1-z)}\left[-1-\sum_{k=1}^{c-2}\frac{(\lambda(1-z))^k}{k!}-\frac{(\lambda(1-z))^{c-1}}{(c-1)!}\right]<0,\end{aligned}$$

which gives that ϕ is decreasing. Furthermore, we have $\phi(1-1/\lambda)>1$ for $c\ge 2$ and $\phi(1)=0$. So the equation $\phi(z)=1$ has unique root, denoted by $z_0\in(1-1/\lambda,1)$. ■

Lemma 1.5 We have that

$$\lim_{N\to\infty}\mathbb{P}(\tau_0^N\ge t_0)=1.\tag{1.80}$$

Proof. For $\varepsilon>0$, let

$$\tau_\varepsilon^N:=\inf\{t>0,A_t^N\le\varepsilon\}\tag{1.81}$$

and

$$t_\varepsilon:=\inf\{t>0,a_t\le\varepsilon\}.\tag{1.82}$$

Because A^N is càdlàg and a is continuous, $\inf_{t\in[0,1]}a_t\le\lim_{N\to\infty}\inf_{t\in[0,1]}A^N_{t\wedge\tau_\varepsilon^N}$. Then for any $0<\varepsilon<\varepsilon'$, by Fatou's lemma:

$$1 = \mathbb{P}(\inf_{t\in[0,t_{\varepsilon'}]} A_t^N > \varepsilon) \leq \mathbb{P}(\lim_{N\to\infty} \inf_{t\in[0,t_{\varepsilon'}]} A_{t\wedge\tau_\varepsilon^N}^N > \varepsilon) = \lim_{N\to\infty} \mathbb{P}(\tau_\varepsilon^N > t_{\varepsilon'}).$$

Let $\varepsilon' \to 0$, we have

$$\lim_{N\to\infty} \mathbb{P}(\tau_0^N \geq t_0) = 1. \tag{1.83}$$

■

2. The RDS process on Stochastic Block Model

Contents

The work in this chapter is submitted to the journal ESAIM[1], under the major revision [105].

The discovery of the “hidden population", whose size and membership are unknown, is made possible by assuming that its members are connected in a social network by their relationships. We explore these groups by a chain-referral sampling (CRS) method, where participants recommend the people they know. This leads to the study of a Markov chain on a random graph where vertices represent individuals and edges connecting any two nodes describe the relationships between corresponding people. We are interested in the study of CRS process on the stochastic block model (SBM), which extends the well-known Erdös-Rényi graphs to populations partitioned into communities. The SBM considered here is characterized by a number of vertices N, a number of communities (blocks) m, proportion of each community $\pi = (\pi_1, ..., \pi_m)$

[1] European Series in Applied and Industrial Mathematics

and a pattern for connection between blocks $P = (\lambda_{k\ell}/N)_{(k,\ell)\in\{1,\dots,m\}^2}$. In this paper, we give a precise description of the dynamic of CRS process in discrete time on an SBM. The difficulty lies in handling the heterogeneity of the graph. We prove that when the population's size is large, the normalized stochastic process of the referral chain behaves like a deterministic curve which is the unique solution of a system of ODEs.

2.1 Introduction

In Sociology, some populations may be hidden because their members share common attributes that are illegal or stigmatized. These hidden groups may be hard to approach because these individuals try to conceal their identities due to legal authorities (e.g. drugs users) or because of the social pressure (e.g. men having sex with men). In such populations, all the information is unknown: there is no sampling frame such as lists of the members of the population or of the relationship between the latter. It causes many challenges for researchers to identify these groups. The discovery of the hidden populations is made possible by assuming that its members are connected by a social network. The population is described by a graph (network) where each individual is represented by a vertex and any interaction or relationship (e.g. friendship, partnership) between a couple of individuals is represented by an edge matching the corresponding vertices. Thanks to this important feature, we are allowed to investigate these populations by using a Chain-referral Sampling (CRS) technique, such as snowball sampling, targeting sampling, respondent driven sampling etc. (see the review of [78] or [47, 49, 50]). CRS consists in detecting hidden individuals in a population structured as a random graph, which is modeled by a stochastic process that we study here. The principle of CRS is that from a group of initially recruited individuals, we follow their connections in the social network to recruit the subsequent participants. The exploration proceeds from node to node along the edges of the graph. The interviewees induce a sub-tree of the underlying real graph, and the information coming from the interviews gives knowledge on other non-interviewed individuals and edges, providing a larger sub-graph. We aim at understanding this recruitment process from the properties of the explored random graph. The CRS showed its practicality and efficiency in recruiting a diverse sample of drug users (see [6]).

CRS models are hard to study from a theoretical point of view without any assumption on the graph structure. In this paper, we consider a particular model with latent community structure: the stochastic block model (SBM) proposed by Holland et al.[54]. This model is a useful benchmark for some statistical tasks as recovering community (also called blocks or types in the sequel) structure in network science [7, 42, 46]. By block structure, we mean that the set of vertices in the graph is partitioned into subsets called blocks and nodes connect to each other with probabilities that depend only on their types, *i.e.* the blocks to which they belong. For example, edges may be more common within a block than between blocks (e.g. group of people having sexual contacts). We recall here the definition of SBM (we refer the reader to

the survey in [1]):

Definition 2.1 Let N be a positive integer (number of vertices), m be a positive integer (number of blocks or types), $\pi = (\pi_1, ..., \pi_m)$ be a probability distribution on $\{1, \dots m\}$ (the probabilities of the m types, *i.e.* a vector of $[0,1]^m$ such that $\sum_{k=1}^{m} \pi_k = 1$) and $P = (p_{k\ell})_{(k,\ell)\in\{1,...,m\}^2}$ be a symmetric matrix with entries $p_{kl} \in [0,1]$ (connectivity probabilities). The pair (Γ, G) is drawn under the distribution SBM(N, π, P) if the vector of types Γ is an N-dimensional random vector, whose components are i.i.d., $\{1, \dots, m\}$-valued with the law π, and G is a simple graph of size N where vertices i and j are connected independently of other pairs of vertices with probability $p_{\Gamma_i\Gamma_j}$. We also denote the blocks (community sets) by: $[\ell] := \{v \in \{1, ..., N\} : \Gamma_v = \ell\}$ with the size $N_\ell := |[\ell]|, l \in \{1, ..., m\}$.

Notice that when $m = 1$, *i.e.* there is only one type. Any arbitrary pair of vertices is connected independently to the others with the same probability p_{11}, SBM becomes the Erdös-Rényi graph, which is studied in [102].
Here, we consider the Poisson case where the connectivity probabilities $p_{k\ell}$ depend on N and are given by $p_{k\ell} = \lambda_{k\ell}/N$. This means that each individual of the block k contacts in average $\lambda_{k\ell}\pi_\ell$ individuals of the block ℓ. This implies that the network examined is sparse. In the present work, we give a rigorous description of a CRS on such SBM and study the propagation of the referral chain on this sparse model.

The CRS relies on a random peer-recruitment process. To handle the two sources of randomness, the graph and the exploring process on it are constructed simultaneously. In the construction, the vertices of the graph will be in 3 different states: inactive vertices that have not being contacted for interviews, active vertices that constitute the next interviewees and off-mode vertices that have been already interviewed. The idea to describe the random graph as a Markov exploration process with active, explored and unexplored nodes is classical in random graphs theory. It has been used as a convenient technique to expose the connections inside a cluster, especially to discover the giant component in a random graph models, for example see [92, 94]. In our case, there is a slight difference in the recruiting process: the number of nodes being switched to the active mode is set to be bounded by a constant. This trick helps to improve the bias towards high-degree nodes in the population (see [50]). At the beginning of the survey, all individuals in the population are hidden and are marked as inactive vertices. We choose some people as seeds of the investigation and activate them. During the interview these individuals name their contacts and a maximum number c of coupons are distributed to the latter, who become active nodes. One by one, every carrier of a coupon can come to a private interview and is asked in turn to give the names of her/his peers. Whenever a new person is named, one edge connecting the interviewee and her/his contact is added but they remain inactive until they receive a coupon. After finishing the interview, a maximum number of c new contacts receive one coupon each and are activated. So if the interviewee

names more than c people, a number of them are not given any coupon and can be still explored later provided another interviewee mentions them. After that, the node associated to the person who has just been interviewed is switched to off-mode and is no longer recruited again, see Figure 1.1. We repeat the procedure of interviewing, referring, distributing coupons until there is no more active vertex in the graph (no more coupon is returned). Each person returning a coupon receives some money as a reward for her/his participation, and an extra bonus depending on the number contacts that will later return the coupons. Notice that each individual in the population is interviewed just once and we assume here that there is no restriction on the total number of coupons.

The process of interest counts the number of coupons present in the population. We also want to know how many people are detected, which leads to the number of people explored but without coupons. Denote by the discrete time $n \in \mathbb{N} = \{0, 1, 2, \dots\}$ the number of interviews completed, A_n corresponds to the number of individuals that have received coupons but that have not been interviewed yet (number of active vertices); B_n to the number of individuals cited in the interviews but who have not been given any coupon (number of found but still inactive vertices) and U_n to the total number of individuals having been interviewed (number of off-mode nodes). Because of the connectivity properties of the SBM graphs, we need to keep track of the types of the interviewees and the coupons distributed not only to one community but also in general to each of the m communities at every step. We then associate to the chain-referral the following stochastic vector process $X_n := (A_n, B_n, U_n), \quad n \in \mathbb{N}$:

$$X_n := \begin{pmatrix} A_n \\ B_n \\ U_n \end{pmatrix} = \begin{pmatrix} A_n^{(1)} & \cdots & A_n^{(m)} \\ B_n^{(1)} & \cdots & B_n^{(m)} \\ U_n^{(1)} & \cdots & U_n^{(m)} \end{pmatrix}, \quad n \in \mathbb{N},$$

where $A_n^{(\ell)}$ (resp. $B_n^{(\ell)}$ and $u_n^{(\ell)}$) corresponds to the number of active nodes (resp. of found but inactive nodes and of off-mode nodes) of type ℓ at step n. In all the paper, we will use the notation $(X_n^{1,(\ell)}, X_n^{2,(\ell)}, X_n^{3,(\ell)}) = (A_n^{(\ell)}, B_n^{(\ell)}, U_n^{(\ell)})$.

The main object of the paper is to establish an approximation result when the size N of the SBM graph tends to infinity. In this case, the chain-referral process correctly renormalized is:

$$X_t^N := \frac{1}{N} X_{\lfloor Nt \rfloor} = \left(\frac{A_{\lfloor Nt \rfloor}}{N}, \frac{B_{\lfloor Nt \rfloor}}{N}, \frac{U_{\lfloor Nt \rfloor}}{N} \right) \in [0,1]^{3\times m}, \quad t \in [0,1]. \tag{2.1}$$

In all the paper, we consider spaces $\mathbb{R}^d$ equipped with the L^1-norm defined for $x = (x^1, ..., x^d)$ as $\|x\| = \sum_{k=1}^d |x^k|$. For all N, the process $(X_t^N)_{t\in[0,1]}$ lives in the space of càdlàg processes $\mathcal{D}([0,1], [0,1]^{3\times m})$ equipped with Skorokhod topology (see [93, 55, 59]).

There exist to our knowledge a few works of studying CRS form a probabilistic point of view, for example Athreya and Röllin [5]. In their work, they obtained a result in a slightly different framework: they consider random walks on the limiting

graphon to construct a sequence of sub-graphs, which converges almost surely to the graphon underlying the network in the cut-metric. Whereas we take here to the limit both the graph and its exploring random walk simultaneously. The main result of this paper is that the sequence of processes $(X^N)_{N\geq 1}$ converges to a system of ordinary differential equations (ODEs). There has also been literature on random walks exploring graphs possibly with different mechanism (see [11, 35] for instance). Here we allow the exploring Markov process to branch. Also, our process bares similarities with epidemics spreading on graphs (see [8, 90, 30, 57]) but with the additional constraint of a maximum number of distributed coupons here.

The CRS is constructed by the similar principle of an epidemic spread and starts with a single individual. There are two main phases of evolution (see [8]): the initial phase is well approximated by a branching process (which we are neglecting here) and the second phase is when the stochastic process is approximated by an deterministic curve. In this paper, we focus on the second phase, but let us comment quickly on the first phase. In the sequel, we will assume that:

Assumption 2.1 For each $\ell, k \in \{1, \ldots, m\}$, denote $\mu_{\ell k} = \lambda_{\ell k}\pi_k$. We assume that the matrix $\mu = (\mu_{\ell k})_{\ell,k\in\{1,\ldots,m\}}$ is *irreducible* and the largest eigenvalue of μ is larger than 1.

Remark 2.1 Under the Assumption 0.2, from the proof of Theorem 3.2 of Barbour and Reinert [8], the early stages of the CRS now can be approximated by a multitype branching process with the offspring distributions determined by the matrix μ. Thanks to the Assumption 2.2 the multitype branching process associated with the offspring matrix μ is supercritical. The analogous results for the extinction probability and for the number of offspring at the n^{th} generation as in the single branching process have been proved in Chapter 5 of [82]: the mean matrix of the population size at time n is proportional to μ^n. And follow the claim (3.11) of Barbour and Reinert [8], we can deduce that if we start with a single individual, then after a finite steps, we can reach a positive fraction of explored individuals in the population with a positive probability.

Assumption 2.2 Set $a_0, b_0, u_0 \in [0,1]^m$, $a_0 = (a_0^{(1)}, \ldots, a_0^{(m)})$ such that $\sum_{i=1}^m a_0^{(i)} = \|a_0\| \in (0,1]$, and set $b_0, u_0 \in [0,1]^m$, with $b_0 = (0, \ldots, 0)$ and $u_0 = (0, \ldots, 0)$. We assume that the sequence $X_0^N = \frac{1}{N}X_0$ converges in probability to the vector (a_0, b_0, u_0), as $N \to +\infty$.

It means that the initial number of individuals with type i at the beginning of the survey is approximately $\lfloor a_0^{(i)} N \rfloor$. A possible way to initializing the process is to draw A_0 from a multinomial distribution $\mathcal{M}(\lfloor \|a_0\| N \rfloor; \pi_1, \ldots, \pi_m)$.

Theorem 2.1 Under the assumptions 2.2 and 2.2, we have: when N tends to infinity, the process $(X^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}([0,1],[0,1]^{3\times m})$ to a deterministic vectorial function $x = (x_\cdot^{(\ell)})_{1\leq \ell\leq m} = (a_\cdot^{(\ell)}, b_\cdot^{(\ell)}, u_\cdot^{(\ell)})_{1\leq \ell\leq m}$ in $\mathcal{C}([0,1],[0,1]^{3\times m})$, which is the unique solution of the system of differential equations

$$x_t = x_0 + \int_0^t f(x_s)ds, \tag{2.2}$$

where $f(x_s) := (f_{i\ell}(x_s))_{\substack{1\leq i\leq 3\\ 1\leq \ell\leq m}}$ has an explicit formula described as follows. Denote

$$t_0 := \inf\{t \in [0,1] : \|a_t\| := a_t^{(1)} + \ldots + a_t^{(m)} = 0\}. \tag{2.3}$$

For $s \in [0, t_0]$,

$$f_{1\ell}(x_s) = \sum_{k=1}^{m} \frac{a_s^{(k)}}{\|a_s\|} \frac{\lambda_s^{k,\ell}}{\Lambda_s^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right) - \frac{a_s^{(\ell)}}{\|a_s\|}; \tag{2.4}$$

$$f_{2\ell}(x_s) = \sum_{k=1}^{m} \frac{a_s^{(k)}}{\|a_s\|} \mu_s^{k,\ell} - \sum_{k=1}^{m} \frac{a_s^{(k)}}{\|a_s\|} \frac{\lambda_s^{k,\ell}}{\Lambda_s^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right); \tag{2.5}$$

$$f_{3\ell}(x_s) = \frac{a_s^{(\ell)}}{\|a_s\|}; \tag{2.6}$$

with

$$\lambda_s^{k,\ell} := \lambda_{k\ell} \left(\pi_\ell - a_s^{(\ell)} - u_s^{(\ell)} \right); \quad \Lambda_s^k := \sum_{\ell=1}^{m} \lambda_s^{k,\ell} \tag{2.7}$$

and $\mu_s^{k,\ell} := \lambda_{k\ell}(\pi_\ell - a_s^{(l\ell)} - b_s^{(\ell)} - u_s^{(\ell)})$.

For $s \in [t_0, 1], x_s = x_{t_0}$.

Remark 2.2 Notice that in this model, the time corresponds to the fraction of the population interviewed. The time t_0 is the first time at which $|a_t|$ reaches 0 and can be seen as the proportion of the population interviewed when there is no more coupon to keep the CRS going. Necessarily, $t_0 \leq 1$. We see that $\|a_t\| = 0$ only if $a_t^{(1)} = \ldots = a_t^{(m)} = 0$. It implies that $f(x_t) = 0, \forall t \in [t_0, 1]$. Then, the solution of the system of ODEs (2.2) becomes constant over the interval $[t_0, 1]$.

The rest of this paper is organized in the following manner. First, in Section 2, we give a precise description of the chain-referral process on a SBM random graph. This relies heavily on the structure of the random graph that we construct progressively

when the exploration process spreads on it. In Section 3, we prove the limit theorem. The proof uses limit theory of càdlàg semi-martingale vector processes equipped with Skorokhod topology (see [93]) and Poisson approximations (see [84]). Then in Section 4, we present simulation results of the stochastic process and the solution of the system of limiting ODEs. We conclude with some discussions on the impacts of changing parameters of the models on the evolution of the chain-referral process.

2.2 Definition of the chain-referral process

Let us describe the dynamics of $X = (X_n)_{n\in\mathbb{N}}$. Recall that $\|A_n\| := \sum_{l=1}^{m} A_n^{(\ell)}$ is the total number of individuals having coupons but who have not yet been interviewed. We start with A_0 seeds, whose types are chosen independently according to π. A_0 is an m-dimensional random vector with multinomial distribution $\mathcal{M}(\lfloor\|a_0\|N\rfloor; \pi_1, ..., \pi_m)$, *i.e.* $\mathbb{P}\big((A_0^{(1)}, ..., A_0^{(m)}) = (k_1, ..., k_m)\big) = \pi_1^{k_1}...\pi_m^{k_m}, k_i \in \mathbb{N}$ such that $\sum_{i=1}^{m} k_i = \lfloor\|a_0\|N\rfloor$ and Assumption 2.2 is satisfied. Also $B_0 = U_0 = (0, ..., 0)$ and we set $X_0 = (A_0, B_0, U_0)$.

We now define X_n given the state X_{n-1} previous to the n^{th}-interview and given the number $N_1, \ldots, N_m$ of nodes of each type. At step $n \geq 1$, after the n^{th}-interview, the type of the upcoming interviewee is chosen uniformly at random according to the number of active coupons of each type in the present time. To choose the type of the next interviewee, we define an m-dimensional vector $I_n := (I_n^{(1)}, ..., I_n^{(m)})$, which takes value 1 at coordinate ℓ and 0 elsewhere if the n^{th} interviewee belongs to block ℓ. This n^{th}-interviewee is chosen uniformly among the $\|A_{n-1}\|$ active coupons of m types *i.e.* I_n has multinomial distribution

$$I_n = (I_n^{(1)}, ..., I_n^{(m)}) \overset{(d)}{=} \mathcal{M}\left(1; \frac{A_{n-1}^{(1)}}{\|A_{n-1}\|}, ..., \frac{A_{n-1}^{(m)}}{\|A_{n-1}\|}\right). \tag{2.8}$$

If the chosen one belongs to block $[\ell]$, $A_n^{(\ell)}$ is reduced by 1 and a number of new coupons distributed are added up, depending on how many new contacts he/she has. In the meantime, the number of interviewees of type ℓ is increased by 1. *i.e.* $U_n^{(\ell)} = U_{n-1}^{(\ell)} + I_n^{(\ell)}$. Among the new contacts of the $n^{\text{th}}-$interviewee, define $H_n^{(\ell)}$ the number of new contacts of type ℓ, who have not been mentioned before; $K_n^{(\ell)}$ the number of new contacts of type ℓ whose identities are already known but who are still inactive. The $H_n^{(\ell)}$ new connections are chosen independently among $N_\ell - A_{n-1}^{(\ell)} - B_{n-1}^{(\ell)} - U_n^{(\ell)}$ individuals in the hidden population where probability of each successful connection is $\sum_{k=1}^{m} I_n^{(k)} p_{kl}$. Hence, conditioning on $(N_1, \ldots, N_m), X_{n-1}$, the random variable $H_n^{(\ell)}$ follows the binomial distribution:

$$H_n^{(\ell)} \overset{(d)}{=} \mathcal{B}in\left(N_\ell - A_{n-1}^{(\ell)} - B_{n-1}^{(\ell)} - U_n^{(\ell)}, \sum_{k=1}^{m} I_n^{(k)} p_{kl}\right). \tag{2.9}$$

And the $K_n^{(\ell)}$ individuals are chosen independently of $H_n^{(\ell)}$ from $B_{n-1}^{(\ell)}$ individuals and independently of the others with probability $\sum_{k=1}^{m} I_n^{(k)} p_{kl}$. In that way, condi-

tioning on $(N_1, \ldots, N_m), X_{n-1}, K_n^{(\ell)}$ also has the binomial distribution:

$$K_n^{(\ell)} \overset{(d)}{=} \mathcal{B}in\left(B_{n-1}^{(\ell)}, \sum_{k=1}^{m} I_n^{(k)} p_{kl}\right). \tag{2.10}$$

In total, there are $Z_n := H_n + K_n$ candidates, who can possibly receive coupons at step n. Notice that, conditioning on $(N_1, \ldots, N_m), X_{n-1}, (H_n^{(\ell)})_{\ell=1,\ldots,m}$ and $(K_n^{(\ell)})_{\ell=1,\ldots,m}$ are independent, henceforth,

$$Z_n^{(\ell)} \overset{(d)}{=} \mathcal{B}in\left(N_\ell - A_{n-1}^{(\ell)} - U_n^{(\ell)}, \sum_{k=1}^{m} I_n^{(k)} p_{k\ell}\right). \tag{2.11}$$

Let $C_n = (C_n^{(1)}, ..., C_n^{(m)})$ $(\ell = 1, ..., m)$ be the numbers of coupons that are distributed at step n. By the setting of the survey, the total coupons $|C_n|$ must be maximum c. If the number Z_n of candidates is less than or equal to c, we deliver exactly Z_n coupons. Otherwise, we choose new people to be enrolled in the study by an $m-$dimensional random variable $C_n'^{(\ell)} = (C_n'^{(1)}, ..., C_n'^{(m)})$ having the multivariate hypergeometric distribution with parameters $(m; c, (Z_n^{(1)}, ..., Z_n^{(m)}))$ and the support $\{(c_1, ..., c_m) \in \mathbb{N}^m : \forall \ell \leq m, c_\ell \leq Z_n^{(\ell)}, \sum_{\ell=1}^{m} c_i = c\}$, that is

$$\mathbb{P}\left((C_n'^{(1)}, ..., C_n'^{(m)}) = (c_1, ..., c_m)\right) = \frac{\prod_{l=1}^{m} \binom{Z_n^{(\ell)}}{c_l}}{\binom{\sum_{l=1}^{m} Z_n^{(\ell)}}{c}}.$$

In another words,

$$C_n^{(\ell)} := \begin{cases} Z_n^{(\ell)} & \text{if} \quad \sum_{l=1}^{m} Z_n^{(\ell)} \leq c \\ C_n'^{(\ell)} & \text{otherwise} \end{cases}. \tag{2.12}$$

Let define by

$$n_0 := \inf\{n \in \{1, ..., N\}, A_n = 0\} \tag{2.13}$$

the first step that $|A_n|$ reaches zero. The dynamics of X_n can be described by the following recursion:

$$\begin{cases} A_n &= A_{n-1} - I_n + C_n \\ B_n &= B_{n-1} + H_n - C_n \\ U_n &= \sum_{i=1}^{n} I_i \end{cases}, \quad \text{for} \quad n \in \{1, ..., n_0\} \tag{2.14}$$

$$\text{and} \qquad X_n = X_{n-1} \quad \text{when} \quad n > n_0.$$

The random network is progressively discovered when the referrals chain process explores it.

Proposition 2.1 Consider the discrete-time process $(X_n)_{1 \leq n \leq N}$ defined in

(2.14). For $n \in \mathbb{N}$, we denote by $\mathcal{F}_n := \sigma\big(\{X_i, i \leq n, (N_1, \dots, N_m)\}\big)$ the canonical filtration associated with $(X_n)_{1\leq n\leq N}$. Then the process $(X_n)_n$ is an inhomogeneous Markov chain with respect to the filtration $(\mathcal{F}_n)_n$.

Proof. The proposition is deduced from the recursion (2.14) of $(X_n)_{1\leq n\leq N}$ and the fact that the random variables C_n, I_n, H_n are defined conditionally on X_{n-1} and $(N_1, \dots, N_m)$. The fact that the Markov process is inhomogeneous comes from the setting of the CRS survey: there is no replacement in the recruitment procedure. For example, when $m = 1$, the definition of $H_n^{(\ell)}$ in (2.9) depends on time as $U_n^{(\ell)} = n$. ■

2.3 Asymptotic behavior of the chain-referral process

Let us now consider the renormalized chain-referral process given in (2.1) in the time interval $[0, t_0]$. The main theorem (Theorem 2.1) shows the convergence of the sequence $(X^N)_{N\geq 1}$ to a deterministic process. For this, we look for an expression of the equations (2.14) as a vector of semi-martingales. We start by writing the Markov chain $(X_n)_{1\leq n\leq N}$ as the sum of its increments in discrete time.

$$X_n = X_0 + \sum_{i=1}^{n}(X_i - X_{i-1}) = \begin{pmatrix} A_0 \\ B_0 \\ U_0 \end{pmatrix} + \sum_{i=1}^{n} \begin{pmatrix} C_i - I_i \\ H_i - C_i \\ I_i \end{pmatrix}.$$

Each element of the increment $X_{n+1} - X_n$ are binomial variables conditioned on all the events having been occurring until step n. When we fix n and let N tend to infinity, the conditional binomial random variables can be approximated by some Poisson random variables. The normalization X_t^N of X_n becomes:

$$X_t^N = \frac{1}{N}\begin{pmatrix} A_0 \\ B_0 \\ U_0 \end{pmatrix} + \frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor} \begin{pmatrix} C_i - I_i \\ H_i - C_i \\ I_i \end{pmatrix}.$$

The Doob decomposition of the renormalized processes $(X_t^N)_{t\in[0,t_0]}$ given in Section 2.3.1 consists of a finite variation process and an $\mathbb{L}^2$-martingale. We use Aldous criteria (conditionally on the past see e.g. [93, 98]) to show the tightness of the distributions of these processes in Section 2.3.2. Once the tightness is established, we identify the limiting values of this tight sequence and finally we prove that the limiting values of all converging subsequences are the same, hence it is the limit of processes $(X^N)_{N\geq 1}$. This proves Theorem 2.1.

Denote by $(\mathcal{F}_t^N)_{t\in[0,1]} := (\mathcal{F}_{\lfloor Nt\rfloor})_{t\in[0,1]}$ the canonical filtration associated to $(X_t^N)_{t\in[0,1]}$.

2.3.1 Doob's decomposition

Lemma 2.1 The process $(X_t^N)_{t\in[0,1]}$ admits the Doob's decomposition: $X_t^N = X_0^N + \Delta_t^N + M_t^N$, $X_0^N = \frac{1}{N}X_0$. $(\Delta_t^N)_{t\in[0,1]}$ is an $\mathcal{F}_t^N-$predictable process defined by

$$\Delta_t^N = \begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \\ \Delta_t^{N,3} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} \begin{pmatrix} \mathbb{E}[C_n - I_n|\mathcal{F}_{n-1}] \\ \mathbb{E}[H_n - C_n|\mathcal{F}_{n-1}] \\ \mathbb{E}[I_n|\mathcal{F}_{n-1}] \end{pmatrix}; \tag{2.15}$$

$(M_t^N)_{t\in[0,1]}$ is an $\mathcal{F}_t^N-$ square integrable centered martingale with quadratic variation process $(\langle M_\cdot^N\rangle_t)_{t\in[0,1]}$ given by: for every $(\ell,k)\in\{1,...,m\}^2$,

$$\begin{aligned}\langle M_\cdot^{(\ell),N}, M_\cdot^{(k),N}\rangle_t =& \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\left(X_n^{(\ell)} - \mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\right)\right. \\ &\left.\left(X_n^{(k)} - \mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\right)^T \middle| \mathcal{F}_{n-1}\right], \quad t\in[0,1] \end{aligned} \tag{2.16}$$

where X is a column vector and X^T is its transpose.

Proof. In order to obtain the Doob's decomposition, we write for $t\in[0,1]$,

$$\begin{aligned} X_t^N &= \frac{X_0}{N} + \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}(X_n - X_{n-1}) \\ &= X_0^N + \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] + \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}(X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]) \\ &= X_0^N + \Delta_t^N + M_t^N. \end{aligned}$$

It is clear that the conditional expectations above are all well-defined since the components of X_n and X_{n-1} are all bounded by N, that Δ_t^N is $\mathcal{F}_t^N-$predictable and that $(M_t^N)_{t\in[0,1]}$ is an $\mathcal{F}_t^N-$martingale. We first check that $(\Delta^N)_{N\geq 1}$ is a sequence of finite variation processes and then we can conclude that $X_t^N = X_0^N + \Delta_t^N + M_t^N$ is the Doob's decomposition.

Denote by $\lambda := \max_{\ell,k\in\{1,...,m\}}\lambda_{k\ell}$. Notice that

$$\|\mathbb{E}[A_n - A_{n-1}|\mathcal{F}_{n-1}\| = \|\mathbb{E}[C_n - I_n|\mathcal{F}_{n-1}\| \leq c, \tag{2.17}$$

$$\|\mathbb{E}[B_n - B_{n-1}|\mathcal{F}_{n-1}\| = \|\mathbb{E}[H_n - C_n|\mathcal{F}_{n-1}\| \leq m(\max_{\ell,k\in\{1,...,m\}}\lambda_{k\ell}) + c = m\lambda + c, \tag{2.18}$$

$$\|\mathbb{E}[U_n - U_{n-1}|\mathcal{F}_{n-1}\| \leq 1, \tag{2.19}$$

then $\|\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]\| \leq 2c + m\lambda + 1$. So the total variation of $(\Delta_t^N)_{t\in[0,1]}$ is

$$\begin{aligned} V^N(\Delta_t^N) &= \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\|\Delta_{nt/N}^N - \Delta_{(n-1)t/N}^N\| = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\|\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]\| \\ &\leq (2c + m\lambda + 1)t, \end{aligned}$$

which is finite. It follows that $(\Delta_t^N)_{t\in[0,1]}$ is an $\mathcal{F}_t^N-$ predictable with finite variations.

The quadratic variation of $(M_t^N)_{t\in[0,1]}$ is computed as follow. For every $k,\ell \in [\![1,m]\!]$,

$$\begin{aligned}
M_t^{(\ell),N}\left(M_t^{(k),N}\right)^T &= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\Big\{\Big(X_n^{(\ell)}-X_{n-1}^{(\ell)}-\mathbb{E}[X_n^{(\ell)}-X_{n-1}^{(\ell)}|\mathcal{F}_{n-1}]\Big)\\
&\qquad\times\Big(X_n^{(k)}-X_{n-1}^{(k)}-\mathbb{E}[X_n^{(k)}-X_{n-1}^{(k)}|\mathcal{F}_{n-1}]\Big)^T\Big\}\\
&\quad+\frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{\substack{n'=1\\n'\neq n}}^{\lfloor Nt\rfloor}\Big\{\Big(X_n^{(\ell)}-X_{n-1}^{(\ell)}-\mathbb{E}[X_n^{(\ell)}-X_{n-1}^{(\ell)}|\mathcal{F}_{n-1}]\Big)\\
&\qquad\times\Big(X_{n'}^{(k)}-X_{n'-1}^{(k)}-\mathbb{E}[X_{n'}^{(k)}-X_{n'-1}^{(k)}|\mathcal{F}_{n'-1}]\Big)^T\Big\}\\
&=: L_t^N+L_t'^N.
\end{aligned}$$

The term $L_t'^N$ is an $\mathcal{F}_t^N-$martingale since whenever $n'<n$, $\Big(X_{n'}^{(k)}-X_{n'-1}^{(k)}-\mathbb{E}[X_{n'}^{(k)}-X_{n'-1}^{(k)}|\mathcal{F}_{n'-1}]\Big)$ is $\mathcal{F}_{n-1}-$measurable. To see that the quadratic variation of M_t^N has the form (2.16), we write the term L_t^N as follows:

$$\begin{aligned}
L_t^N &:= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\Big(X_n^{(\ell)}-\mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\Big)\Big(X_n^{(k)}-\mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T\Big|\mathcal{F}_{n-1}\right]\\
&\quad+\frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\Big(X_n^{(\ell)}-\mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\Big)\Big(X_n^{(k)}-\mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T\\
&\quad-\frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\Big(X_n^{(\ell)}-\mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\Big)\Big(X_n^{(k)}-\mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T\Big|\mathcal{F}_{n-1}\right]\\
&= \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\Big(X_n^{(\ell)}-\mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\Big)\Big(X_n^{(k)}-\mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T\Big|\mathcal{F}_{n-1}\right]+L_t''^N\\
&= \langle M^N\rangle_t+L_t''^N.
\end{aligned}$$

As a result,

$$M_t^{(\ell),N}\left(M_t^{(k),N}\right)^T=\langle M^N\rangle_t+L_t'^N+L_t''^N. \tag{2.20}$$

Because both $L_t'^N$ and $L_t''^N$ are $\mathcal{F}_t^N-$martingale, $L_t'^N+L_t''^N$ is an $\mathcal{F}_t^N-$martingale as well. The term $(\langle M^N\rangle_t)_t$ is $\mathcal{F}_t^N-$adapted with the variation

$$\begin{aligned}
V^N(\langle M_\cdot^N\rangle_t) = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k,\ell=1}^{m}\Bigg\|&\mathbb{E}\left[\Big(X_n^{(\ell)}-\mathbb{E}[X_n^{(\ell)}|\mathcal{F}_{n-1}]\Big)\right.\\
&\left.\times\Big(X_n^{(k)}-\mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T\Big|\mathcal{F}_{n-1}\right]\Bigg\|.
\end{aligned}\tag{2.21}$$

The integrand in the right hand side is the conditional covariance between $X_n^{(\ell)}$ and $X_n^{(k)}$ conditionally to $\mathcal{F}_{n-1}$. Because $X_n^{(\ell)}$ and $X_n^{(k)}$ are vectors, this covariance is a matrix of size 3×3 and for $1\leq i,j\leq 3$, the term (i,j) of this matrix is:

$$\mathbb{E}\left[\left(X_n^{i,(\ell)}-\mathbb{E}[X_n^{i,(\ell)}|\mathcal{F}_{n-1}]\right)\left(X_n^{j,(k)}-\mathbb{E}[X_n^{j,(k)}|\mathcal{F}_{n-1}]\right)\Big|\mathcal{F}_{n-1}\right]$$
$$\leq \left(\mathrm{Var}(X_n^{i,(\ell)}-X_{n-1}^{i,(\ell)}|\mathcal{F}_{n-1})\right)^{1/2}\left(\mathrm{Var}(X_n^{j,(k)}-X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2},$$

by the Cauchy-Schwarz inequality. Thus:

$$V^N(\langle M_\cdot^N\rangle_t)\leq \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k,\ell=1}^{m}\Bigg|\sum_{i,j=1}^{3}\left(\mathrm{Var}(X_n^{i,(\ell)}-X_{n-1}^{i,(\ell)}|\mathcal{F}_{n-1})\right)^{1/2}$$
$$\times\left(\mathrm{Var}(X_n^{j,(k)}-X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2}\Bigg|,$$

where $(X_n^{1,(\ell)},X_n^{2,(\ell)},X_n^{3,(\ell)})=(A_n^{(\ell)},B_n^{(\ell)},U_n^{(\ell)})$. By Cauchy-Schwarz's inequality, we have

$$\sum_{i,j=1}^{3}\left(\mathrm{Var}(X_n^{i,(\ell)}-X_{n-1}^{i,(\ell)}|\mathcal{F}_{n-1})\right)^{1/2}\left(\mathrm{Var}(X_n^{j,(k)}-X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2}$$
$$=\left(\sum_{i=1}^{3}\left(\mathrm{Var}(X_n^{i,(\ell)}-X_{n-1}^{i,(\ell)}|\mathcal{F}_{n-1})\right)^{1/2}\right)\left(\sum_{j=1}^{3}\left(\mathrm{Var}(X_n^{j,(k)}-X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2}\right)$$
$$\leq\frac{3}{2}\sum_{i=1}^{3}\left(\mathrm{Var}(X_n^{i,(\ell)}-X_{n-1}^{i,(\ell)}|\mathcal{F}_{n-1})+\mathrm{Var}(X_n^{i,(k)}-X_{n-1}^{i,(k)}|\mathcal{F}_{n-1})\right). \tag{2.22}$$

From (2.17)-(2.19) and by Cauchy-Schwarz's inequality, we obtain the following inequalities

$$\mathrm{Var}(C_n^{(\ell)}-I_n^{(\ell)}|\mathcal{F}_{n-1})\leq c^2,\quad \mathrm{Var}(I_n^{(\ell)}|\mathcal{F}_{n-1})\leq 1,$$
$$\text{and}\quad \mathrm{Var}(H_n^{(\ell)}-C_n^{(\ell)}|\mathcal{F}_{n-1})\leq 2(\max_{\ell,k\in\{1,\dots,m\}}\lambda_{\ell k}^2+c^2), \tag{2.23}$$

As a consequence,

$$V^N\left(\langle M_\cdot^N\rangle_t\right)\leq\frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}3m^2(c^2+2(\max_{\ell,k\in\{1,\dots,m\}}\lambda_{lk}^2+c^2)+1)$$
$$\leq\frac{1}{N}3m^2(3c^2+2\lambda^2+1)<\infty.$$

Thus, the proof of the Lemma is completed. ■

2.3.2 Tightness of the renormalized process

Lemma 2.2 The sequence of processes $(X^N)_{N\geq 1}$ is tight in the Skorokhod

space $\mathcal{D}([0,1],[0,1]^{3\times m})$.

Proof. To prove the tightness of $(X^N)_{N\geq 1}$, we use the criteria of tightness for semi-martingales in [98, Theorem 2.3.2 (Rebolledo)]: first, we verify the marginal tightness of each sequence $(X_t^N)_{N\geq 1}$ for each $t\in[0,1]$, then we show the tightness for each process in the Doob's decomposition of X^N, the finite variation process $(\Delta^N)_{N\geq 1}$ and the quadratic variation of the martingale $(M^N)_{N\geq 1}$. For any $t\in[0,1]$, the tightness of marginal sequence $(X_t^N)_N$ is easily deduced from the compactness of a sequence of random variables taking values in a compact set $[0,1]^{3\times m}$. Since the sequence of martigales $(M^N)_{N\geq 1}$ is proved to be convergent (to zero) in $\mathbb{L}^2$ as $N\to\infty$ (which is done by Proposition 2.2), we have the tightness of $(M^N)_{N\geq 1}$. Thus, it is sufficient to check the tightness condition for the modulus of continuity of $(\Delta^N)_{N\geq 1}$ (see, *e.g.* , [85, Theorem 13.2, p.139]).

For all $0<\delta<1$ and for every $s,t\in[0,1]$ such that $|t-s|<\delta$, we have that

$$\begin{aligned}\|\Delta_t^N-\Delta_s^N\| &= \left\|\frac{1}{N}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\mathbb{E}[X_n-X_{n-1}|\mathcal{F}_{n-1}]\right\| \\ &\leq \frac{1}{N}\sum_{n=\lfloor Ns\rfloor+1}^{\lfloor Nt\rfloor}\|\mathbb{E}[X_n-X_{n-1}|\mathcal{F}_{n-1}]\|.\end{aligned}$$

By (2.17)-(2.19), we get

$$\|\Delta_t^N-\Delta_s^N\|\leq\frac{\lfloor Nt\rfloor-\lfloor Ns\rfloor}{N}(c+m\lambda+c+1)\leq(2c+m\lambda+1)\left(\delta+\frac{1}{N}\right).$$

Thus, for each $\varepsilon>0$, choose $\delta_0\leq\frac{\varepsilon}{2(2c+m\lambda+1)}$, we have that

$$\mathbb{P}\left(\sup_{\substack{|t-s|<\delta\\0\leq s<t\leq 1}}\|\Delta_t^N-\Delta_s^N\|>\varepsilon\right)=0,\quad\forall\delta\leq\delta_0,\forall N>\frac{1}{\delta_0},$$

which allows us to conclude that the sequence $(\Delta_\cdot^N)_N$ is tight and finishes the proof of the lemma. ■

To complete the proof of Lemma 2.2, we now prove that:

Proposition 2.2 The sequence of martingale $(M^N)_{N\geq 1}$ converges to 0 in $\mathbb{L}^2$ as N goes to infinity.

Proof. Consider the quadratic variation of $(M^N)_{N\geq 1}$: According to the fomula (2.16), we apply the Cauchy-Schwarz's inequality and then use the inequality (2.22) to obtain that for every $t\in[0,1]$,

$$\|\langle M^{(\ell),N},M^{(k),N}\rangle_t\|$$

$$
\begin{aligned}
&= \left\| \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \mathbb{E}\left[\left(X_n^{(\ell)} - \mathbb{E}[X_n^{(\ell)} | \mathcal{F}_{n-1}] \right) \left(X_n^{(k)} - \mathbb{E}[X_n^{(k)} | \mathcal{F}_{n-1}] \right)^T \middle| \mathcal{F}_{n-1} \right] \right\| \\
&\leq \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \left| \sum_{i,j=1}^{3} \left(\mathrm{Var}(X_n^{i,(\ell)} - X_{n-1}^{i,(\ell)} | \mathcal{F}_{n-1}) \right)^{1/2} \left(\mathrm{Var}(X_n^{j,(k)} - X_{n-1}^{j,(k)} | \mathcal{F}_{n-1}) \right)^{1/2} \right| \\
&\leq \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \frac{3}{2} \sum_{i=1}^{3} \left(\mathrm{Var}(X_n^{i,(\ell)} - X_{n-1}^{i,(\ell)} | \mathcal{F}_{n-1}) + \mathrm{Var}(X_n^{i,(k)} - X_{n-1}^{i,(k)} | \mathcal{F}_{n-1}) \right),
\end{aligned}
$$

where $(X_n^{1,(\ell)}, X_n^{2,(\ell)}, X_n^{3,(\ell)}) = (A_n^{(\ell)}, B_n^{(\ell)}, U_n^{(\ell)})$. From (2.17)-(2.19) and (2.23), we deduce that

$$
\begin{aligned}
\| \langle M_\cdot^N \rangle_t \| &\leq \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \frac{3m^2}{2} \left(c^2 + 2(\max_{\ell,k \in \{1,\dots,m\}} \lambda_{\ell k}^2 + c^2) + 1 \right) \\
&\leq \frac{1}{N} \frac{3m^2}{2} (3c^2 + 2\lambda^2 + 1) t. \qquad (2.24)
\end{aligned}
$$

Applying the Doob's inequality for martingale, for every $t \in [0,1]$, we have

$$
\mathbb{E}\left[\max_{0 \leq s \leq t} \| M_s^N \|^2 \right] \leq 4\mathbb{E}\left[\| \langle M_\cdot^N \rangle_t \| \right] \leq \frac{1}{N} 6m^2 (3c^2 + 2\lambda^2 + 1) \to 0 \quad \text{as} \quad N \to \infty.
$$

This concludes the proof of Prop. 2.2 and hence of Lemma 2.2. ■

2.3.3 Identify the limiting value

Since the sequence $(X^N)_{N \geq 1}$ is tight, for any limiting value $x = (a, b, u)$ of the sequence $(X^N)_{N \geq 1}$, there exists an increasing sequence $(\varphi_N)_N$ in $\mathbb{N}$ such that $(X^{\varphi_N})_{N \geq 1}$ converges in distribution to x in $\mathcal{D}([0,1], [0,1]^{3 \times m})$. Because the sizes of the jumps converge to zero with N, the limit is in fact in $\mathcal{C}([0,1], [0,1]^{3 \times m})$. We want to identify that limit. In order to simplify the notations, we also write the subsequence $(X^{\varphi_N})_{N \geq 1}$ as $(X^N)_{N \geq 1} = (A^N, B^N, U^N)_{N \geq 1}$.

We consider separately the martingale and finite variation parts. Proposition 2.2 implies that the sequence martingale $(M^N)_{N \geq 1}$ converges to 0 in distribution and hence $(M^N)_{N \geq 1}$ converges to zero in probability. It remains to find the limit of the finite variation process $(\Delta^N)_{N \geq 1}$ given in Equation (2.15) and prove that the limit found is the same (which is done later in the proof for the uniqueness of the system of the ODEs (2.1)) for every convergent subsequence extracted from the tight sequence $(X^N)_{N \geq 1}$.

Proposition 2.3 When N goes to infinity, we have the following convergences in distribution in $\mathcal{D}([0,1], [0,1]^{3 \times m})$:

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[C_n^{(\ell)}|\mathcal{F}_{n-1}] \xrightarrow{(d)} \int_0^t \left\{\sum_{k=1}^{m}\frac{a_s^{(k)}}{\|a_s\|}\frac{\lambda_s^{k,\ell}}{\Lambda_s^k}\left(c-\sum_{h=0}^{c}(c-h)\frac{(\Lambda_s^k)^h}{h!}e^{-\Lambda_s^k}\right)\right\} ds, \tag{2.25}$$

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[H_n^{(\ell)}|\mathcal{F}_{n-1}] \xrightarrow{(d)} \int_0^t \sum_{k=1}^{m}\frac{a_s^{(k)}}{\|a_s\|}\mu_s^{k,\ell}ds, \tag{2.26}$$

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[I_n^{(\ell)}|\mathcal{F}_{n-1}] = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left(\frac{A_{n-1}^{(\ell)}}{N}\right)\Big/\left(\frac{\|A_{n-1}\|}{N}\right) \xrightarrow{(d)} \int_0^t \frac{a_s^{(\ell)}}{\|a_s\|}ds, \tag{2.27}$$

where $\lambda_s^{k,\ell}, \Lambda_s^k, \mu_s^{k,\ell}$ are defined as in Theorem 2.1. This provides the convergence of $(\Delta^N)_{N\geq 1}$ to a solution x of (2.2).

Since the limits are deterministic, the convergences hold in probability. Moreover the uniqueness of the solution of (2.2) will be proved later, which will imply the convergence of the whole sequence $(X^N)_{N\geq 1}$ to this solution.

Proof. Recall that since the sequence $(X^N)_{N\geq 1}$ is tight, we have extracted a converging subsequence also denoted by $(X^N)_{N\geq 1}$ of which we study the limit.

The proof of the Proposition 2.3 is separated into three steps.

Step 1: We consider the most complicated term $\mathbb{E}[C_n|\mathbb{F}_{n-1}]$. We prove that: for each $\ell \in \{0, ..., m\}$,

$$\left|\mathbb{E}[C_n^{(\ell)}|\mathcal{F}_{n-1}] - \frac{\lambda_n^{(\ell)}}{\Lambda_n}\left(c-\sum_{h=0}^{c}(c-h)\frac{(\Lambda_n)^h}{h!}e^{-\Lambda_n}\right)\right| \leq \frac{m(c+1)\lambda}{N}, \tag{2.28}$$

where

$$\lambda_n^{(\ell)} := \left(\sum_{k=1}^{m} I_n^{(k)}\lambda_{k\ell}\right)\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N}\right) \qquad \text{and} \qquad \Lambda_n := \sum_{j=1}^{m}\lambda_n^{(j)}. \tag{2.29}$$

Notice that $\Lambda_n = 0$ only if for each $l \in \{1, \ldots, m\}$, $\lambda_n^{(\ell)} = 0$. It happens when $A_{n-1}^{(\ell)} + U_n^{(\ell)} = N_\ell$, meaning that all the nodes of type ℓ have been discovered. In this case, $C_n^{(\ell)} = 0$ and (2.28) is satisfied.
Let us write

$$\mathbb{E}[C_n^{(\ell)}|\mathcal{F}_{n-1}] = \mathbb{E}\left[Z_n^{(\ell)}\mathbf{1}_{\sum_{j=1}^m Z_n^{(j)}\leq c}\Big|\mathcal{F}_{n-1}\right] + \mathbb{E}\left[\frac{cZ_n^{(\ell)}}{\sum_{j=1}^m Z_n^{(j)}}\mathbf{1}_{\sum_{j=1}^m Z_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right]. \tag{2.30}$$

For every $\ell = 1, ..., m$ and every fixed n, when all the parameters are positive, we have that $(N_\ell - A_{n-1}^{(\ell)} - U_n^{(\ell)}) \xrightarrow[\text{a.s.}]{N\to\infty} +\infty$. Then we work conditionally on $N_\ell, A_{n-1}^{(\ell)}, U_n^{(\ell)}$ and $I_n^{(\ell)}$ and use the Poisson approximation (*e.g.* see Equation (1.23) and Theorem 2.A, 2.B by Barbour, Holst and Janson in [84]) for the approximation: the binomial random variable $Z_n^{(\ell)}$ may be approximated by a Poisson random variable $\tilde{Z}_n^{(\ell)} \overset{(d)}{=} \mathcal{P}\big((\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell})(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N})\big)$ such that

$$\begin{aligned} d_{\text{TV}}(Z_n^{(\ell)}, \tilde{Z}_n^{(\ell)}) &\leq \frac{2}{(N_\ell - A_n^{(\ell)} - U_n^{(\ell)})\left(\frac{\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}}{N}\right)} \sum_{i=1}^{N_\ell - A_n^{(\ell)} - U_n^{(\ell)}} \left(\frac{\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}}{N}\right)^2 \\ &\leq \frac{2\max_{k,\ell}\lambda_{k\ell}}{N} = \frac{2\lambda}{N}. \end{aligned}$$

As a consequence, the first term in the right hand side of (2.30) can be approximated as

$$\left|\mathbb{E}\left[Z_n^{(\ell)}\mathbf{1}_{\sum_{j=1}^m Z_n^{(j)}\leq c}\Big|\mathcal{F}_{n-1}\right] - \mathbb{E}\left[\tilde{Z}_n^{(\ell)}\mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}\leq c}\Big|\mathcal{F}_{n-1}\right]\right| \leq \frac{2mc\lambda}{N}, \tag{2.31}$$

and

$$\left|\mathbb{E}\left[\frac{Z_n^{(\ell)}}{\sum_{j=1}^m Z_n^{(j)}}\mathbf{1}_{\sum_{j=1}^m Z_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right] - \mathbb{E}\left[\frac{\tilde{Z}_n^{(\ell)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}}\mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right]\right| \leq \frac{2m\lambda}{N}. \tag{2.32}$$

It follows that we need to deal with the Poisson random variables $\tilde{Z}_n^{(\ell)} (\ell \in \{1, ..., m\})$. Because of the result that the sum of two independent Poisson random variables is a Poisson random variable whose parameter is the sum of the two parameters, we have that $\sum_{j\neq\ell}\tilde{Z}_n^{(j)} =: \hat{Z}_n^{(\ell)}$ has a Poisson distribution with parameter $\hat{\lambda}_n^{(\ell)} := \sum_{j\neq\ell}\lambda_n^{(j)}$. And hence,

$$\begin{aligned} \mathbb{E}\left[\tilde{Z}_n^{(\ell)}\mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}\leq c}\Big|\mathcal{F}_{n-1}\right] &= \sum_{h=1}^{c}\sum_{h_1=1}^{h} h_1 \frac{(\lambda_n^{(\ell)})^{h_1}(\hat{\lambda}_n^{(\ell)})^{h-h_1}}{h_1!(h-h_1)!}e^{-\Lambda_n} \\ &= \lambda_n^{(\ell)}\sum_{h=1}^{c}\frac{(\Lambda_n)^{h-1}}{(h-1)!}e^{-\Lambda_n} = \lambda_n^{(\ell)}\sum_{h=0}^{c}\frac{h}{\Lambda_n}\frac{(\Lambda_n)^h}{h!}e^{-\Lambda_n} \end{aligned}$$

and

$$\mathbb{E}\left[\frac{\tilde{Z}_n^{(\ell)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}}\mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right] = \sum_{h=c+1}^{\infty}\sum_{k=0}^{h}\frac{k}{h}\frac{(\lambda_n^{(\ell)})^k}{k!}\frac{(\hat{\lambda}_n^{(\ell)})^{h-k}}{(h-k)!}e^{-\lambda_n^{(\ell)}}e^{-\hat{\lambda}_n^{(\ell)}}$$

$$
\begin{aligned}
&= \lambda_n^{(\ell)} \sum_{h=c+1}^{\infty} \sum_{k=0}^{h-1} \frac{1}{h} \frac{(\lambda_n^{(\ell)})^k}{k!} \frac{(\hat{\lambda}_n^{(\ell)})^{h-1-k}}{(h-1-k)!} e^{-\lambda_n^{(\ell)}} e^{-\hat{\lambda}_n^{(\ell)}} \\
&= \lambda_n^{(\ell)} \sum_{h=c+1}^{\infty} \frac{1}{h} \frac{(\lambda_n^{(\ell)} + \hat{\lambda}_n^{(\ell)})^{h-1}}{(h-1)!} e^{-(\lambda_n^{(\ell)} + \hat{\lambda}_n^{(\ell)})} \\
&= \frac{\lambda_n^{(\ell)}}{\Lambda_n} \sum_{h=c+1}^{\infty} \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \\
&= \frac{\lambda_n^{(\ell)}}{\Lambda_n} (1 - \sum_{h=0}^{c} \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n}). \qquad (2.33)
\end{aligned}
$$

Using (2.30), we obtain:

$$
\begin{aligned}
\mathbb{E}[C_n^{(\ell)} | \mathcal{F}_{n-1}] &= \mathbb{E}\left[\tilde{Z}_n^{(\ell)} \mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)} \leq c} + \frac{\tilde{Z}_n^{(\ell)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}} \mathbf{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)} > c} \middle| \mathcal{F}_{n-1} \right] \\
&= \frac{\lambda_n^{(\ell)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \right),
\end{aligned}
$$

which finishes step 1.

Step 2: We decompose the second term in the left hand side of (2.28) as follow

$$
\frac{\lambda_n^{(\ell)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \right) = \alpha_n^{(\ell)} + \xi_n^{(\ell)}, \quad \ell = 1, ..., m. \qquad (2.34)
$$

where

$$
\begin{aligned}
\alpha_n^{(\ell)} &:= \mathbb{E}\left[\frac{\lambda_n^{(\ell)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \right) \middle| \mathcal{F}_{n-1} \right] \\
\xi_n^{(\ell)} &:= \frac{\lambda_n^{(\ell)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \right) \\
&\qquad - \mathbb{E}\left[\frac{\lambda_n^{(\ell)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} e^{-\Lambda_n} \right) \middle| \mathcal{F}_{n-1} \right].
\end{aligned}
$$

By writing

$$
\alpha_n^{(\ell)} = \sum_{k=1}^{m} \mathbb{P}(I_n^{(k)} = 1) \frac{\lambda_n^{k,\ell}}{\Lambda_n^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n^k)^h}{h!} e^{-\Lambda_n^k} \right),
$$

where

$$\lambda_n^{k,\ell} := \lambda_{k\ell}\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{U_{n-1}^{(\ell)}}{N} - \frac{\mathbf{1}_{\{k=\ell\}}}{N}\right) \text{ and } \Lambda_n^k := \sum_{j=1}^m \lambda_n^{k,j} \quad (\ell = 1, ..., m), \tag{2.35}$$

we obtain that for every $t \in [0,1]$,

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} \alpha_n^{(\ell)} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left\{\sum_{k=1}^m \frac{A_{n-1}^{(k)}}{|A_{n-1}|}\frac{\lambda_n^{k,\ell}}{\Lambda_n^k}\left(c - \sum_{h=0}^c (c-h)\frac{(\Lambda_n^k)^h}{h!}e^{-\Lambda_n^k}\right)\right\}. \tag{2.36}$$

It is obvious that $\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\xi_n$ is an $\mathcal{F}_t^N-$martigale with the quadratic variation,

$$\langle \frac{1}{N}\sum_{n=1}^{\lfloor N\cdot\rfloor}\xi_n\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}\left[\xi_n^2|\mathcal{F}_{n-1}\right] \le \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor} m(c+1)^2 \le \frac{m(c+1)^2}{N}.$$

By the Doob's inequality, we have

$$\mathbb{E}\left[\max_{0\le s\le t}\|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\xi_n\|^2\right] \le 4\mathbb{E}\left[\|\langle\frac{1}{N}\sum_{n=1}^{\lfloor N\cdot\rfloor}\xi_n\rangle_t\|\right] \le \frac{4m(c+1)^2}{N} \overset{N\to\infty}{\longrightarrow} 0,$$

which deduces that as N tends to infinity, we have that

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\xi_n \overset{(\mathbb{L}^2)}{\to} 0 \tag{2.37}$$

uniformly in $t \in [0,1]$. Together with the points given in (2.28), (2.34) and (2.37), take the limit as $N \to \infty$ in the right hand side of (2.36), we obtain the right hand side of (2.25).

Step 3: We use similar arguments as in step 2 to obtain the limit in right hand side of (2.26). Denote by

$$\mu_n^{(\ell)} := \left(\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}\right)\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N}\right).$$

Recall from (2.9) that conditioning on $\mathcal{F}_{n-1}$,

$$H_n^{(\ell)} \overset{(d)}{=} \mathcal{B}in\left(N_\ell - A_{n-1}^{(\ell)} - B_{n-1}^{(\ell)} - U_n^{(\ell)}, \frac{\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}}{N}\right),$$

then

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[H_n^{(\ell)}|\mathcal{F}_{n-1}] = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mu_n^{(\ell)}. \tag{2.38}$$

We write

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mu_n^{(\ell)} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}(\beta_n^{(\ell)} + \zeta_n^{(\ell)}) \tag{2.39}$$

where

$$\beta_n^{(\ell)} := \mathbb{E}\left[\left(\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}\right)\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N}\right)\Big|\mathcal{F}_{n-1}\right];$$

$$\zeta_n^{(\ell)} := \left(\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}\right)\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N}\right) - \mathbb{E}\left[\left(\sum_{k=1}^m I_n^{(k)}\lambda_{k\ell}\right)\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_n^{(\ell)}}{N}\right)\Big|\mathcal{F}_{n-1}\right].$$

Using a similar argument as in step 2, we have

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\beta_n^{(\ell)} &= \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m \mathbb{P}(I_n^{(k)}=1)\lambda_{k\ell}\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_{n-1}^{(\ell)}}{N} - \frac{\mathbf{1}_{\{k\neq\ell\}}}{N}\right)\\ &= \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m \frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,\ell} - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m \frac{A_{n-1}^{N,k}}{\|A_{n-1}^N\|}\lambda_{k\ell}\frac{\mathbf{1}_{\{k\neq\ell\}}}{N},\end{aligned}$$

where $\mu_n^{k,\ell} := \lambda_{k\ell}\left(\frac{N_\ell}{N} - \frac{A_{n-1}^{(\ell)}}{N} - \frac{B_{n-1}^{(\ell)}}{N} - \frac{U_{n-1}^{(\ell)}}{N}\right)$. Then,

$$\left|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left(\beta_n^{(\ell)} - \sum_{k=1}^m \frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,\ell}\right)\right| \le \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m \frac{A_{n-1}^{N,k}}{\|A_{n-1}^N\|}\lambda_{k\ell}\frac{\mathbf{1}_{\{k\neq\ell\}}}{N} \le \frac{\lambda}{N}. \tag{2.40}$$

Take the limit as $N\to+\infty$, we have that

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m \frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,\ell} = \int_0^t \sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\mu_s^{k,\ell}ds.$$

Further, the $\mathcal{F}_t^N-$martingale $\frac{1}{N}\sum_{n=1}^{\lfloor N\cdot\rfloor}\zeta_n^{(\ell)}$ converges in $\mathbb{L}^2$ to 0 uniformly in $t\in[0,1]$. Thus, (2.26) is proved.

For the proof of (2.27), by the definition of I_n as in (2.8), we have

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[I_n^{(\ell)}|\mathcal{F}_{n-1}] = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\frac{A_{n-1}^{(\ell)}}{\|A_{n-1}\|} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\frac{A_{n-1}^{(\ell)}/N}{\|A_{n-1}\|/N}.$$

Take the limit as $N\to+\infty$, we obtain the limit in the right hand side of (2.27).

The preceding steps allow to conclude the proof of Proposition 2.3. ■

2.3.4 The uniqueness

It remains to prove that the limiting value $x = (a, b, u)$ we have found is unique solution of the system of he ODEs (2.2). If it is the case, then the process $(X^N)_{N\geq 1}$ admits a unique limiting value and thus converges to x.
Assume that there exist two solutions x^1 and x^2 to ODEs (2.2) on the interval $[0, t_0']$, where

$$t_0' = \inf\{t \in [0,1] : a_{t_0'}^1 = 0 \text{ or } a_{t_0'}^2 = 0\}.$$

Then using the intermediate value theorem, there exist $\xi_{ij}(s) \in [x_{ij}^1(s), x_{ij}^2(s)]$ such that

$$\|x_t^1 - x_t^2\| = \left\| \int_0^t (f(x_s^1) - f(x_s^2))ds \right\| \leq \int_0^t \sum_{i=1}^{3}\sum_{j=1}^{m} \left| \frac{\partial f}{\partial x_{ij}}(\xi_{ij}(s)) \right| \left| x_{ij}^1(s) - x_{ij}^2(s) \right| ds$$
$$\leq \int_0^t L(s)\|x_s^1 - x_s^2\|ds,$$

where $x_s^k = (x_{ij}(s))_{\substack{1\leq i\leq 3 \\ 1\leq j\leq m}}$ $(k \in \{1,2\})$ and $L(s) = \sum_{i=1}^{3}\sum_{j=1}^{m} \max \left| \frac{\partial f}{\partial x_{ij}}(x_s) \right|$, of which the maximum is over $x(s) = (x_{ij}(s))_{ij} \in [0,1]^{3m}$ such that $\forall i, j : x_{ij} \in [x_{ij}^1, x_{ij}^2]$, where by an abuse of notation, the bounds of interval $[x_{ij}^1, x_{ij}^2]$ can be switched depending on the minimum or maximum of the bounds.
By the Grönwall's inequality, we get

$$\|x_t^1 - x_t^2\| \leq \|x_0^1 - x_0^2\| \exp\left(\int_0^t L(s)ds\right) = 0.$$

This shows that $x_t^1 \equiv x_t^2$ for all $t \in [0, t_0']$. It also follows $t_0' = t_0$.

2.4 Simulation

The simulations show that the deterministic solution of the system of ODEs (2.2) fits well with our stochastic process, see Figure 2.1. The sequence of stochastic process $(X^N)_{N\geq 1}$ that we have constructed describes how the chain-referral process works on a network. When we consider the population with a very large number of people, the process $(X^N)_{N\geq 1}$ is asymptotically a deterministic function, which is a solution of a system of (2.2). To see numerically the convergence, we do a simulation: for $c = 3$, we vary N from 500 to 50000 and plot as a function of N the $\log$ of the quantity:

$$\int_0^1 (\|A_t^N - a_t\| + \|B_t^N - b_t\| + \|U_t^N - u_t\|)dt,$$

Figure 2.3. The speed of convergence has been studied in the case of Erdös-Rényi graphs in the PhD-thesis, by establishing a central limit theorem: Theorem 0.10.

By studying the solution of (2.2), we can obtain an approximation of the fraction of the population that has been interviewed when the CRS process stops. The proportion of the population discovered is then approximated by t_0.

The number of maximum coupon c plays an important role in how many people we could explore before there is no distributed coupons any more (when $\|a_t\| = 0$). By keeping all other parameters fixed and changing c, in the simulations of Figure 2.1, we see that the time t_0 are different. For example, with $m = 2$, $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{22} = 4, \lambda_{12} = 3$, we obtain the table 2.1.

c	1	2	3	4	5	6	...
t_0	0.18	0.91	0.94	0.95	0.95	0.95	...

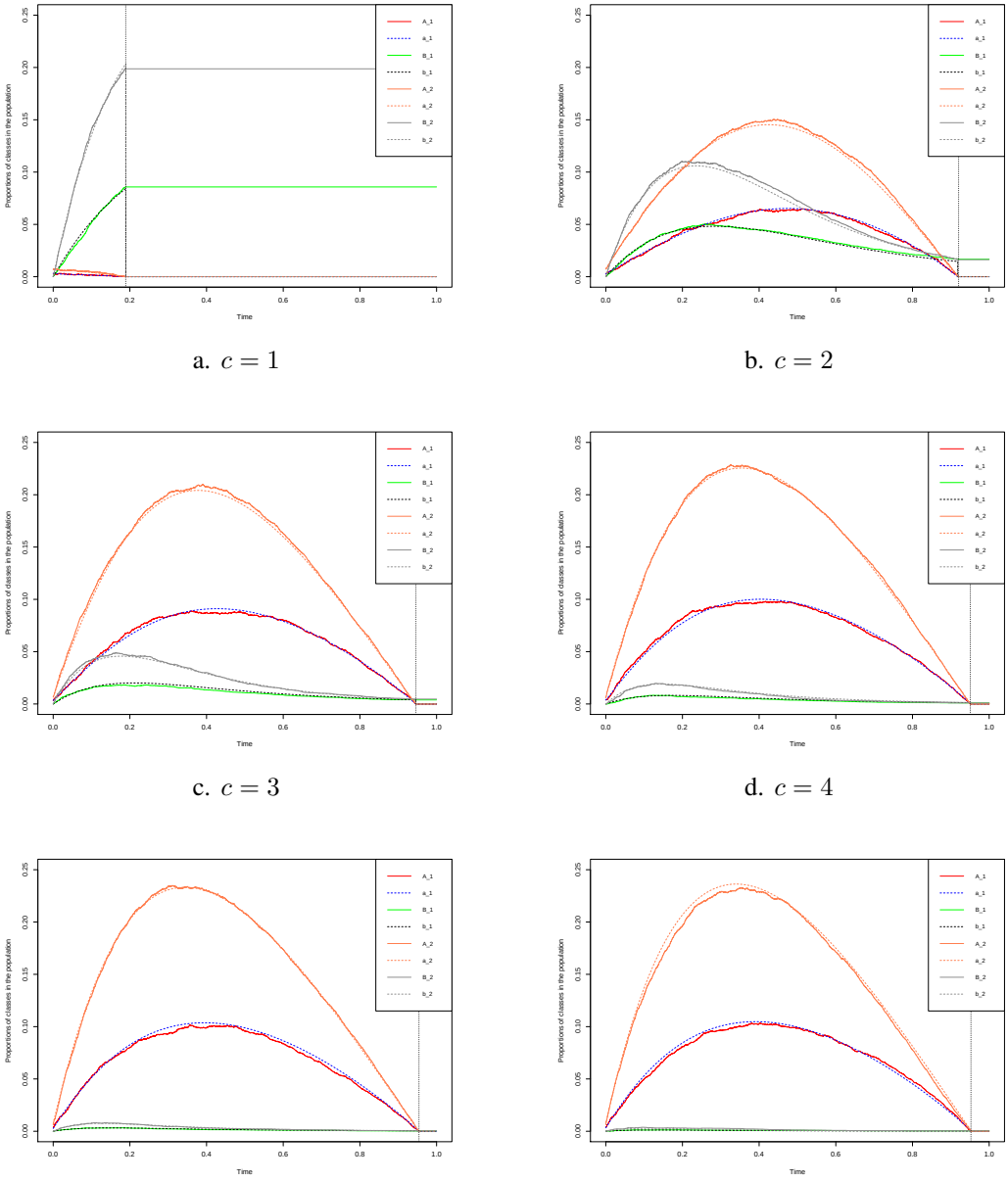
Table 2.1. Numerical computation of t_0 for varying parameters c.

If $c = 1$, even though the average number of neighbors are bigger than 1, the simple random walk describing the survey reaches only a very small number of people, see Figure 2.1.a.. The random walk stops when it encounters a node of degree 1 and can not propagate any more.
Furthermore, the parameter c also impacts the peaks (time and size) the curves corresponding to the number of distributed coupons. In case of a limited budget with a fixed number of interviews, a higher value of c can imply that we discover a larger fraction of the population since it allows more flexibility in the interviewees. From the Figure 2.4, we observe that the proportion of people receiving coupons gets bigger as c increases. If $c = 1$, the fraction of discovered population is small, which means that the survey is not so efficient. When c takes values from 4 to 6, the corresponding curves of $\|a_t\|$ are "close" and so are the times t_0. However, in these cases, the number of coupons spent during the CRS survey is large. We can also be interested in seeing how c impacts the part of population discovered when the survey stops after a fixed number of interviewed individuals. For example, consider the case when $N = 1000$ and assume that we start with $A_0 = 10$. The parameters of the SBM are $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{22} = 4, \lambda_{12} = 3$. Then when there have been approximately $\lfloor 0.2N \rfloor$ individuals interviewed, the proportion of the explored individuals: $\|A^N_{0.2}\| + \|B^N_{0.2}\|$ for each c varying from 1 to 6 is given in Table 2.2.

c	1	2	3	4	5	6
$\|A^{1000}_{0.2}\| + \|B^{1000}_{0.2}\|$	0.213	0.308	0.268	0.308	0.310	0.260

Table 2.2. Numerical computation of $\|A^N_t\| + \|B^N_t\|$ for varying parameters $c \in \{1, \dots, 6\}$ at time $t = 0.2$ and $N = 1000, A_0 = 10, \pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{22} = 4, \lambda_{12} = 3$.

Changing the parameters $\lambda_{k\ell}$ impacts the discovered proportion of types. For instant, let us take a bipartite random model $\pi = (1/3, 2/3), c = 3$ and $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 4$, which means that the people between communities are highly connected and there is no connection within community. In this case, the number of explored people without coupon of type 1 is quite small compared to the one of type 2, see Figure 2.2.



a. $c = 1$ b. $c = 2$

c. $c = 3$ d. $c = 4$

e. $c = 5$ f. $c = 6$

Figure 2.1. Plots of the proportions of classes in the population of size $N = 10000$ when c varies from 1 to 6 and all the others parameters are fixed: $\|A_0\| = 100$ the parameters $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$.

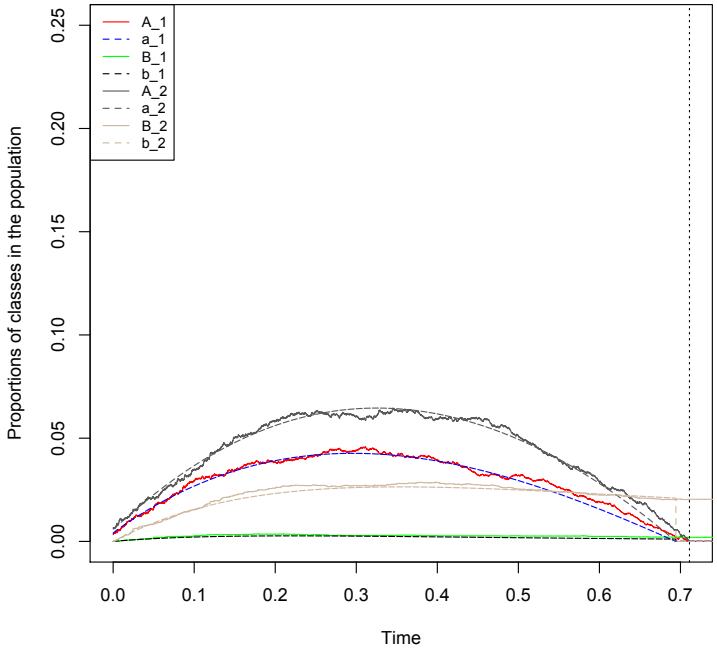


Figure 2.2. Plot the proportion of classes in the case $c = 3, N = 1000, A_0 = 10, \pi = (1/3, 2/3)$ and the graph is bipartite $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 4$.

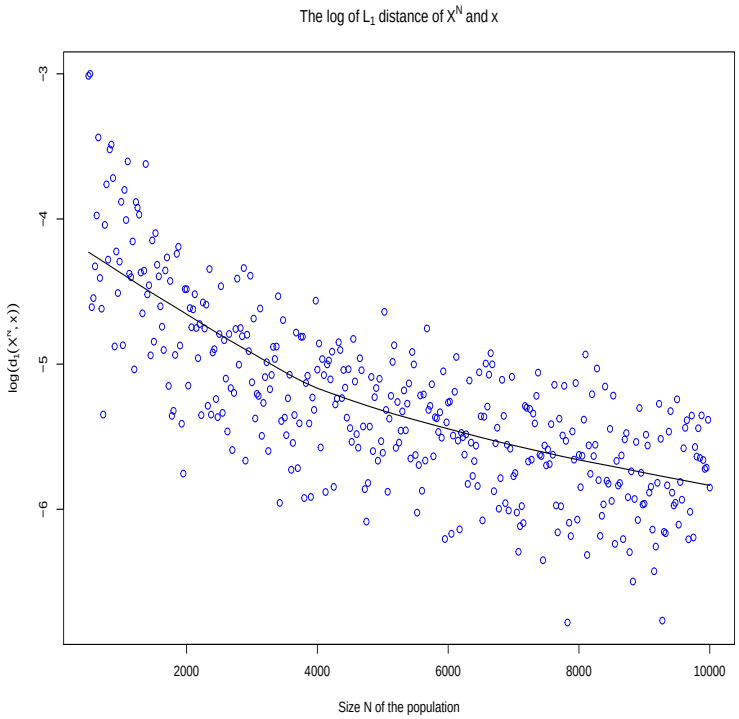


Figure 2.3. Scatter plot of $\ln d_1(X^N, x)$ along with the smoothing line suggesting the linear relationship between $\ln d_1(X^N, x)$ and N. The plot is done for the case $c = 3$, the number of initial individuals are 1% of the population and the size N varies from 500 to 10000. All other parameters are fixed: $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$.

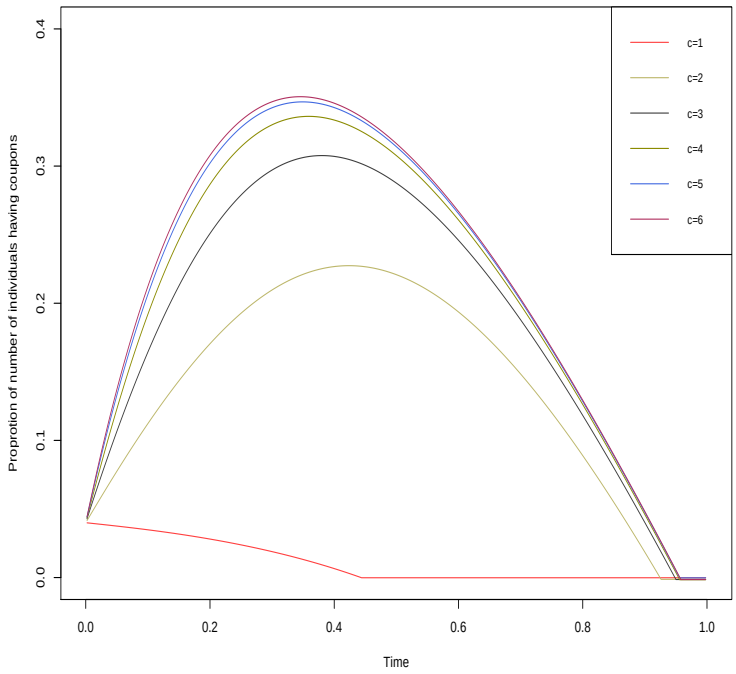


Figure 2.4. Plot the function $\|a\|$ for 6 cases: c takes values from 1 to 6. All other parameters are fixed: $\|a_0\| = 0.05, \pi = (1/3, 2/3), \lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$. The values $\|a_t\|$ represents the proportion of individuals having coupons at time t.

3. Estimation of dense stochastic block models visited by random walks

Contents

The work in this chapter is in collaboration with Viet Chi Tran[1] and is submitted to Electronic Journal of Statistics (EJS) [104].

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), by Labex Bézout

[1]University Gustave Eiffel

(ANR-10-LABX-0058) and by the Chair “Modélisation Mathématique et Biodiversité" of Veolia Environnement-Ecole Polytechnique-Museum National d’Histoire Naturelle-Fondation X.

3.1 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. [72]). 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 [47] or the ‘respondent-driven sampling’ (RDS) introduced by Heckathorn [49]. 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 [80] (see also [65]). 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 be in practice without replacement, an important numerical literature on the subject has followed (see e.g. [44, 45, 70]). Gile [43] proposed an improved estimator for population means taking into account the without replacement sampling, and Rohe established critical threshold for the design effects [65]. 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 [27] 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. [76] and Jauffret-Roustide et al. [58]. 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. [61]). 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 [54] 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 [20, 62]. Second, we propose a new estimation using new theoretical probabilistic results by Athreya and Röllin [4] 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. [29] 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. [66]). Recall that SBM graphs are a generalization of Erdös-Rényi 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 variables (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. [15, 16, 66]. Let us recall the definition of the SBM graphon.

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

$$I_q = \big[\sum_{k=1}^{q-1}\alpha_k, \sum_{k=1}^{q}\alpha_k\big), \qquad q \in \{1,\dots Q\}. \tag{3.1}$$

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

$$\kappa(x,y) = \sum_{q=1}^{Q}\sum_{r=1}^{Q} \pi_{qr}\ \mathbf{1}_{I_q}(x)\mathbf{1}_{I_r}(y). \tag{3.2}$$

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:

$$K_\kappa(x,dy) = \frac{\kappa(x,y)dy}{\int_0^1 \kappa(x,v)dv} = \frac{\sum_{q=1}^{Q}\Big(\sum_{r=1}^{Q}\pi_{qr}\ \mathbf{1}_{I_r}(y)\Big)\mathbf{1}_{I_q}(x)\ dy}{\sum_{q=1}^{Q}\Big(\sum_{r=1}^{Q}\pi_{qr}\alpha_r\Big)\mathbf{1}_{I_q}(x)}. \tag{3.3}$$

This random walk is the analogous of the classical random walk on a graph that jumps from a vertex to one of its neighboring 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 Röllin [4], 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 3.1 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 [29, 69, 79]. 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 3.5).

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 3.3 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 [29].
The second approach developed in Section 3.4 is inspired by the recent work of Athreya and Röllin [4]. 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 [29]. Based on this estimator that is not consistent in our case, we propose a new consistent estimator of κ.

3.2 Probabilistic setting

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

estimating the graphon κ from the RDS subgraph G_n.

3.2.1 Exploration by a random walk

Assumption 3.1 In all the paper, we assume that κ is the graphon of an SBM graph (see (3.2)) 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 3.1 Under Assumptions 3.1, the random walk $X = (X_n)_{n\geq 1}$ admits a unique invariant probability measure

$$m(dx) = \frac{\int_0^1 \kappa(x,v)dv}{\int_0^1 \int_0^1 \kappa(u,v)du\, dv}\, dx = \frac{\sum_{q=1}^Q \left(\sum_{r=1}^Q \pi_{qr}\alpha_r\right)\mathbf{1}_{I_q}(x)\, dx}{\sum_{q=1}^Q \sum_{r=1}^Q \pi_{qr}\, \alpha_q\, \alpha_r}. \quad (3.4)$$

The general proof is given in [4, Prop. 4.1] but for the case of SBM graphons, the result is easy to prove.

From expression (3.4), 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.

3.2.2 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 [66].

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:

$$t(F,G) = \frac{\#\{\text{injections from } F \text{ to } G\}}{(n)_k} = \frac{1}{(n)_k} \sum_{(i_1,\cdots i_k)\in[\![1,n]\!]} \prod_{\{\ell,\ell'\}\in E(F)} \mathbf{1}_{i_\ell \sim_G i_{\ell'}} \quad (3.5)$$

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:

$$t(F,\kappa)=\int_{[0,1]^k}\prod_{\{\ell,\ell'\}\in E(F)}\kappa(x_\ell,x_{\ell'})dx_1\cdots dx_k. \tag{3.6}$$

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:

$$d_{\text{sub}}(G,G')=\sum_{i\geq 0}\frac{1}{2^i}\left|t(F_i,G)-t(F_i,G')\right| \tag{3.7}$$

The convergence of the large graphs to graphons can be expressed with this distance [66, Chapter 11].

3.2.3 Biases in the discovery of κ

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

$$\Gamma(x)=\frac{\sum_{q=1}^{Q}\sum_{r=1}^{Q}\left(\pi_{qr}\alpha_r\right)\left[\min\left(\alpha_q,\ x-\sum_{k=1}^{q-1}\alpha_k\right)\right]_+}{\sum_{q=1}^{Q}\sum_{r=1}^{Q}\pi_{qr}\,\alpha_q\,\alpha_r} \tag{3.8}$$

Athreya and Röllin [4] have proved that the graphon discovered by the RDS is biased:

Proposition 3.2 — Corollary 2.2 [4]. We have under Assumptions 3.1 that:

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

where the generalized inverse of Γ is

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

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

$$\kappa_{\Gamma^{-1}}(x,y)=\kappa\left(\Gamma^{-1}(x),\Gamma^{-1}(y)\right). \tag{3.9}$$

This proposition, that is true not only for SBM graphons but also in more general cases, as developed in [4], 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 3.1 When $Q=2$, the graphon is given:

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

The invariant probability measure is:

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

Then the cumulative distribution of m is:

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

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

$$\kappa_{\Gamma^{-1}}(x,y) := \begin{cases}\pi_{11}, & \text{if} \quad (x,y) \in [0,\Gamma(\alpha)] \times [0,\Gamma(\alpha)]; \\ \pi_{22}, & \text{if} \quad (x,y) \in [\Gamma(\alpha),1] \times [\Gamma(\alpha),1]; \\ \pi_{12}, & \text{otherwise;}\end{cases} \tag{3.10}$$

where

$$\Gamma(\alpha) = \frac{(\pi_{11}\alpha + \pi_{12}(1-\alpha))\alpha}{\pi_{11}\alpha^2 + 2\pi_{12}\alpha(1-\alpha) + \pi_{22}(1-\alpha)^2}. \tag{3.11}$$

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-Rényi) or when $\alpha = 1/2$ and $\pi_{11} = \pi_{22}$ (both types are symmetric).

3.2.4 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:

$$\Gamma_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i\leq x} \qquad \text{and} \qquad \Gamma_n^{-1}(y) = \inf\left\{x \in [0,1] \ : \ \Gamma_n(x) \geq y\right\}. \tag{3.12}$$

Lemma 3.1 Γ_n and Γ_n^{-1} converge a.s. uniformly to Γ and Γ^{-1} respectively.

Proof. The almost sure point-wise 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 non-decreasing and piece-wise affine bijection and the inverse bijection Γ^{-1} is also non-decreasing and piece-wise 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 $n \geq n_0$,

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

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,

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

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

3.3 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 3.3 makes only use of the connectivity information carried by the random variables Y_{ij}, where $Y = (Y_{ij})_{i,j\in[\![1,N]\!]}$ is the associated adjacency matrix of G_n. 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:

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

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

3.3.1 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 3.3 The complete likelihood of the observations is

$$\mathcal{L}(Z, Y, X, \theta) = \prod_{q=1}^{Q} \left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}} (1-\pi_{qq})^{N_n^q(N_n^q-1)/2}$$

$$\times \prod_{q\neq r}\left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}} (1-\pi_{qr})^{N_n^q N_n^r} \times \prod_{q=1}^{Q} \frac{\alpha_q^{N_n^q}}{(\sum_{q'=1}^{Q}\pi_{qq'}\alpha_{q'})^{N_n^q - \mathbf{1}_{Z_n=q}}}. \tag{3.13}$$

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:

$$\mathcal{L}(Z_i, Y_{ij}; i,j \in X^{(n)}; \theta) = \frac{\prod_{i=1}^{n}\alpha_{Z_i}}{\prod_{i=1}^{n-1}\sum_{q=1}^{Q}\pi_{Z_i q}\alpha_q} \times \prod_{\substack{i,j\in[\![1,n]\!]\\ X_i,X_j\in X^{(n)}}} b(Y_{ij}, \pi_{Z_i Z_j}), \tag{3.14}$$

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, rewriting the above likelihood using N_n^q, $N_n^{q\leftrightarrow r}$, we obtain (3.13). ∎

Proposition 3.4 The MLE $\widehat{\theta} = (\widehat{\alpha}, \widehat{\pi})$ is the solution of the following system of equations:

$$\sum_{m=1}^{n}\frac{\mathbf{1}_{Z_m=q}}{\alpha_q} - \sum_{m=1}^{n-1}\frac{\pi_{Z_m q}}{\sum_{q'=1}^{Q}\pi_{Z_m q'}\alpha'_q} = 0; \tag{3.15}$$

$$\sum_{m=1}^{n-1}\left(\frac{\mathbf{1}_{(Z_m,Z_{m+1})=(qr)}}{\pi_{qr}} - \frac{\alpha_r \mathbf{1}_{Z_m=q}}{\sum_{q'=1}^{Q}\pi_{qq'}\alpha_{q'}}\right) + \sum_{\substack{i,j: X_i,X_j\in X^{(n)}\\ \{X_i,X_j\}\notin H_n}}\left(\frac{Y_{i,j}}{\pi_{qr}} - \frac{1-Y_{i,j}}{1-\pi_{qr}}\right)\mathbf{1}_{(Z_i,Z_j)=(qr)} = 0. \tag{3.16}$$

Proof. The log likelihood of the observations is:

$$\log\mathcal{L} = \sum_{q=1}^{Q}\left(N_n^q \log\alpha_q - (N_n^q - \mathbf{1}_{Z_n=q})\log\Big(\sum_{q'=1}^{Q}\pi_{qq'}\alpha_{q'}\Big)\right) + \sum_{q=1}^{Q}\left(N_n^{q\leftrightarrow q}\log\Big(\frac{\pi_{qq}}{1-\pi_{qq}}\Big) + \frac{N_n^q(N_n^q-1)}{2}\log(1-\pi_{qq})\right)$$

$$+\sum_{q=1}^{Q}\Big(\sum_{r\neq q} N_n^{q\leftrightarrow r}\log\big(\frac{\pi_{qr}}{1-\pi_{qr}}\big)+N_n^q N_n^r\log(1-\pi_{qr})\Big)$$

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

$$\frac{N_n^q}{\alpha_q}-\sum_{p=1}^{Q}\frac{N_n^p\pi_{pq}}{\sum_{q'=1}^{Q}\pi_{pq'}\alpha_{q'}}=0; \tag{3.17}$$

$$\frac{N_n^{q\leftrightarrow r}}{\pi_{qr}}-\frac{N_n^{q\leftrightarrow r}}{1-\pi_{qr}}-N_n^q\frac{\alpha_r}{\sum_{q'=1}^{Q}\pi_{qq'}\alpha_{q'}}=0. \tag{3.18}$$

The identifiability of the model is a result by Allman et al. [33]. Since the likelihood is differentiable, there exists a sequence of solutions of (3.17) that converge to the true parameter θ. ■

Remark 3.1 Notice that in absence of bias, the classical likelihood, as obtained in Daudin et al. [29] is:

$$\begin{aligned}\mathcal{L}^{\text{class}}(Z_i,Y_{ij};\theta)&=\prod_{i=1}^{n}\alpha_{Z_i}\times\prod_{i,j\in(X_n)}b(Y_{ij},\pi_{Z_iZ_j})\\&=\prod_{q=1}^{Q}\alpha_q^{N_n^q}\times\prod_{q=1}^{Q}\left(\frac{\pi_{qq}}{1-\pi_{qq}}\right)^{N_n^{q\leftrightarrow q}}(1-\pi_{qq})^{N_n^q(N_n^q-1)/2} \qquad (3.19)\\&\quad\times\prod_{q\neq r}\left(\frac{\pi_{qr}}{1-\pi_{qr}}\right)^{N_n^{q\leftrightarrow r}}(1-\pi_{qr})^{N_n^qN_n^r}. \qquad (3.20)\end{aligned}$$

The difference between (3.19) and (3.14) 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:

$$\widehat{\alpha}_q^{\text{class}}=\frac{N_n^q}{n},\qquad \widehat{\pi}_{qr}^{\text{class}}=\frac{N_n^{q\leftrightarrow r}}{N_n^qN_n^r},\qquad \widehat{\pi}_{qq}^{\text{class}}=\frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q-1)}. \tag{3.21}$$

Here, for the likelihood (3.13), the MLE which solves (3.15) is not explicit any more. In Section 3.3.1, we detail in the case of two classes ($Q=2$) the computation of the MLE.

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

$$\frac{N_n^1}{\widehat{\alpha}} - \frac{N_n^1\widehat{\pi_{11}}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} - \frac{N_n^2\widehat{\pi_{12}}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{3.22}$$

$$\frac{N_n^2}{1-\widehat{\alpha}} - \frac{N_n^1\widehat{\pi_{12}}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} - \frac{N_n^2\widehat{\pi_{22}}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{3.23}$$

$$\frac{N_n^{1\leftrightarrow 1}}{\widehat{\pi_{11}}} - \frac{N_n^{1\nleftrightarrow 1}}{1-\widehat{\pi_{11}}} - \frac{N_n^1\widehat{\alpha}}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} = 0; \tag{3.24}$$

$$\frac{N_n^{1\leftrightarrow 2}}{\widehat{\pi_{12}}} - \frac{N_n^{1\nleftrightarrow 2}}{1-\widehat{\pi_{12}}} - \frac{N_n^1(1-\widehat{\alpha})}{\widehat{\pi_{11}}\widehat{\alpha} + \widehat{\pi_{12}}(1-\widehat{\alpha})} = 0; \tag{3.25}$$

$$\frac{N_n^{2\leftrightarrow 1}}{\widehat{\pi_{12}}} - \frac{N_n^{1\nleftrightarrow 2}}{1-\widehat{\pi_{12}}} - \frac{N_n^2\widehat{\alpha}}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0; \tag{3.26}$$

$$\frac{N_n^{2\leftrightarrow 2}}{\widehat{\pi_{22}}} - \frac{N_n^{2\nleftrightarrow 2}}{1-\widehat{\pi_{22}}} - \frac{N_n^2(1-\widehat{\alpha})}{\widehat{\pi_{12}}\widehat{\alpha} + \widehat{\pi_{22}}(1-\widehat{\alpha})} = 0. \tag{3.27}$$

Proposition 3.5 The MLE $\widehat{\theta} = (\widehat{\alpha}, \widehat{\pi_{11}}, \widehat{\pi_{12}}, \widehat{\pi_{22}})$ can be expressed as a function of $\widehat{\pi}_{12}$:

$$\widehat{\pi_{11}} = \frac{(N_n^{1\leftrightarrow 1} + N_n^{1\leftrightarrow 2} - N_n^1) - (N_n^1 N_n^2 - N_n^1 + N_n^{1\leftrightarrow 1})\widehat{\pi_{12}}}{(\frac{N_n^1(N_n^1-1)}{2} - N_n^1 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^1(N_n^1-1)}{2} + N_n^1 N_n^2 - N_n^1)\widehat{\pi_{12}}}, \tag{3.28}$$

$$\widehat{\pi_{22}} = \frac{(N_n^{2\leftrightarrow 2} + N_n^{1\leftrightarrow 2} - N_n^2) - (N_n^{2\leftrightarrow 2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}}{(\frac{N_n^2(N_n^2-1)}{2} - N_n^2 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^2(N_n^2-1)}{2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}}, \tag{3.29}$$

$$\widehat{\alpha} = \frac{\widehat{\beta}}{1+\widehat{\beta}}, \tag{3.30}$$

with

$$\widehat{\beta} = \frac{(N_n^1 - N_n^2)\widehat{\pi_{12}} + \sqrt{(N_n^1 - N_n^2)^2\widehat{\pi_{12}}^2 + 4N_n^1 N_n^2\widehat{\pi_{11}}\widehat{\pi_{22}}}}{2N_n^2\widehat{\pi_{11}}}, \tag{3.31}$$

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

$$\begin{aligned}\widehat{\pi_{12}}^2 =& \frac{(N_n^{1\leftrightarrow 1} + N_n^{1\leftrightarrow 2} - N_n^1) - (N_n^1 N_n^2 - N_n^1 + N_n^{1\leftrightarrow 1})\widehat{\pi_{12}}}{(\frac{N_n^1(N_n^1-1)}{2} - N_n^1 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^1(N_n^1-1)}{2} + N_n^1 N_n^2 - N_n^1)\widehat{\pi_{12}}} \\ &\times \frac{(N_n^{2\leftrightarrow 2} + N_n^{1\leftrightarrow 2} - N_n^2) - (N_n^{2\leftrightarrow 2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}}{(\frac{N_n^2(N_n^2-1)}{2} - N_n^2 + N_n^{1\leftrightarrow 2}) - (\frac{N_n^2(N_n^2-1)}{2} + N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}} \\ &\times \frac{(N_n^{1\leftrightarrow 2} - N_n^1 N_n^2\widehat{\pi_{12}})^2}{[(N_n^{1\leftrightarrow 2} - N_n^1) - (N_n^1 N_n^2 - N_n^1)\widehat{\pi_{12}}][(N_n^{1\leftrightarrow 2} - N_n^2) - (N_n^1 N_n^2 - N_n^2)\widehat{\pi_{12}}]}.\end{aligned} \tag{3.32}$$

Proof. Multiply (3.24) by $\widehat{\pi_{11}}$ and (3.25) by $\widehat{\pi_{12}}$, and sum them up, we have

$$N_n^{1\nleftrightarrow 1}\frac{\widehat{\pi_{11}}}{1-\widehat{\pi_{11}}} + N_n^{1\nleftrightarrow 2}\frac{\widehat{\pi_{12}}}{1-\widehat{\pi_{12}}} = N_n^{1\leftrightarrow 1} + N_n^{1\leftrightarrow 2} - N_n^1. \tag{3.33}$$

Similarly, from equations (3.26) and (3.27), we deduce

$$N_n^{1\nleftrightarrow 2}\frac{\widehat{\pi_{12}}}{1-\widehat{\pi_{12}}}+N_n^{2\nleftrightarrow 2}\frac{\widehat{\pi_{22}}}{1-\widehat{\pi_{22}}}=N_n^{1\leftrightarrow 2}+N_n^{2\leftrightarrow 2}-N_n^2. \tag{3.34}$$

Also, the system of equations (3.24)-(3.27) gives

$$\left(\frac{N_n^{1\leftrightarrow 1}}{\widehat{\pi_{11}}}-\frac{N_n^{1\nleftrightarrow 1}}{1-\widehat{\pi_{11}}}\right)\left(\frac{N_n^{2\leftrightarrow 2}}{\widehat{\pi_{22}}}-\frac{N_n^{2\nleftrightarrow 2}}{1-\widehat{\pi_{22}}}\right)=\left(\frac{N_n^{1\leftrightarrow 2}}{\widehat{\pi_{12}}}-\frac{N_n^{1\nleftrightarrow 2}}{1-\widehat{\pi_{12}}}\right)^2. \tag{3.35}$$

Notice that $N^{1\nleftrightarrow 2}+N^{1\leftrightarrow 2}=N_n^1N_n^2$, $N^{1\nleftrightarrow 1}+N^{1\leftrightarrow 1}=\frac{N_n^1(N_n^1-1)}{2}$ and $N^{2\nleftrightarrow 2}+N^{2\leftrightarrow 2}=\frac{N_n^2(N_n^2-1)}{2}$ and we consider $\widehat{\pi_{12}}$ as a parameter. Solving the system (3.33)-(3.34) for $\widehat{\pi_{11}}, \widehat{\pi_{22}}$ provides the two first equations of (3.28). Using this, (3.35) is equivalent to:

$$\frac{(N_n^{1\leftrightarrow 2}-N_n^1N_n^2\widehat{\pi_{12}})^2}{[(N_n^{1\leftrightarrow 2}-N_n^1)-(N_n^1N_n^2-N_n^1)\widehat{\pi_{12}}][(N_n^{1\leftrightarrow 2}-N_n^2)-(N_n^1N_n^2-N_n^2)\widehat{\pi_{12}}]}\frac{\widehat{\pi_{11}}\widehat{\pi_{22}}}{\widehat{\pi_{12}}^2}=1. \tag{3.36}$$

This gives the (3.32).

For the estimator of α, let us denote $\beta:=\frac{\alpha}{(1-\alpha)}$. Then equations (3.22) and (3.23) are the same and equivalent to

$$\frac{N_n^1}{\widehat{\pi_{11}}\hat{\beta}+\widehat{\pi_{12}}}=\frac{N_n^2\hat{\beta}}{\widehat{\pi_{12}}\hat{\beta}+\widehat{\pi_{22}}}. \tag{3.37}$$

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 (3.32), see Fig. 3.1. 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 (3.32), we use the bisection method with the following constraints:

- The equation (3.32) has 4 excluded values that make the denominator zero:

$$\bar{\pi}_{12}^1=\frac{N_n^{1\leftrightarrow 2}-N_n^2}{N_n^1N_n^2-N_n^2}\qquad \bar{\pi}_{12}^2=\frac{N_n^{1\leftrightarrow 2}-N_n^1}{N_n^1N_n^2-N_n^1} \tag{3.38}$$

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

  It is observed that $\max(\bar{\pi}_{12}^1,\bar{\pi}_{12}^2)<\min(\bar{\pi}_{12}^3,\bar{\pi}_{12}^4)$. And if $N_n^1<N_n^2$, we have them ordered: $\bar{\pi}_{12}^1<\bar{\pi}_{12}^2<\bar{\pi}_{12}^3<\bar{\pi}_{12}^4$.
- All the estimators $\widehat{\pi_{11}}, \widehat{\pi_{12}}, \widehat{\pi_{22}}$ and $\widehat{\alpha}$ take values in the interval $(0,1)$.

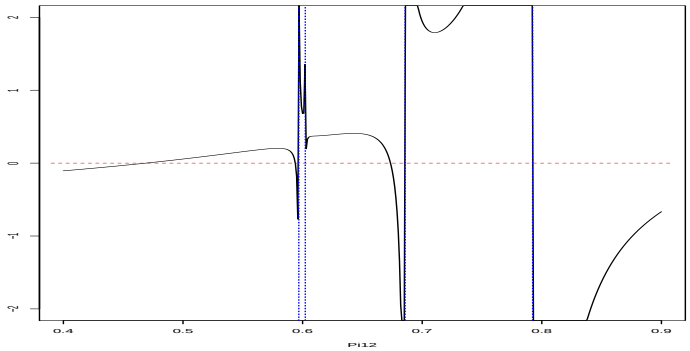


Figure 3.1. *E*quation (3.32) can be rewritten as $\phi(\pi_{12}) = 0$. The function ϕ is represented graphically on the figure above as a function of π_{12}. The vertical dotted lines correspond to the excluded values $\bar{\pi}^1_{12}, \dots \bar{\pi}^4_{12}$ given in (3.38).

Taking care of the points above, we solve (3.32) 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 (3.32), 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 3.5.

3.3.2 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 (3.14) over all the possible values of the unobserved variables Z:

$$\mathcal{L}(Y_{ij}; i,j \in [\![1,n]\!]; \theta) = \sum_{q_1,\cdots q_n=1}^{Q} \Big[\prod_{i=1}^{n} \mathbf{1}_{Z_i=q_i} \frac{\prod_{i=1}^{n} \alpha_{q_i}}{\prod_{i=1}^{n-1} \sum_{q=1}^{Q} \pi_{q_i q} \alpha_q} \times \prod_{i,j:X_i,X_j \in X^{(n)}} b(Y_{ij}, \pi_{q_i q_j}) \Big], \quad (3.39)$$

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 [29] by adapting the expression to our setting with the likelihood (3.13).

Let us sum up the EM algorithm (see e.g. [20, 21, 62]). 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 [20, 62], proceeding as follows.

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 3.3.2, where we use a variational approach.
- Calculate the acceptance probability

$$\omega(Z^{(k-1)}, Z^c) := \min\left\{1, \frac{\mathcal{L}(Z^c, Y, \theta^{(k-1)}) \cdot q_{\theta^{(k-1)}}(Z^{(k-1)}|Z^c)}{\mathcal{L}(Z^{(k-1)}, Y, \theta^{(k-1)}) \cdot q_{\theta^{(k-1)}}(Z^c|Z^{(k-1)})}\right\}; \tag{3.40}$$

- 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

$$\mathcal{Q}^{(k)}(\theta) = \mathcal{Q}^{(k-1)}(\theta) + s_k\left(\log \mathcal{L}(Z_i^{(k)}, Y_{ij}, \theta) - \mathcal{Q}^{(k-1)}(\theta)\right), \tag{3.41}$$

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 $\mathcal{Q}^{(k)}$

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{Q}^{(k)}(\theta). \tag{3.42}$$

Kuhn and Lavielle studied the convergence of the sequence $\theta^{(k)}$ in [62]. In the particular case of SBM, the consistency of EM and variational methods has been studied by Célisse et al. [22] and the asymptotic normality has been studied by Bickel et al. [9]. The likelihood that is considered here differs and these results can not be directly applied, but a study along these lines could be investigated.

Variational approach

For the proposal distribution $q_{\theta^{(k-1)}}(. \mid Z^{(k-1)})$ of $Z^{(k)}$, we follow Daudin et al. [29], who use a variational approach. Let us recall the main idea of this approach. The general strategy has been described in Jordan et al. [60] or Jaakkola [96].

Recall the likelihood $\mathcal{L}(Y, \theta)$ of the incomplete data (3.39). The idea of the variational approach is to replace the likelihood by a lower bound:

$$\mathcal{J}(R_{Y,\theta}) = \mathcal{L}(Y, \theta) - \mathrm{KL}(R_{Y,\theta}(Z), \mathcal{L}(Z|Y, \theta)), \tag{3.43}$$

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,

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

Following [29], 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 (3.43), 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,

$$R_{Y,\theta}(Z) = \prod_{i=1}^{n} \tau_{i,Z_i}. \tag{3.44}$$

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

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

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 (3.44) with the parameters $\widehat{\tau}$ given in the next proposition:

Proposition 3.6 Given α, π, the optimal parameter

$$\hat{\tau} := \arg\max_{\tau} \mathcal{J}(R_{Y,\theta}), \tag{3.45}$$

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

$$\tau_{iq} \propto \frac{\alpha_q}{\sum_{\ell=1}^{Q} \pi_{q\ell}\alpha_\ell} \prod_{i \neq j} \prod_{\ell=1}^{Q} b(Y_{ij}, \pi_{q\ell})^{\tau_{j\ell}}. \tag{3.46}$$

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

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

Similarly,

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

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,

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

$$-\sum_{i=1}^{n-1}\sum_{q=1}^{Q}\log\left(\sum_{r=1}^{Q}\pi_{qr}\alpha_r\right)\tau_{iq}. \tag{3.47}$$

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 $\text{Lag}(\tau,\lambda) := \mathcal{J}(R_{Y,\theta}) + \sum_{i=1}^{n}\lambda_i(\sum_{q=1}^{Q}\tau_{iq} - 1)$, where λ_i is the Lagrange multiplier. Take the derivative of $\mathcal{L}ag$ w.r.t. λ_i and τ, we have

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

The optimal solution must satisfy $\dfrac{\partial \text{Lag}}{\partial \lambda_i} = \dfrac{\partial \text{Lag}}{\partial \tau_{iq}} = 0$, 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,

$$\tau_{iq} = e^{\lambda_i - 1}\frac{\alpha_q}{\sum_{r=1}^{Q}\pi_{qr}\alpha_r}\prod_{i\neq j}\prod_{r=1}^{Q} b(Y_{ij},\pi_{qr})^{\tau_{jr}}. \tag{3.48}$$

■

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 3.7 When $Q = 2$, the variational parameter τ_i has formula:

$$\tau_i = \frac{\phi_i(\tau)}{1+\phi_i(\tau)} =: \Phi_i(\tau), \tag{3.49}$$

where

$$\phi_i(\tau) := \frac{\alpha}{1-\alpha}\frac{\alpha\pi_{21} + (1-\alpha)\pi_{22}}{\alpha\pi_{11} + (1-\alpha)\pi_{12}}\prod_{j\neq i}\left(\frac{b(Y_{ij},\pi_{12})}{b(Y_{ij},\pi_{22})}\right)^{1/2} \times \prod_{j\neq i}\left(\frac{b(Y_{ij},\pi_{11})b(Y_{ij},\pi_{22})}{b(Y_{ij},\pi_{12})^2}\right)^{\tau_j/2}. \tag{3.50}$$

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:

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

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

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

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

$$\begin{aligned}\frac{\tau_i}{1-\tau_i} &= \frac{\alpha}{1-\alpha}\times\frac{\alpha\pi_{11}+(1-\alpha)\pi_{12}}{\alpha\pi_{21}+(1-\alpha)\pi_{22}}\times\prod_{j\neq i}\left(\frac{b(Y_{ij},\pi_{12})}{b(Y_{ij},\pi_{22})}\right)^{1/2}\\ &\qquad\times\prod_{j\neq i}\left(\frac{b(Y_{ij},\pi_{11})b(Y_{ij},\pi_{22})}{b(Y_{ij},\pi_{12})^2}\right)^{\tau_j/2}\\ &= \phi_i(\tau). \qquad (3.51)\end{aligned}$$

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

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 3.7. The parameters in (3.50) 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

$$\mu(Z^{(k-1)}, Z^c) := \min\left\{1, \frac{\mathcal{L}_{\text{complete}}(Z^c, Y, \theta^{(k-1)}) q_{\theta^{(k-1)}}(Z^{(k-1)}|Z^c)}{\mathcal{L}_{\text{complete}}(Z^{(k-1)}, Y, \theta^{(k-1)}) q_{\theta^{(k-1)}}(Z^c|Z^{(k-1)})}\right\}, \quad (3.52)$$

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

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

and

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

3.4 Estimation via biased graphon and 'classical likelihood'

In Section 3.3, 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 (3.21). It turns out that we can study the asymptotic bias of this estimator thanks to the result of Athreya and Röllin [4]. 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.

3.4.1 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 $\rho = (\rho_{qr})_{q,r\in[\![1,Q]\!]}$:

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

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

$$J_q = [\sum_{k=1}^{q-1}\gamma_k, \sum_{k=1}^{q}\gamma_k), \qquad q \in [\![1,Q]\!]. \quad (3.53)$$

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 3.2 Denote by

$$\widehat{\gamma}_q^n := \frac{N_n^q}{n}; \quad \widehat{\rho}_{qr}^n := \frac{N_n^{q\leftrightarrow r}}{N_n^q N_n^r} \quad \text{for} \quad q \neq r \quad \text{and} \quad \widehat{\rho}_{qq}^n := \frac{2N_n^{q\leftrightarrow q}}{N_n^q(N_n^q-1)}. \tag{3.54}$$

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

$$\widehat{\chi}_n(x,y) := \sum_{q=1}^{Q}\sum_{r=1}^{Q} \widehat{\rho}_{qr}^n \mathbf{1}_{J_q^n}(x)\mathbf{1}_{J_r^n}(y), \tag{3.55}$$

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

We notice that this estimator corresponds to the MLE in the 'classical case' (see (3.21)). Thanks to the Proposition 3.2 (due to [4]), we can study the asymptotic limit of $\widehat{\chi}_n$.

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 3.8 We have under Assumption 3.1 that:
(i) when $n \to +\infty$,

$$\lim_{n\to+\infty} d_{sub}(G_n, \widehat{\chi}_n) = 0. \tag{3.56}$$

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

$$\lim_{n\to+\infty} d_{\text{sub}}(\widehat{\chi}_n, \kappa_{\Gamma^{-1}}) = 0. \tag{3.57}$$

Proof. We postpone the proof of Proposition 3.8 (i) to the Section 3.4.1. For the point (ii), we have:

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

The first term in the right hand side is upper bounded by C/n by Proposition 3.8. The second term is the Proposition 3.2 shown in [4, Corollary 2.2]. ∎

As a consequence, using the result of Athreya and Röllin [4] (see Proposition 3.2), we obtain:

Proposition 3.9 Under Assumptions 3.1,
(i) $\widehat{\rho}$ is a consistent estimator of π, and for $q, r \in [\![1, Q]\!]$,

$$\lim_{n\to+\infty} \widehat{\rho}^n_{qr} = \pi_{qr}, \qquad \text{and} \qquad \lim_{n\to+\infty} \widehat{\gamma}^n_q = \Gamma(\sum_{r=1}^{q} \alpha_r) - \Gamma\big(\sum_{r=1}^{q-1} \alpha_r\big) =: \gamma_q. \tag{3.58}$$

It follows that a consistent estimator of α_q is

$$\widehat{\alpha}^n_q = \Gamma_n^{-1}\big(\sum_{r=1}^{q} \widehat{\gamma}^n_r\big) - \Gamma_n^{-1}\big(\sum_{r=1}^{q-1} \widehat{\gamma}^n_r\big). \tag{3.59}$$

(ii) In the special case of $Q = 2$, an estimator of α_1 is $\widehat{\alpha}^n_1 = \Gamma_n^{-1}(\widehat{\gamma}^n_1)$.

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

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

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 $\widehat{\rho}^n_{qr}$ is a consistent estimator of π_{qr}. Rewrite $\widehat{\rho}^n_{qr}$ as

$$\widehat{\rho}^n_{qr} = \frac{N^{q\leftrightarrow r}_n/n^2}{\frac{N^q_n}{n}\frac{N^r_n}{n}} = \frac{1}{\widehat{\gamma}^n_q\widehat{\gamma}^n_r}\frac{1}{n^2}N^{q\leftrightarrow r}_n.$$

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 $N^{q\leftrightarrow r}_n$: two cases have to be distinguished.

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

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

Then,

$$\widehat{\rho}^n_{qr} = \frac{1}{\widehat{\gamma}^n_q\widehat{\gamma}^n_r n}\left(\frac{1}{n}\sum_{i=1}^{n-1} \mathbf{1}_{X_i\in I_q, X_{i+1}\in I_r}\right) + \frac{1}{n^2}\sum_{\substack{1\le i,j\le n \\ \{X_i,X_j\}\notin E(H_n)}} \frac{\mathbf{1}_{i\sim_{G_n} j}\mathbf{1}_{X_i\in I_q, X_j\in I_r}}{\widehat{\gamma}^n_q\widehat{\gamma}^n_r} \tag{3.60}$$

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 (3.60) converges to 0 in probability.

Consider now the second term in the r.h.s. of (3.60). 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

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

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

$$\begin{aligned}
&\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon\widehat{\gamma}_q^n\widehat{\gamma}_r^n-\pi_{qr}\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|\right)\\
&\quad\le\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>c^2\varepsilon-\frac{c^3}{2}\varepsilon\right)\\
&\qquad+\mathbb{P}\left(\left|\frac{(n-1)(n-2)}{n^2}\gamma_q\gamma_r-\widehat{\gamma}_q^n\widehat{\gamma}_r^n\right|>\frac{c^3\varepsilon}{2\pi_{qr}}\right)+\mathbb{P}(\widehat{\gamma}_q^n\widehat{\gamma}_r^n<c^2).
\end{aligned}\tag{3.61}$$

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 (3.61) 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 [97] for function f , we obtain:

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

Note that $0<c<1$ then $c^2(1-c/2)>0$. We use Borel-Cantelli's Theorem to conclude that

$$\lim_{n\to+\infty}\mathbb{P}\left(\left|f(G_n)-\mathbb{E}[f(G_n)]\right|>c^2\varepsilon-\frac{c^3}{2}\varepsilon\right)=0$$

and hence,

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

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:

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

Then,

$$\widehat{\rho}_{qq}^n=\frac{1}{\widehat{\gamma}_q^n\big(n\widehat{\gamma}_q^n-1\big)}\left(\frac{1}{n}\sum_{i=1}^{n-1}\mathbf{1}_{X_i\in I_q,X_{i+1}\in I_q}\right)+\frac{1}{n^2}\sum_{\substack{1\leq i,j\leq n\\ \{X_i,X_j\}\notin E(H_n)}}\frac{\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_q}}{\widehat{\gamma}_q^n\big(\widehat{\gamma}_q^n-1/n\big)}\tag{3.62}$$

We have that the first term on r.h.s. of (3.62) converges in probability to 0 as in case 1. For the second term on r.h.s. of (3.62), we define the function f as in Case 1 by

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

For a fixed $\varepsilon>0$,

$$\mathbb{P}\left(\Big|\frac{1}{n^2}\sum_{\substack{1\le i,j\le n\\ \{X_i,X_j\}\notin E(H_n)}}\frac{\mathbf{1}_{i\sim_{G_n}j}\mathbf{1}_{X_i\in I_q,X_j\in I_q}}{\widehat{\gamma}_q^n\big(\widehat{\gamma}_q^n-1/n\big)}-\pi_{qq}\Big|>\varepsilon\right)$$

$$\le\mathbb{P}\Bigg(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>\varepsilon\widehat{\gamma}_q^n\big(\widehat{\gamma}_q^n-1/n\big) - \pi_{qq}\left|\frac{(n-1)(n-2)}{n^2}(\gamma_q)^2-\widehat{\gamma}_q^n\big(\widehat{\gamma}_q^n-1/n\big)\right|\Bigg)$$

$$\le\mathbb{P}\left(\big|f(G_n)-\mathbb{E}[f(G_n)]\big|>c\big(c-\frac{1}{n}\big)\varepsilon-\frac{c^3}{2}\varepsilon\right)+\mathbb{P}(\widehat{\gamma}_q^n<c) + \mathbb{P}\left(\left|\frac{(n-1)(n-2)}{n^2}(\gamma_q)^2-\widehat{\gamma}_q^n\big(\widehat{\gamma}_q^n-\frac{1}{n}\big)\right|>\frac{c^3\varepsilon}{2\pi_{qq}}\right).$$

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. ■

Proof of Proposition 3.8

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 $\gamma_n^1=:\gamma_n$ and $\gamma_\infty^1=:\gamma_\infty=\Gamma(\alpha)$.

Our purpose is to prove a convergence of graphons for the distance d_{sub} introduced in (3.7) using the densities (3.5). 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,

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

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

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

$$t(F, G_n) = \frac{1}{(n)_k} \sum_{(i_1,\cdots i_k)\in[\![1,n]\!]} \prod_{\{\ell,\ell'\}\in E(F)} \mathbf{1}_{i_\ell \sim_G i_{\ell'}} \tag{3.63}$$

$$t(F, \chi_n) = \int_{[0,1]^k} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}_n^{qr} \mathbf{1}_{J_q^n \times J_r^n}(x_\ell, x_{\ell'}) \right) dx_1 \cdots dx_k \tag{3.64}$$

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

$$\begin{aligned}
|t(F, G_n) - t(F, \chi_n)| &= \left| \frac{1}{(n)_2} \sum_{(i,j)\in[\![1,n]\!]} \mathbf{1}_{i\sim_{G_n} j} - \int_{[0,1]^2} \widehat{\chi}_n(x_1, x_2)\, dx_1 dx_2 \right| \\
&\leq \left| \frac{1}{(n)_2} \sum_{(i,j)\in[\![1,n]\!]} \left(\mathbf{1}_{i\sim_{G_n} j} - \widehat{\rho}_{Z_i,Z_j} \right) \right| \\
&\quad + \left| \frac{1}{(n)_2} \sum_{(i,j)\in[\![1,n]\!]} \widehat{\rho}_{Z_i,Z_j} - (\widehat{\gamma}_1^n)^2 \widehat{\rho}_{11}^n - 2\widehat{\gamma}_1^n(1-\widehat{\gamma}_1^n)\widehat{\rho}_{12}^n - (1-\widehat{\gamma}_1^n)^2 \widehat{\rho}_{22}^n \right| \\
&\leq \left| \frac{1}{(n)_2} \sum_{(i,j)\in[\![1,n]\!]} \left(\mathbf{1}_{i\sim_{G_n} j} - \widehat{\rho}_{Z_i,Z_j} \right) \right| + \left| \widehat{\rho}_{11}^n \left(\sum_{(i,j) \mid (Z_i,Z_j)=(1,1)} \frac{1}{(n)_2} - (\widehat{\gamma}_1^n)^2 \right) \right| \\
&\quad + \left| \widehat{\rho}_{22}^n \left(\sum_{(i,j) \mid (Z_i,Z_j)=(2,2)} \frac{1}{(n)_2} - (1-\widehat{\gamma}_1^n)^2 \right) \right| \\
&\qquad + \left| \widehat{\rho}_{12}^n \left(\sum_{\substack{(i,j) \mid (Z_i,Z_j)=(1,2) \\ \text{or}(Z_i,Z_j)=(2,1)}} \frac{1}{(n)_2} - 2\widehat{\gamma}_1^n(1-\widehat{\gamma}_1^n) \right) \right|.
\end{aligned}$$

By the law of large numbers and using (3.58) whose proof does not depend on the Proposition 3.8, the four terms converge to zero.

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

$$|t(F, G_n) - t(F, \chi_n)|$$

$$\leq \left| \frac{1}{(n)_k} \sum_{(i_1,\cdots i_k)\in[\![1,n]\!]} \prod_{\{\ell,\ell'\}\in E(F)} \mathbf{1}_{i_\ell\sim_G i_{\ell'}} - \frac{1}{(n)_k} \sum_{(i_1,\cdots,i_k)} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{Z_{i_\ell}=q, Z_{i_{\ell'}}=r} \right) \right|$$
$$+ \left| \frac{1}{(n)_k} \sum_{(i_1,\cdots,i_k)} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{Z_{i_\ell}=q, Z_{i_{\ell'}}=r} \right) - \frac{1}{n^k} \sum_{1\leq i_1,\cdots,i_k\leq n} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{Z_{i_\ell}=q, Z_{i_{\ell'}}=r} \right) \right|$$
$$+ \left| \frac{1}{n^k} \sum_{1\leq i_1,\cdots,i_k\leq n} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{Z_{i_\ell}=q, Z_{i_{\ell'}}=r} \right) - \int_{[0,1]^k} \prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{J^n_q\times J^n_r}(x_\ell, x_{\ell'}) \right) dx_1\cdots dx_k \right|.$$

As $\prod_{\{\ell,\ell'\}\in E(F)} \mathbf{1}_{i_\ell\sim_G i_{\ell'}}$ and $\prod_{\{\ell,\ell'\}\in E(F)} \left(\sum_{q,r=1}^{Q} \widehat{\rho}^n_{qr} \mathbf{1}_{Z_{i_\ell}=q, Z_{i_{\ell'}}=r} \right)$ 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

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

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

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

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.

3.4.2 Incomplete observations and graphon de-biasing

In Proposition 3.9, it is shown that the 'classical' SBM estimator (3.21) 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 [29]:

- We follow the algorithm described in Section 3.3.2, but with the likelihood $\mathcal{L}^{\text{class}}(Z_i, Y_{ij}; \theta)$ given in (3.19). 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 1 corresponds to a variational EM for the classical likelihood, for which the consistency and asymptotic normality have been established by Celisse et al. [22] and Bikel et al. [9].

Step 2: We estimate the cumulative distribution function Γ_n (see (3.12)) and deduce the graphon estimator $\widehat{\alpha}_1^n$ of α_1 using (3.59). This provides the estimator of κ:

$$\widehat{\kappa}_n(x,y) := \sum_{q=1}^{Q}\sum_{r=1}^{Q} \widehat{\rho}_{qr}^n \mathbf{1}_{[\sum_{k=1}^{q-1}\widehat{\alpha}_k^n, \sum_{k=1}^{q}\widehat{\alpha}_k^n)}(x) \mathbf{1}_{[\sum_{k=1}^{r-1}\widehat{\alpha}_k^n, \sum_{k=1}^{r}\widehat{\alpha}_k^n)}(y). \tag{3.65}$$

3.5 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 3.3.1 for complete observations by assuming that the types $Z_i \in \{1, 2\}$ are observed. In the estimation, the system of equations (3.22)-(3.27) is solved. For this, we look numerically for the zeros of (3.32) and choose the solution corresponding to the highest likelihood. For the bisection method ([34]), we use a grid of step 10^{-2}.
- the SAEM algorithm of Section 3.3.2 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 3.9 assuming complete observations,
- the debiasing of the Variational EM Algorithm (VEM) of Daudin et al. presented in Section 3.4.2. Again, we use $K = 200$ iterations for the EM 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. 3.2, and this allows us to estimate the associated mean squares errors (MSE) for each method, see Table 3.1.

Parameters	Complete likelihood	SAEM	De-biased graphon	De-biased graphon with VEM
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 3.1. *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 (3.59) 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 (3.30)). The shape of function $\beta = \frac{\alpha}{1-\alpha}$ (see figure 3.3) 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 (3.59)) 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 3.1). Notice first that the columns 2 and 4 of Table 3.1 are not completely equivalent, since the debiasing methods of Section 3.4 necessitate the knowledge of the positions X_i of the Markov chain, when the likelihood (3.13) necessitates only the connections Y_{ij} and the types Z_i's. Second, the updating of the types in the SAEM algorithm

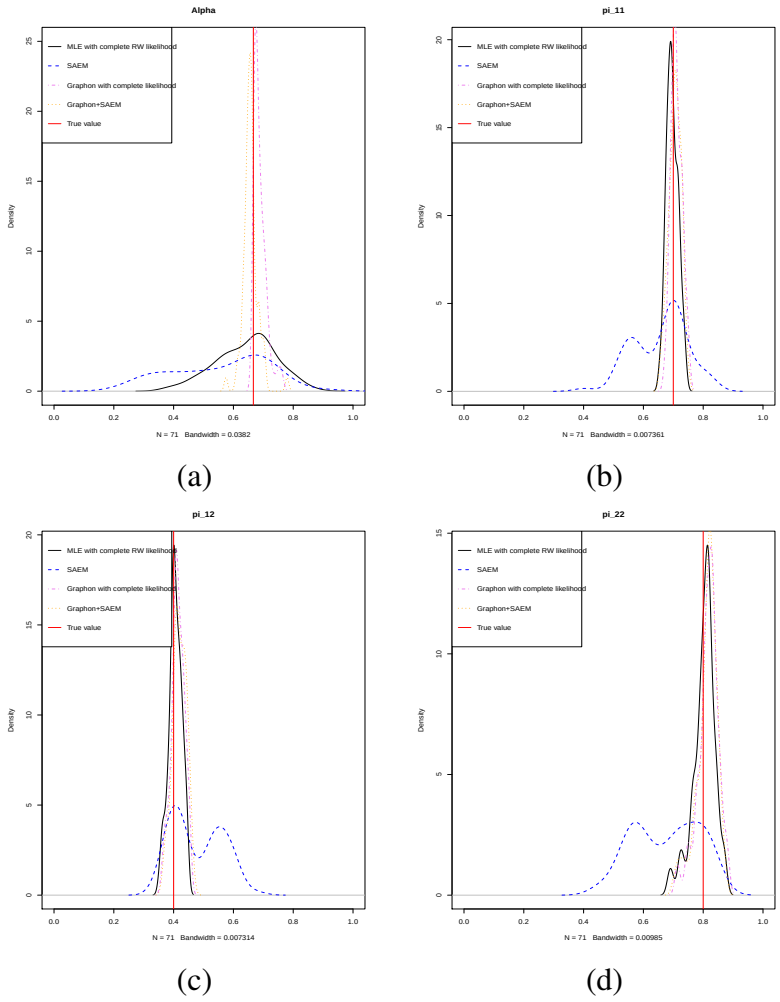


Figure 3.2. *E*stimation on complete data for a graph of $n = 50$ vertices with $Q = 2$ classes and parameters $\alpha_1 = 2/3$, $\pi_{11} = 0.7$, $\pi_{12} = \pi_{21} = 0.4$ and $\pi_{22} = 0.8$. 200 such graphs are simulated and the empirical distributions of the estimators are represented here with 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}.

is easier in Section 3.4.2 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 (3.19) seems to converge more easily than for the likelihood (3.13).

3.5.1 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 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

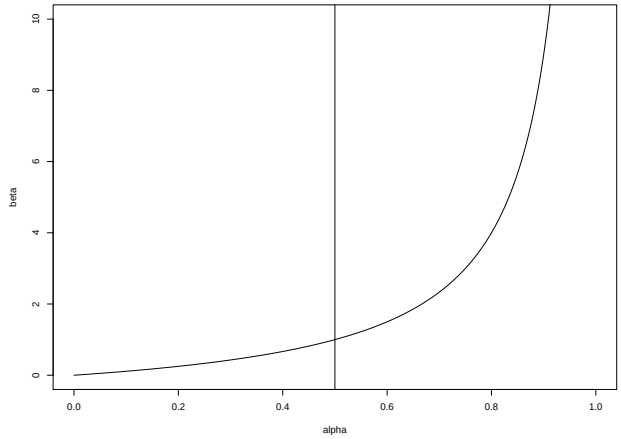


Figure 3.3. The correlation of β and α.

very robust and provides relatively large MSEs. An alternative approach is proposed by taking advantage of recent results by Athreya and Röllin [4]: 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

List of Figures

List of Tables

Bibliography

Articles

[1] E. Abbe. “Community detection and stochastic block models: recent developments”. In: *The Journal of Machine Learning Research* 18.1 (Jan. 2017), pages 6446–6531 (*cited on pages 13, 14, 16, 73, 98*).

[2] E. Abbe. “Community Detection and Stochastic Block Models”. In: *Foundations and Trends® in Communications and Information Theory* 14.1-2 (2018), pages 1–162 (*cited on pages 4, 13, 14*).

[3] E. Abbe, A. S. Bandeira, and G. Hall. “Exact Recovery in the Stochastic Block Model”. In: *IEEE Transactions on Information Theory* 62.1 (Jan. 2016), pages 471–487 (*cited on pages 4, 13*).

[4] S. Athreya and A. Röllin. “Dense graph limits under respondent-driven sampling”. In: *Annals of Applied Probability* 44 (2016), pages 2193–2210 (*cited on pages 30–32, 99–102, 115, 116, 126*).

[5] S. Athreya and A. Röllin. “Respondent driven sampling and sparse graph convergence”. In: *Electronic Communications in Probability* 23 (2018) (*cited on pages 30, 36, 74*).

[6] A. Bagheri and M. Saadati. “Exploring the effectiveness of chain referral methods in sampling hidden populations”. In: *Indian J. Sci. Technol* 8.30 (2015) (*cited on page 72*).

[7] P. Barbillon et al. “Stochastic Block Models for Multiplex networks: an application to networks of researchers”. In: *arXiv:1501.06444* (2015) (*cited on page 72*).

[8] A. D. Barbour and G. Reinert. “Approximating the epidemic curve”. In: *Electronic Journal of Probability* 18.54 (2013), page 2557 (*cited on pages 25, 27, 75*).

[9] P. Bickel et al. “Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels”. In: *The Annals of Statistics* 41.4 (2013), pages 1922–1943 (*cited on pages 110, 123*).

[10] B. Bollobás and O. Riordan. “Sparse graphs: metrics and random models”. In: *Random Structures and Algorithms* 29.1 (2011), pages 1–38 (*cited on page 20*).

[11] B. Bollobás and O. Riordan. “Asymptotic Normality of the Size of the Giant Component via a Random Walk”. In: *Journal of Combinatorial Theory Serie B* 102.1 (Jan. 2012), pages 53–61 (*cited on pages 21, 75*).

[12] C. Borgs, J. Chayes, and D. Gamarnik. “Convergent Sequences of Sparse Graphs: A Large Deviations Approach”. In: *Random Structures and Algorithms* 51.1 (Aug. 2017), pages 52–89 (*cited on pages 16, 20*).

[13] C. Borgs et al. “Random subgraphs of finite graphs. II. The lace expansion and the triangle condition”. In: *The Annals of Probability* 33.5 (Mar. 2005), pages 1886–1994 (*cited on page 20*).

[14] C. Borgs et al. “Counting Graph Homomorphisms”. In: *Topics in discrete Mathematics* 26 (2006), pages 315–371 (*cited on pages 16, 18*).

[15] C. Borgs et al. “Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing”. In: *Advances in Mathematics* 219.6 (Dec. 2008), pages 1801–1851 (*cited on pages 16–19, 21, 99*).

[16] C. Borgs et al. “Convergent Sequences of Dense Graphs II: Multiway Cuts and Statistical Physics”. In: *Annals of Mathematics* 176.1 (Dec. 2012), pages 151–219 (*cited on pages 16–18, 99*).

[17] C. Borgs et al. “An L^p theory of sparse graph convergence II: LD convergence, quotients and right convergence”. In: *The Annals of Probability* 46 (2018), pages 337–396 (*cited on pages 18, 20*).

[18] H. Cohn C. Borgs J. T. Chayes and Y. Zhao. “An theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions”. In: *Transactions of the American Mathematical Society* 372.5 (2019), pages 3019–306 (*cited on pages 17, 18, 20*).

[19] J. Kahn C. Borgs J. Chayes and L. Lovász. “Left and right convergence of graphs with bounded degree”. In: *Random Structures and Algorithms* 42 (Apr. 2012), pages 1–28 (*cited on page 20*).

[20] G. Celeux, D. Chauveau, and J. Diebolt. “Stochastic versions of the em algorithm: an experimental study in the mixture case”. In: *Journal of Statistical Computation and Simulation* 55.4 (1996), pages 287–314 (*cited on pages 99, 109*).

[21] G. Celeux and J. Diebolt. “The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem”. In: *Computational Statistics Quarterly* 2 (1985), pages 73–82 (*cited on page 109*).

[22] A. Celisse, J. J. Daudin, and L. Pierre. “Consistency of maximum-likelihood and variational estimators in the stochastic block model”. In: *Electronic Journal of Statistics* 6 (2012), pages 1847–1899 (*cited on pages 110, 123*).

[23] S. Clémençon et al. “A statistical network analysis of the HIV/AIDS epidemics in Cuba”. In: *Social Network Analysis and Mining* 5 (2015), Art.58 (*cited on pages 3, 40*).

[24] A. Cousien et al. “Dynamic modelling of HCV transmission among people who inject drugs: a methodological review”. In: *Journal of Viral Hepatitis* 22.3 (2015), pages 213–229 (*cited on pages 3, 40*).

[25] A. Cousien et al. “Hepatitis C treatment as prevention of viral transmission and level-related morbidity in persons who inject drugs”. In: *Hepatology* 63.4 (2016), pages 1090–1101 (*cited on pages 3, 40*).

[26] F. W. Crawford. “The graphical structure of respondent-driven sampling”. In: *Sociological methodology* 46.1 (2016), pages 187–211 (*cited on page 30*).

[27] F. W. Crawford, J. Wu, and R. Heimer. “Hidden population size estimation from respondent-driven sampling: a network approach”. In: *Journal of the American Statistical Association* 113.522 (2018), pages 755–766 (*cited on pages 30, 40, 98*).

[28] Gabor Csardi and Tamas Nepusz. “The igraph software package for complex network research”. In: *InterJournal* Complex Systems (2006), page 1695 (*cited on page 14*).

[29] J. J. Daudin, F. Picard, and S. Robin. “A mixture model for random graphs”. In: *Statistics and Computing* 18.2 (2008), pages 173–183 (*cited on pages 99, 100, 106, 109–111, 123*).

[30] L. Decreusefond et al. “Large graph limit for a SIR process in random network with heterogeneous connectivity”. In: *Annals of Applied Probability* 22.2 (2012), pages 541–575 (*cited on page 75*).

[31] B. Delyon, M. Lavielle, and E. Moulines. “Convergence of a stochastic approximation version of the EM algorithm”. In: *Annals of statistics* (1999), pages 94–128 (*cited on page 34*).

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incoplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B* 39.1 (1977), pages 1–38 (*cited on page 34*).

[33] C. Matias E. Allman and J. Rhodes. “Parameter identifiability in a class of random graph mixture models”. In: *Journal of Statistical Planning and Inference* 141.5 (2011), pages 1719–1736 (*cited on page 106*).

[34] A. Eiger, K. Sikorski, and F. Stenger. “A bisection method for systems of nonlinear equations”. In: *ACM Transactions on Mathematical Software (TOMS)* 10.4 (1984), pages 367–377 (*cited on page 124*).

[35] N. Enriquez, G. Faraud, and L. Ménard. “Limiting shape of the depth first search tree in an Erdős-Rényi graph”. In: *Random Structures & Algorithms* 56.2 (2020), pages 501–516 (*cited on pages 21, 75*).

[36] A. Erdös and P. Rényi. “On random graphs”. In: *Publicationes Mathematicae* 6 (1959), pages 290–297 (*cited on pages 4, 7*).

[37] A. Erdös and P. Rényi. “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of Hungarian Academy of Science* 5 (1961) (*cited on pages 4, 7*).

[38] S. E. Fienberg and S. S. Wasserman. “Categorical data analysis of single sociometric relations”. In: *Sociological methodology* 12 (1981), pages 156–192 (*cited on page 13*).

[39] M. Freedman, L. Lovász, and A. Schrijver. "Reflection Positivity, Rank Connectivity, and Homomorphism of Graphs". In: *Journal of the American Mathematical Society* 20.1 (2007), pages 37–51 (*cited on page 16*).

[40] A. Frieze and R. Kannan. "Quick approximation to matrices and applications". In: *Combinatorica* 19.2 (1999), pages 175–220 (*cited on page 20*).

[41] D. M. Frost, J. T. Parsons, and J. E. Nanín. "Stigma, concealment and symptoms of depression as explanations for sexually transmitted infections among gay men". In: *Journal of health psychology* 12.4 (2007), pages 636–640 (*cited on page 40*).

[42] A. Gadde et al. "Active learning for community detection in stochastic block models". In: *IEEE International Symposium on Information Theory (ISIT)* (June 2016), pages 1889–18936 (*cited on pages 13, 72*).

[43] K. J. Gile. "Improved Inference for Respondent-Driven Sampling data with application to HIV prevalence estimation". In: *Journal of the American Statistical Association* 106.493 (2011), pages 135–146 (*cited on pages 29, 40, 98*).

[44] K. J. Gile and M. S. Handcock. "Respondent-driven sampling: An assessment of current methodology". In: *Sociological methodology* 40.1 (2010), pages 285–327 (*cited on pages 29, 98*).

[45] K. J. Gile and M. Salganik. "Diagnostics for respondent-driven sampling". In: *ournal of the Royal Statistical Society A* 178 (2015), pages 241–269 (*cited on pages 29, 98*).

[46] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *PNAS* 99.12 (June 2002), pages 7821–7826 (*cited on pages 13, 72*).

[47] L. A. Goodman. "Snowball sampling". In: *Annals of Mathematical Statistics* 32.1 (1961), pages 148–170 (*cited on pages 3, 40, 72, 98*).

[48] M.S. Handcock, K.J. Gile, and C.M. Mar. "Estimating hidden population size using Respondent-Driven Sampling data". In: *Electronic Journal of Statistics* 8.1 (2014), pages 1491–1521 (*cited on page 40*).

[49] D. D. Heckathorn. "Respondent-driven Sampling: a new approach to the study of hidden populations". In: *Social Problems* 44.1 (1997), pages 74–99 (*cited on pages 3, 21, 22, 40, 72, 98*).

[50] D. D. Heckathorn. "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations". In: *Social Problems* 49.1 (2002), pages 11–34 (*cited on pages 3, 72, 73*).

[51] D. D. Heckathorn, S. Semaan R. S. Broadhead, and J. J. Hudges. "Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25". In: *AIDS ans Behavior* 6.1 (Mar. 2002) (*cited on page 3*).

[52] M. J. Salganik D. D. Heckathorn. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling". In: *Sociological Methodology* 34.1 (2004), pages 193–239 (*cited on page 29*).

[53] M. Hellard et al. "The impact of injecting networks on hepatitis C transmission and treatment in people who inject drugs". In: *Hepatology* 60.6 (2014), pages 1861–1870 (*cited on pages 3, 40*).

[54] P. W Holland, K. B. Laskey, and S. Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pages 109–137 (*cited on pages 4, 13, 72, 98*).

[55] A. Jakubowski. "On the Skorokhod topology". In: *Annales de l'Institut Henri Poincaré* 22.3 (1986), pages 263–285 (*cited on pages 24, 46, 74*).

[56] S. Janson and P. Diaconis. "Graph limits and exchangeable random graphs". In: *Rendiconti di Matematica, Serie VII* 28 (2008), pages 33–61 (*cited on page 16*).

[57] S. Janson, M. Luczak, and P. Windridge. "Law of large numbers for the SIR epidemic on a random graph with given degrees". In: *Random Structures & Algorithms* 45.4 (2014), pages 726–763 (*cited on page 75*).

[58] M. Jauffret-Roustide et al. "A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design". In: *BMC Infectious Diseases* 0.1 (2009), page 113 (*cited on pages 40, 98*).

[59] A. Joffe and M. Métivier. "Weak Convergence of Sequences of Semimartingales with Applications to Multitype Branching Processes". In: *Advances in Applied Probability* 18 (1986), pages 20–65 (*cited on pages 24, 74*).

[60] M. Jordana et al. "An introduction to variational methods for graphical models". In: *Machine Learning* 37 (1999), pages 183–233 (*cited on page 110*).

[61] M. Khabbazian et al. "Novel sampling design for respondent-driven sampling". In: *Electronic Journal of Statistics* 11.2 (2017), pages 4769–4812 (*cited on page 98*).

[62] E. Kuhn and M. Lavielle. "Coupling a stochastic approximation version of EM with an MCMC procedure". In: *ESAIM: PS* 8 (2004), pages 115–131 (*cited on pages 99, 109, 110*).

[63] A. Lansky et al. "Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System". In: *Public Health Reports* 122.1 (2007), pages 48–55 (*cited on page 22*).

[64] P. Latouche, E. Birmele, and C. Ambroise. "Variational Bayesian inference and complexity control for stochastic block models". In: *Statistical Modelling* 12.1 (2012), pages 93–115 (*cited on page 35*).

[65] X. Li and K. Rohe. "Central limit theorems for network driven sampling". In: *Electronic Journal of Statistics* 11.2 (2017), pages 4871–4895 (*cited on pages 30, 40, 98*).

[66] L. Lovász. "Large networks and graph limits". In: *Colloquium Publications, American Mathematical Society, Rhode Island* 60 (2012) (*cited on pages 99, 101, 102*).

[67] L. Lovász and B. Szegedy. "Limits of dense graph sequences". In: *Journal of Combinatorial Theory, Series B* 96 (2006), pages 933–957 (*cited on pages 16–19*).

[68] L. Lovász and B. Szegedy. "Testing properties of graphs and functions". In: *Israel Journal of Mathematics* 178 (2010), pages 113–156 (*cited on page 16*).

[69] M. Mariadassou and T. Tabouy. "Consistency and asymptotic normality of stochastic block models estimators from sampled data". In: *arXiv preprint:1903.12488* (2019) (*cited on page 100*).

[70] T. Mouw and A.M. Verdery. "Network sampling with memory: a proposal for more efficient sampling from social networks". In: *Sociological Methodology* 42 (2012), pages 206–256 (*cited on pages 3, 22, 29, 40, 98*).

[71] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. "Random graph models of social networks". In: *Proceedings of the National Academy of Sciences* 99.suppl 1 (2002), pages 2566–2572. eprint: `https://www.pnas.org/content/99/suppl_1/2566.full.pdf` (*cited on pages 3, 8*).

[72] O. Riordan. "The phase transition in the configuration model". In: *Combinatorics, Probability and Computing* 21.1-2 (2012), pages 265–299 (*cited on pages 12, 98*).

[73] O. Robineau et al. "HIV transmission and pre-exposure prophylaxis in a high risk MSM population: A simulation study of location-based selection of sexual partners". In: *PLoS ONE* 12.11 (2017), e0189002 (*cited on page 40*).

[74] O. Robineau et al. "Model-based respondent driven sampling analysis for HIV prevalence in brazilian MSM". In: *Scientific Reports* 10 (2020), page 2646 (*cited on page 40*).

[75] D. A. Rolls et al. "Hepatitis C Transmission and Treatment in Contact Networks of People Who Inject Drugs". In: *PLOS ONE* 8.11 (Nov. 2013), pages 1–15 (*cited on page 40*).

[76] D. A. Rolls et al. "Modelling a disease-relevant contact network of people who inject drugs". In: *Social Networks* 35.4 (2013), pages 699–710 (*cited on pages 40, 98*).

[77] M. J. Salganik and D. D. Heckathorn. "Sampling and estimation in hidden populations using respondent-driven sampling". In: *Sociological methodology* 34.1 (2004), pages 193–240 (*cited on page 29*).

[78] A. Shaghaghi, R. S. Bhopal, and A. Sheikh. "Approaches to recruiting 'hard-to-reach' populations into research: a review of the literature". In: *Health promotion perspectives* 1.2 (2011), page 86 (*cited on page 72*).

[79] T. Tabouy, P. Barbillon, and J. Chiquet. "Variational inference for stochastic block models from sampled data". In: *Journal of the American Statistical Association* (2019), pages 1–23 (*cited on page 100*).

[80] E. Volz and D.D. Heckathorn. "Probability-based estimation theory for respondent-driven sampling". In: *Journal of Official Statistics* 24 (2008), pages 79–97 (*cited on pages 22, 29, 40, 98*).

[81] H. White, S. Boorman, and R. L. Breiger. "Social structure from multiple networks. I. Blockmodels of roles and positions". In: *American Journal of Sociology* 81.4 (1976), pages 730–780 (*cited on page 13*).

Books

[82] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, 1970 (*cited on pages 27, 75*).

[83] F. Ball et al. *Stochastic epidemic models with inference*. MathBiosciences. Springer, 2019 (*cited on pages 3, 40*).

[84] A. D. Barbour, Lars Holst, and Svante Janson. *Poisson approximation*. Volume 2. Oxford Studies in Probability. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1992, pages x+277 (*cited on pages 26, 77, 86*).
[85] P. Billingsley. *Convergence of Probability Measures*. New York: John Wiley and Sons, 1968 (*cited on pages 53, 83*).
[86] P. Billingsley. *Probability and Measure*. 3rd edition. New York: John Wiley and Sons, 1995 (*cited on pages 46, 68*).
[87] B. Bollobás. *Graph theory*. Volume 63. New York: Springer, 1979 (*cited on pages 4, 16*).
[88] B. Bollobás. *Modern Graph Theory*. Graduate texts in mathematics. New York: Springer, 1998 (*cited on page 4*).
[89] B. Bollobás. *Random graphs*. 2nd edition. Cambridge University Press, 2001 (*cited on page 42*).
[90] T. Britton and E. Pardoux. *Stochastic Epidemic models with Inference*. 2019 (*cited on page 75*).
[91] R. Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Graduate texts in mathematics. Springer, Aug. 2005 (*cited on page 4*).
[92] R. Durrett. *Random graph dynamics*. New York: Cambridge University Press, 2007 (*cited on pages 4, 8, 73*).
[93] S. N. Ethier and T. G. Kurtz. *Markov Processes, Characterization and Convergence*. New York: John Wiley and Sons, 1986 (*cited on pages 24, 26, 74, 77, 79*).
[94] R. Van der Hofstad. *Random Graphs and Complex Networks*. Volume 1. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2017 (*cited on pages 4, 8–13, 42, 73*).
[95] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014 (*cited on page 67*).
[96] T. Jaakkola. *Tutorial on variational approximation methods*. Advanced Mean Field Methods: Theory and Practice, Cambridge, MIT Press, 2000 (*cited on page 110*).
[97] C. McDiarmid. *London Mathematical Society Lecture Note Series*. Volume 141. Cambridge University Press, 1989, pages 148–188 (*cited on page 119*).
[98] M. Métivier. *Semimartingales: a course on stochastic processes*. Berlin, New-York: de Gruyter, 1982 (*cited on pages 52, 79, 83*).
[99] M. Newman. *Networks*. Oxford university press, 2018 (*cited on page 3*).
[100] R. Rebolledo. *La méthode des martingales appliquée à l'étude de la convergence en loi de processus*. Mémoire de la SMF. 1979 (*cited on page 64*).
[101] T. Łuczak S. Janson and A. Rucinski. *Theory of Random graphs*. New York-Chichester: John Wiley and Sons, Feb. 2000 (*cited on pages 8, 16*).

Preprints

[102] A. Cousien et al. "Responding Driven Sampling on sparse Erdös-Rényi graphs". In progress. 2020 (*cited on page 73*).

[103] J.F. Delmas, D. Dronnier, and P.A. Zitt. “An Infinite-Dimensional SIS Model”. working paper or preprint. June 2020 (*cited on page 3*).

[104] V. C. Tran and T. P. T. Vo. “Estimation of dense stochastic block models visited by random walks”. submitted to EJMS. 2020 (*cited on page 97*).

[105] T. P. T Vo. “Chain-referral sampling on Stochastic Block Models”. to appear in ESAIM:P&S. 2020 (*cited on page 71*).