Chain-referral sampling on stochastic block models

Thi Phuong Thuy Vo

ESAIM: PS 24 (2020) 718–738
https://doi.org/10.1051/ps/2020025

ESAIM: Probability and Statistics
www.esaim-ps.org

CHAIN-REFERRAL SAMPLING ON STOCHASTIC BLOCK MODELS*

Thi Phuong Thuy Vo**

Abstract. 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 $\pi = (\pi_1, \ldots, \pi_m)$ and a pattern for connection between blocks $P = (\lambda_{kl}/N)_{(k,l)\in\{1,\ldots,m\}^2}$. 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.

Mathematics Subject Classification. 05C80, 60J05, 60F17, 90B15, 92D30, 91D30.

Received June 9, 2019. Accepted September 18, 2020.

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

* This work was done during the Ph.D. thesis of the author under the supervision of Jean-Stéphane Dhersin and Tran Viet Chi. The author was partially supported by the Chaire MMB (Modélisation Mathématique et Biodiversité of Veolia-Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X) and by the ANR Econet (ANR-18-CE02-0010).

Keywords and phrases: Chain-referral sampling, random graph, social network, stochastic block model, exploration process, large graph limit, respondent driven sampling.

Vo Thi Phuong Thuy, Univ. Paris 13, CNRS, UMR 7539 - LAGA, 99 avenue J.-B. Clément, 93430 Villetaneuse, France.

** Corresponding author: vo@math.univ-paris13.fr

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 [25] or [16–18]). 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 [4]).

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.* [19]. 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 [14, 15, 23]. 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 1.1. Let N be a positive integer (number of vertices), m be a positive integer (number of blocks or types), $\pi = (\pi_1, \dots, \pi_m)$ 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 $P = (p_{kl})_{(k,l)\in\{1,\dots,m\}^2}$ 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: $[l] := \{v \in \{1, \dots, N\} : \Gamma_v = l\}$ with the size $N_l := |[l]|, l \in \{1, \dots, m\}$.

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 [10].

Here, we consider the Poisson case where the connectivity probabilities p_{kl} depend on N and are given by $p_{kl} = \lambda_{kl}/N$. This means that each individual of the block k contacts in average $\lambda_{kl}\pi_l$ 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 [11, 27]. 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 [18]). 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

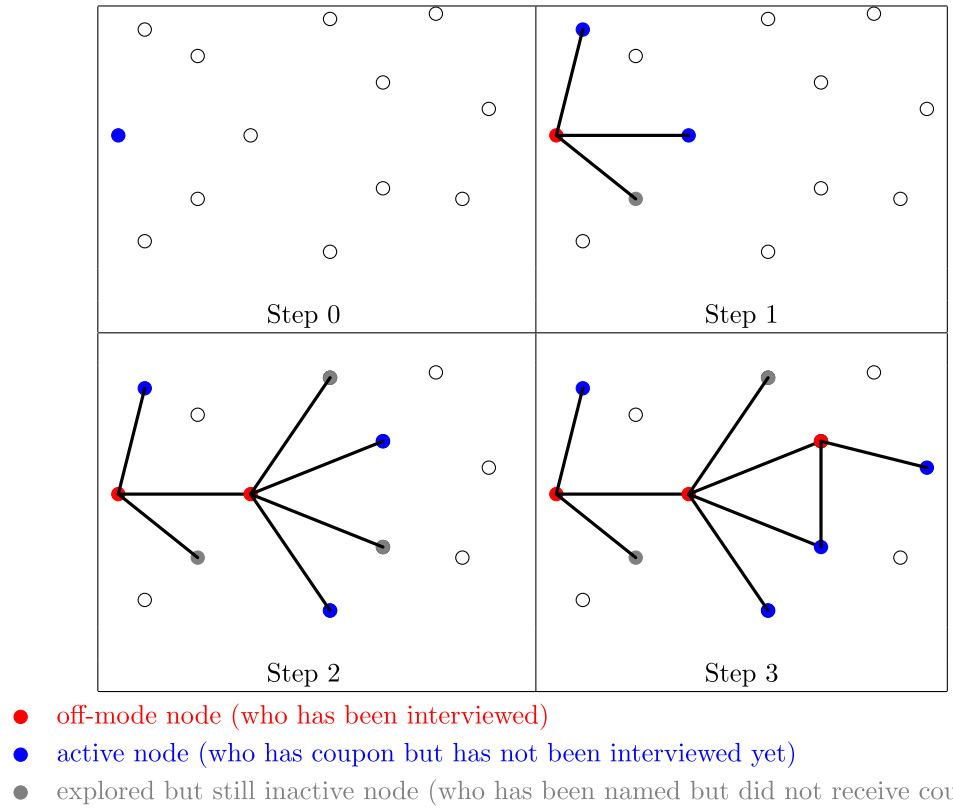


FIGURE 1. Description of how the chain-referral sampling works. In our model, the random network and the CRS are constructed simultaneously. For example, at Step 3, an edge between two vertices who are already known at Step 2 is revealed.

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. 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 $A_n^{(l)}$ (resp. $B_n^{(l)}$ and $u_n^{(l)}$) 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 $(X_n^{1,(l)}, X_n^{2,(l)}, X_n^{3,(l)}) = (A_n^{(l)}, B_n^{(l)}, U_n^{(l)})$.

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:

$$X_t^N := \frac{1}{N} X_{\lfloor Nt \rfloor} = \left(\frac{A_{\lfloor Nt \rfloor}}{N}, \frac{B_{\lfloor Nt \rfloor}}{N}, \frac{U_{\lfloor Nt \rfloor}}{N} \right) \in [0,1]^{3\times m}, \quad t \in [0,1]. \tag{1.1}$$

In all the paper, we consider spaces $\mathbb{R}^d$ equipped with the L^1-norm defined for $x = (x^1, \ldots, x^d)$ as $\|x\| = \sum_{k=1}^d |x^k|$. For all N, the process $X_\cdot^N$ lives in the space of càdlàg processes $\mathcal{D}([0,1], [0,1]^{3\times m})$ equipped with Skorokhod topology (see [13, 20, 22]).

There exist to our knowledge a few works of studying CRS form a probabilistic point of view, for example, Athreya and Röllin [3]. 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 process $(X_\cdot^N)_N$ converges to a system of ordinary differential equations (ODEs). There has also been literature on random walks exploring graphs possibly with different mechanism (see [7, 12] for instance). Here we allow the exploring Markov process to branch. Also, our process bares similarities with epidemics spreading on graphs (see [6, 9, 21, 26]) 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 [6]): 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 1.2. 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 1.3. Under the Assumption 1.2, from the proof of Theorem 3.2 of Barbour and Reinert [6], the early stages of the CRS is now can be associated approximated by a multitype branching process with the offspring distributions determined by the matrix μ. Thanks to the Assumption 1.2 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 nth generation as in the single branching process have been proved in Chapter 5 of [2]: 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 [6], 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 1.4. 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 1.5. *Under the Assumptions 1.2 and 1.4, we have: when N tends to infinity, the process $(X_\cdot^N)_N$ converges in distribution in $\mathcal{D}([0,1],[0,1]^{3\times m})$ to a deterministic vectorial function $x = (x_\cdot^{(l)})_{1\le l\le m} = (a_\cdot^{(l)}, b_\cdot^{(l)}, u_\cdot^{(l)})_{1\le l\le m}$ in $\mathcal{C}([0,1],[0,1]^{3\times m})$, which is the unique solution of the system of differential equations*

$$x_t = x_0 + \int_0^t f(x_s)\mathrm{d}s, \tag{1.2}$$

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

$$t_0 := \inf\{t \in [0,1] : \|a_t\| := a_t^{(1)} + \ldots + a_t^{(m)} = 0\}. \tag{1.3}$$

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

$$f_{1l}(x_s) = \sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\frac{\lambda_s^{k,l}}{\Lambda_s^k}\left(c - \sum_{h=0}^c (c-h)\frac{(\Lambda_s^k)^h}{h!}\mathrm{e}^{-\Lambda_s^k}\right) - \frac{a_s^{(l)}}{\|a_s\|}; \tag{1.4}$$

$$f_{2l}(x_s) = \sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\mu_s^{k,l} - \sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\frac{\lambda_s^{k,l}}{\Lambda_s^k}\left(c - \sum_{h=0}^c (c-h)\frac{(\Lambda_s^k)^h}{h!}\mathrm{e}^{-\Lambda_s^k}\right); \tag{1.5}$$

$$f_{3l}(x_s) = \frac{a_s^{(l)}}{\|a_s\|}; \tag{1.6}$$

with

$$\lambda_s^{k,l} := \lambda_{kl}\left(\pi_l - a_s^{(l)} - u_s^{(l)}\right); \quad \Lambda_s^k := \sum_{l=1}^m \lambda_s^{k,l} \quad \text{and} \quad \mu_s^{k,l} := \lambda_{kl}(\pi_l - a_s^{(l)} - b_s^{(l)} - u_s^{(l)}). \tag{1.7}$$

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

Remark 1.6. 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, $t_0 \le 1$. 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 (1.5) 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 [13]) and Poisson approximations (see [5]). 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. Definition of the chain-referral process

Let us describe the dynamics of $X = (X_n)_{n\in\mathbb{N}}$. Recall that $\|A_n\| := \sum_{l=1}^m A_n^{(l)}$ 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 $\mathcal{M}(\lfloor\|a_0\|N\rfloor;\pi_1,\dots,\pi_m)$, *i.e.* $\mathbb{P}\big((A_0^{(1)},\dots,A_0^{(m)}) = (k_1,\dots,k_m)\big) = \pi_1^{k_1}\dots\pi_m^{k_m}, k_i \in \mathbb{N}$ such that $\sum_{i=1}^m k_i = \lfloor\|a_0\|N\rfloor$ and Assumption 1.4 is satisfied. Also $B_0 = U_0 = (0,\dots,0)$ 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 nth-interview and given the number $N_1,\dots,N_m$ of nodes of each type. At step $n \geq 1$, after the nth-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 $I_n := (I_n^{(1)},\dots,I_n^{(m)})$, which takes value 1 at coordinate l and 0 elsewhere if the nth interviewee belongs to block l. This nth-interviewee is chosen uniformly among the $\|A_{n-1}\|$ active coupons of m types *i.e.* I_n has multinomial distribution

$$I_n = (I_n^{(1)},\dots,I_n^{(m)}) \overset{(d)}{=} \mathcal{M}\left(1; \frac{A_{n-1}^{(1)}}{\|A_{n-1}\|},\dots,\frac{A_{n-1}^{(m)}}{\|A_{n-1}\|}\right). \tag{2.1}$$

If the chosen one belongs to block $[l]$, $A_n^{(l)}$ 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.* $U_n^{(l)} = U_{n-1}^{(l)} + I_n^{(l)}$. Among the new contacts of the nth$-$interviewee, define $H_n^{(l)}$ the number of new contacts of type l, who have not been mentioned before; $K_n^{(l)}$ the number of new contacts of type l whose identities are already known but who are still inactive. The $H_n^{(l)}$ new connections are chosen independently among $N_l - A_{n-1}^{(l)} - B_{n-1}^{(l)} - U_n^{(l)}$ individuals in the hidden population where probability of each successful connection is $\sum_{k=1}^m I_n^{(k)} p_{kl}$. Hence, conditioning on $(N_1,\dots,N_m), X_{n-1}$, the random variable $H_n^{(l)}$ follows the binomial distribution:

$$H_n^{(l)} \overset{(d)}{=} \mathcal{B}in\left(N_l - A_{n-1}^{(l)} - B_{n-1}^{(l)} - U_n^{(l)}, \sum_{k=1}^m I_n^{(k)} p_{kl}\right). \tag{2.2}$$

And the $K_n^{(l)}$ individuals are chosen independently of $H_n^{(l)}$ from $B_{n-1}^{(l)}$ individuals and independently of the others with probability $\sum_{k=1}^m I_n^{(k)} p_{kl}$. In that way, conditioning on $(N_1,\dots,N_m), X_{n-1}$, $K_n^{(l)}$ also has the binomial distribution:

$$K_n^{(l)} \overset{(d)}{=} \mathcal{B}in\left(B_{n-1}^{(l)}, \sum_{k=1}^m I_n^{(k)} p_{kl}\right). \tag{2.3}$$

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

$$Z_n^{(l)} \overset{(d)}{=} \mathcal{B}in\left(N_l - A_{n-1}^{(l)} - U_n^{(l)}, \sum_{k=1}^m I_n^{(k)} p_{kl}\right). \tag{2.4}$$

Let $C_n = (C_n^{(1)},\dots,C_n^{(m)})$ $(l = 1,\dots,m)$ 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 $C_n'^{(l)} = (C_n'^{(1)},\dots,C_n'^{(m)})$ having the multivariate hypergeometric distribution

with parameters $(m; c, (Z_n^{(1)}, \dots, Z_n^{(m)}))$ and the support $\{(c_1, \dots, c_m) \in \mathbb{N}^m : \forall l \leq m, c_l \leq Z_n^{(l)}, \sum_{l=1}^{m} c_i = c\}$, that is

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

In another words,

$$C_n^{(l)} := \begin{cases} Z_n^{(l)} & \text{if} \quad \sum_{l=1}^{m} Z_n^{(l)} \leq c \\ C_n'^{(l)} & \text{otherwise} \end{cases}. \tag{2.5}$$

Let define by

$$n_0 := \inf\{n \in \{1, \dots, N\}, A_n = 0\} \tag{2.6}$$

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

$$\begin{aligned} &\begin{cases} A_n &= A_{n-1} - I_n + C_n \\ B_n &= B_{n-1} + H_n - C_n \\ U_n &= \sum_{i=1}^{n} I_i \end{cases}, \quad \text{for} \quad n \in \{1, \dots, n_0\} \\ \text{and} \quad &X_n = X_{n-1} \quad \text{when} \quad n > n_0. \end{aligned} \tag{2.7}$$

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 (2.7). 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 (2.7) 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 $H_n^{(l)}$ in (2.2) depends on time as $U_n^{(l)} = n$. □

3. Asymptotic behavior of the chain-referral process

Let us now consider the renormalized chain-referral process given in (1.1) in the time interval $[0, t_0]$. The main theorem (Thm. 1.5) shows the convergence of the sequence $(X_\cdot^N)_N$ to a deterministic process. For this, we look for an expression of the equations (2.7) 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 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.* [13, 24]) to show the tightness of the distributions of these processes in Section 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 $(X_\cdot^N)_N$. This proves Theorem 1.5.

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

3.1. Doob's decomposition

Lemma 3.1. *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*

$$\Delta_t^N = \begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \\ \Delta_t^{N,3} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} \begin{pmatrix} \mathbb{E}[C_n - I_n|\mathcal{F}_{n-1}] \\ \mathbb{E}[H_n - C_n|\mathcal{F}_{n-1}] \\ \mathbb{E}[I_n|\mathcal{F}_{n-1}] \end{pmatrix}; \tag{3.1}$$

$(M_t^N)_{t\in[0,1]}$ 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 $(l,k)\in\{1,\dots,m\}^2$,

$$\langle M_\cdot^{(l),N}, M_\cdot^{(k),N}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt \rfloor} \mathbb{E}\left[\left(X_n^{(l)} - \mathbb{E}[X_n^{(l)}|\mathcal{F}_{n-1}]\right)\left(X_n^{(k)} - \mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\right)^T \middle| \mathcal{F}_{n-1}\right], \quad t\in[0,1] \tag{3.2}$$

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 $(\Delta_\cdot^N)_N$ 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 $\lambda := \max_{l,k\in\{1,\dots,m\}} \lambda_{kl}$. Notice that

$$\|\mathbb{E}[A_n - A_{n-1}|\mathcal{F}_{n-1}\| = \|\mathbb{E}[C_n - I_n|\mathcal{F}_{n-1}\| \le c, \tag{3.3}$$

$$\|\mathbb{E}[B_n - B_{n-1}|\mathcal{F}_{n-1}\| = \|\mathbb{E}[H_n - C_n|\mathcal{F}_{n-1}\| \le m(\max_{l,k\in\{1,\dots,m\}} \lambda_{kl}) + c = m\lambda + c, \tag{3.4}$$

$$\|\mathbb{E}[U_n - U_{n-1}|\mathcal{F}_{n-1}\| \le 1, \tag{3.5}$$

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

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

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 $k, l = 1, \ldots, m$

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

The term $L_t'^N$ is an $\mathcal{F}_t^N-$martingale since whenever $n' < n$, $\left(X_{n'}^{(k)} - X_{n'-1}^{(k)} - \mathbb{E}[X_{n'}^{(k)} - X_{n'-1}^{(k)}|\mathcal{F}_{n'-1}]\right)$ is $\mathcal{F}_{n-1}-$measurable. To see that the quadratic variation of M_t^N has the form (3.2), we write the term L_t^N as follows:

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

As a result,

$$M_t^{(l),N}\left(M_t^{(k),N}\right)^T = \langle M^N\rangle_t + L_t'^N + L_t''^N. \tag{3.6}$$

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

$$V^N(\langle M_\cdot^N\rangle_t) = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k,l=1}^{m}\left\|\mathbb{E}\left[\left(X_n^{(l)} - \mathbb{E}[X_n^{(l)}|\mathcal{F}_{n-1}]\right)\left(X_n^{(k)} - \mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\right)^T\Big|\mathcal{F}_{n-1}\right]\right\|. \tag{3.7}$$

The integrand in the right hand side is the conditional covariance between $X_n^{(l)}$ and $X_n^{(k)}$ conditionally to $\mathcal{F}_{n-1}$. Because $X_n^{(l)}$ 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:

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

by the Cauchy-Schwarz inequality. Thus:

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

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

$$\begin{aligned}
&\sum_{i,j=1}^{3}\left(\mathrm{Var}(X_n^{i,(l)} - X_{n-1}^{i,(l)}|\mathcal{F}_{n-1})\right)^{1/2}\left(\mathrm{Var}(X_n^{j,(k)} - X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2}\\
&\quad= \left(\sum_{i=1}^{3}\left(\mathrm{Var}(X_n^{i,(l)} - X_{n-1}^{i,(l)}|\mathcal{F}_{n-1})\right)^{1/2}\right)\left(\sum_{j=1}^{3}\left(\mathrm{Var}(X_n^{j,(k)} - X_{n-1}^{j,(k)}|\mathcal{F}_{n-1})\right)^{1/2}\right)\\
&\quad\leq \frac{3}{2}\sum_{i=1}^{3}\left(\mathrm{Var}(X_n^{i,(l)} - X_{n-1}^{i,(l)}|\mathcal{F}_{n-1}) + \mathrm{Var}(X_n^{i,(k)} - X_{n-1}^{i,(k)}|\mathcal{F}_{n-1})\right). \qquad (3.8)
\end{aligned}$$

From (3.3)–(3.5) and by Cauchy-Schwarz's inequality, we obtain the following inequalities

$$\mathrm{Var}(C_n^{(l)} - I_n^{(l)}|\mathcal{F}_{n-1}) \leq c^2, \quad \mathrm{Var}(H_n^{(l)} - C_n^{(l)}|\mathcal{F}_{n-1}) \leq 2(\max_{l,k\in\{1,\dots,m\}}\lambda_{lk}^2 + c^2), \quad \mathrm{Var}(I_n^{(l)}|\mathcal{F}_{n-1}) \leq 1. \qquad (3.9)$$

As a consequence,

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

Thus, the proof of the lemma is completed. □

3.2. Tightness of the renormalized process

Lemma 3.2. *The sequence of processes $(X_\cdot^N)_N$ is tight in the Skorokhod space $\mathcal{D}([0,1],[0,1]^{3\times m})$.*

Proof. To prove the tightness of $(X_\cdot^N)_N$, we use the criteria of tightness for semi-martingales in ([24], Thm. 2.3.2 (Rebolledo)): first, we verify the marginal tightness of each sequence $(X_t^N)_N$ 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 $(\Delta^N)_N$ and the quadratic variation of the martingale $(M^N)_N$. 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 $(M^N)_N$ is proved to be convergent (to zero) in $\mathbb{L}^2$ as $N \to \infty$ (which is done by Prop. 3.3), we have the tightness of $(M^N)_N$. Thus, it is sufficient to check the tightness condition for the modulus of continuity of $(\Delta^N)_N$ (see, *e.g.*, [8], Thm. 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

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

By (3.3)–(3.5), 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 3.2, we now prove that:

Proposition 3.3. *The sequence of martingale $(M_t^N, t \in [0,1])_N$ converges locally uniformly in t to 0 in $\mathbb{L}^2$, as N goes to infinity.*

Proof. Consider the quadratic variation of $(M_\cdot^N)_N$: According to the formula (3.2), we apply the Cauchy–Schwarz's inequality and then use the inequality (3.8) to obtain that for every $t \in [0,1]$,

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

where $(X_n^{1,(l)}, X_n^{2,(l)}, X_n^{3,(l)}) = (A_n^{(l)}, B_n^{(l)}, U_n^{(l)})$. From (3.3)–(3.5) and (3.9), we deduce that

$$\|\langle M_\cdot^N \rangle_t\| \leq \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \frac{3m^2}{2} \left(c^2 + 2(\max_{l,k \in \{1,\dots,m\}} \lambda_{lk}^2 + c^2) + 1 \right) \leq \frac{1}{N} \frac{3m^2}{2} (3c^2 + 2\lambda^2 + 1)t. \tag{3.10}$$

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 Proposition 3.3 and hence of Lemma 3.2. □

3.3. Identify the limiting value

Since the sequence $(X_\cdot^N)_N$ is tight, for any limiting value $x = (a, b, u)$ of the sequence $(X^N)_N$, there exists an increasing sequence $(\varphi_N)_N$ in $\mathbb{N}$ such that $(X_\cdot^{\varphi_N})_N$ 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 $(X_\cdot^{\varphi_N})_N$ as $(X_\cdot^N)_N = (A_\cdot^N, B_\cdot^N, U_\cdot^N)_N$.

We consider separately the martingale and finite variation parts. Proposition 3.3 implies that the sequence martingale $(M_\cdot^N)_N$ converges to 0 in distribution and hence $(M^N)_N$ converges to zero in probability. It remains to find the limit of the finite variation process $(\Delta_\cdot^N)_N$ given in equation (3.1) 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 (1.5)) for every convergent subsequence extracted from the tight sequence $(X^N)_N$.

Proposition 3.4. *When N goes to infinity, we have the following convergences in distribution in $\mathcal{D}([0,1],[0,1]^{3\times m})$:*

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[C_n^{(l)}|\mathcal{F}_{n-1}] \xrightarrow{(d)} \int_0^t \left\{\sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\frac{\lambda_s^{k,l}}{\Lambda_s^k}\left(c-\sum_{h=0}^c (c-h)\frac{(\Lambda_s^k)^h}{h!}\mathrm{e}^{-\Lambda_s^k}\right)\right\}\mathrm{d}s, \tag{3.11}$$

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[H_n^{(l)}|\mathcal{F}_{n-1}] \xrightarrow{(d)} \int_0^t \sum_{k=1}^m \frac{a_s^{(k)}}{\|a_s\|}\mu_s^{k,l}\mathrm{d}s, \tag{3.12}$$

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[I_n^{(l)}|\mathcal{F}_{n-1}] = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left(\frac{A_{n-1}^{(l)}}{N}\right)\Big/\left(\frac{\|A_{n-1}\|}{N}\right) \xrightarrow{(d)} \int_0^t \frac{a_s^{(l)}}{\|a_s\|}\mathrm{d}s, \tag{3.13}$$

where $\lambda_s^{k,l}, \Lambda_s^k, \mu_s^{k,l}$ are defined as in Theorem 1.5. This provides the convergence of $(\Delta_\cdot^N)_N$ to a solution $x_\cdot$ of (1.2).

Since the limits are deterministic, the convergences hold in probability. Moreover the uniqueness of the solution of (1.2) will be proved later, which will imply the convergence of the whole sequence $(X_\cdot^N)_N$ to this solution.

Proof. Recall that since the sequence $(X_\cdot^N)_N$ is tight, we have extracted a converging subsequence also denoted by $(X_\cdot^N)_N$ of which we study the limit.

The proof of the Proposition 3.4 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 $l \in \{0,\dots,m\}$,

$$\left|\mathbb{E}[C_n^{(l)}|\mathcal{F}_{n-1}] - \frac{\lambda_n^{(l)}}{\Lambda_n}\left(c-\sum_{h=0}^c (c-h)\frac{(\Lambda_n)^h}{h!}\mathrm{e}^{-\Lambda_n}\right)\right| \le \frac{m(c+1)\lambda}{N}, \tag{3.14}$$

where

$$\lambda_n^{(l)} := \left(\sum_{k=1}^m I_n^{(k)}\lambda_{kl}\right)\left(\frac{N_l}{N} - \frac{A_{n-1}^{(l)}}{N} - \frac{U_n^{(l)}}{N}\right) \qquad \text{and} \qquad \Lambda_n := \sum_{j=1}^m \lambda_n^{(j)}. \tag{3.15}$$

Notice that $\Lambda_n = 0$ only if for each $l \in \{1,\dots,m\}$, $\lambda_n^{(l)} = 0$. It happens when $A_{n-1}^{(l)} + U_n^{(l)} = N_l$, meaning that all the nodes of type l have been discovered. In this case, $C_n^{(l)} = 0$ and (3.14) is satisfied.

Let us write

$$\mathbb{E}[C_n^{(l)}|\mathcal{F}_{n-1}] = \mathbb{E}\left[Z_n^{(l)}\mathbb{1}_{\sum_{j=1}^m Z_n^{(j)}\le c}\Big|\mathcal{F}_{n-1}\right] + \mathbb{E}\left[\frac{cZ_n^{(l)}}{\sum_{j=1}^m Z_n^{(j)}}\mathbb{1}_{\sum_{j=1}^m Z_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right]. \tag{3.16}$$

For every $l = 1,\dots,m$ and every fixed n, when all the parameters are positive, we have that $(N_l - A_{n-1}^{(l)} - U_n^{(l)}) \xrightarrow[\text{a.s.}]{N\to\infty} +\infty$. Then we work conditionally on $N_l, A_{n-1}^{(l)}, U_n^{(l)}$ and $I_n^{(l)}$ and use the Poisson approximation (*e.g.* see Eq. (1.23) and Thm. 2.A, 2.B by Barbour, Holst and Janson in [5]) for the approximation: the binomial random variable $Z_n^{(l)}$ may be approximated by a Poisson random variable $\tilde{Z}_n^{(l)} \overset{(d)}{=} \mathcal{P}\big((\sum_{k=1}^m I_n^{(k)}\lambda_{kl})(\frac{N_l}{N} - \frac{A_{n-1}^{(l)}}{N} - \frac{U_n^{(l)}}{N})\big)$ such that

$$d_{\mathrm{TV}}(Z_n^{(l)}, \tilde{Z}_n^{(l)}) \le \frac{2}{(N_l - A_n^{(l)} - U_n^{(l)})\left(\frac{\sum_{k=1}^m I_n^{(k)}\lambda_{kl}}{N}\right)} \sum_{i=1}^{N_l - A_n^{(l)} - U_n^{(l)}} \left(\frac{\sum_{k=1}^m I_n^{(k)}\lambda_{kl}}{N}\right)^2 \le \frac{2\max_{k,l}\lambda_{kl}}{N} = \frac{2\lambda}{N}.$$

As a consequence, the first term in the right hand side of (3.16) can be approximated as

$$\left|\mathbb{E}\left[Z_n^{(l)}\mathbb{1}_{\sum_{j=1}^m Z_n^{(j)}\le c}\Big|\mathcal{F}_{n-1}\right] - \mathbb{E}\left[\tilde{Z}_n^{(l)}\mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}\le c}\Big|\mathcal{F}_{n-1}\right]\right| \le \frac{2mc\lambda}{N}, \tag{3.17}$$

and

$$\left|\mathbb{E}\left[\frac{Z_n^{(l)}}{\sum_{j=1}^m Z_n^{(j)}}\mathbb{1}_{\sum_{j=1}^m Z_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right] - \mathbb{E}\left[\frac{\tilde{Z}_n^{(l)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}}\mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right]\right| \le \frac{2m\lambda}{N}. \tag{3.18}$$

It follows that we need to deal with the Poisson random variables $\tilde{Z}_n^{(l)} (l \in \{1,\dots,m\})$. 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 $\sum_{j\neq l}\tilde{Z}_n^{(j)} =: \hat{Z}_n^{(l)}$ has a Poisson distribution with parameter $\hat{\lambda}_n^{(l)} := \sum_{j\neq l}\lambda_n^{(j)}$. And hence,

$$\begin{aligned}\mathbb{E}\left[\tilde{Z}_n^{(l)}\mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}\le c}\Big|\mathcal{F}_{n-1}\right] &= \sum_{h=1}^{c}\sum_{h_1=1}^{h} h_1 \frac{(\lambda_n^{(l)})^{h_1}(\hat{\lambda}_n^{(l)})^{h-h_1}}{h_1!(h-h_1)!}\mathrm{e}^{-\Lambda_n} \\ &= \lambda_n^{(l)}\sum_{h=1}^{c}\frac{(\Lambda_n)^{h-1}}{(h-1)!}\mathrm{e}^{-\Lambda_n} = \lambda_n^{(l)}\sum_{h=0}^{c}\frac{h}{\Lambda_n}\frac{(\Lambda_n)^h}{h!}\mathrm{e}^{-\Lambda_n}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\frac{\tilde{Z}_n^{(l)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}}\mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)}>c}\Big|\mathcal{F}_{n-1}\right] &= \sum_{h=c+1}^{\infty}\sum_{k=0}^{h}\frac{k}{h}\frac{(\lambda_n^{(l)})^k}{k!}\frac{(\hat{\lambda}_n^{(l)})^{h-k}}{(h-k)!}\mathrm{e}^{-\lambda_n^{(l)}}\mathrm{e}^{-\hat{\lambda}_n^{(l)}} \\ &= \lambda_n^{(l)}\sum_{h=c+1}^{\infty}\sum_{k=0}^{h-1}\frac{1}{h}\frac{(\lambda_n^{(l)})^k}{k!}\frac{(\hat{\lambda}_n^{(l)})^{h-1-k}}{(h-1-k)!}\mathrm{e}^{-\lambda_n^{(l)}}\mathrm{e}^{-\hat{\lambda}_n^{(l)}}\end{aligned}$$

$$= \lambda_n^{(l)} \sum_{h=c+1}^{\infty} \frac{1}{h} \frac{(\lambda_n^{(l)} + \hat{\lambda}_n^{(l)})^{h-1}}{(h-1)!} \mathrm{e}^{-(\lambda_n^{(l)} + \hat{\lambda}_n^{(l)})} = \frac{\lambda_n^{(l)}}{\Lambda_n} \sum_{h=c+1}^{\infty} \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n}$$
$$= \frac{\lambda_n^{(l)}}{\Lambda_n} (1 - \sum_{h=0}^{c} \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n}). \tag{3.19}$$

Using (3.16), we obtain:

$$\mathbb{E}[C_n^{(l)} | \mathcal{F}_{n-1}] = \mathbb{E}\left[\tilde{Z}_n^{(l)} \mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)} \le c} + \frac{\tilde{Z}_n^{(l)}}{\sum_{j=1}^m \tilde{Z}_n^{(j)}} \mathbb{1}_{\sum_{j=1}^m \tilde{Z}_n^{(j)} > c} \middle| \mathcal{F}_{n-1} \right] = \frac{\lambda_n^{(l)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n} \right),$$

which finishes Step 1.

Step 2: We decompose the second term in the left hand side of (3.14) as follows:

$$\frac{\lambda_n^{(l)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n} \right) = \alpha_n^{(l)} + \xi_n^{(l)}, \quad l = 1, \dots, m. \tag{3.20}$$

with

$$\alpha_n^{(l)} := \mathbb{E}\left[\frac{\lambda_n^{(l)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n} \right) \middle| \mathcal{F}_{n-1} \right]$$
$$\xi_n^{(l)} := \frac{\lambda_n^{(l)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n} \right) - \mathbb{E}\left[\frac{\lambda_n^{(l)}}{\Lambda_n} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n)^h}{h!} \mathrm{e}^{-\Lambda_n} \right) \middle| \mathcal{F}_{n-1} \right].$$

By writing

$$\alpha_n^{(l)} = \sum_{k=1}^{m} \mathbb{P}(I_n^{(k)} = 1) \frac{\lambda_n^{k,l}}{\Lambda_n^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n^k)^h}{h!} \mathrm{e}^{-\Lambda_n^k} \right),$$

where

$$\lambda_n^{k,l} := \lambda_{kl} \left(\frac{N_l}{N} - \frac{A_{n-1}^{(l)}}{N} - \frac{U_{n-1}^{(l)}}{N} - \frac{\mathbb{1}_{\{k=l\}}}{N} \right) \quad \text{and} \quad \Lambda_n^k := \sum_{j=1}^{m} \lambda_n^{k,j} \quad (l = 1, \dots, m), \tag{3.21}$$

we obtain that for every $t \in [0,1]$,

$$\frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \alpha_n^{(l)} = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \left\{ \sum_{k=1}^{m} \frac{A_{n-1}^{(k)}}{|A_{n-1}|} \frac{\lambda_n^{k,l}}{\Lambda_n^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_n^k)^h}{h!} \mathrm{e}^{-\Lambda_n^k} \right) \right\}. \tag{3.22}$$

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

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\xi_n\overset{(\mathbb{L}^2)}{\to}0 \tag{3.23}$$

uniformly in $t\in[0,1]$. Together with the points given in (3.14), (3.20) and (3.23), take the limit as $N\to\infty$ in the right hand side of (3.22), we obtain the right hand side of (3.11).

Step 3: We use similar arguments as in Step 2 to obtain the limit in right hand side of (3.12). Denote by

$$\mu_n^{(l)}:=\left(\sum_{k=1}^m I_n^{(k)}\lambda_{kl}\right)\left(\frac{N_l}{N}-\frac{A_{n-1}^{(l)}}{N}-\frac{B_{n-1}^{(l)}}{N}-\frac{U_n^{(l)}}{N}\right).$$

Recall from (2.2) that conditioning on $\mathcal{F}_{n-1}$, $H_n^{(l)}\overset{(d)}{=}\mathcal{B}in\left(N_l-A_{n-1}^{(l)}-B_{n-1}^{(l)}-U_n^{(l)},\frac{\sum_{k=1}^m I_n^{(k)}\lambda_{kl}}{N}\right)$, then

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[H_n^{(l)}|\mathcal{F}_{n-1}]=\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mu_n^{(l)}. \tag{3.24}$$

We write

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mu_n^{(l)}=\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}(\beta_n^{(l)}+\zeta_n^{(l)}) \tag{3.25}$$

with

$$\beta_n^{(l)}:=\mathbb{E}\left[\left(\sum_{k=1}^m I_n^{(k)}\lambda_{kl}\right)\left(\frac{N_l}{N}-\frac{A_{n-1}^{(l)}}{N}-\frac{B_{n-1}^{(l)}}{N}-\frac{U_n^{(l)}}{N}\right)\Big|\mathcal{F}_{n-1}\right];$$

$$\zeta_n^{(l)}:=\left(\sum_{k=1}^m I_n^{(k)}\lambda_{kl}\right)\left(\frac{N_l}{N}-\frac{A_{n-1}^{(l)}}{N}-\frac{B_{n-1}^{(l)}}{N}-\frac{U_n^{(l)}}{N}\right)-\mathbb{E}\left[\left(\sum_{k=1}^m I_n^{(k)}\lambda_{kl}\right)\left(\frac{N_l}{N}-\frac{A_{n-1}^{(l)}}{N}-\frac{B_{n-1}^{(l)}}{N}-\frac{U_n^{(l)}}{N}\right)\Big|\mathcal{F}_{n-1}\right].$$

Using a similar argument as in Step 2, we have

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\beta_n^{(l)}&=\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m\mathbb{P}(I_n^{(k)}=1)\lambda_{kl}\left(\frac{N_l}{N}-\frac{A_{n-1}^{(l)}}{N}-\frac{B_{n-1}^{(l)}}{N}-\frac{U_{n-1}^{(l)}}{N}-\frac{\mathbb{1}_{\{k\neq l\}}}{N}\right)\\&=\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m\frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,l}-\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^m\frac{A_{n-1}^{N,k}}{\|A_{n-1}^N\|}\lambda_{kl}\frac{\mathbb{1}_{\{k\neq l\}}}{N},\end{aligned}$$

with $\mu_n^{k,l} := \lambda_{kl}\left(\frac{N_l}{N} - \frac{A_{n-1}^{(l)}}{N} - \frac{B_{n-1}^{(l)}}{N} - \frac{U_{n-1}^{(l)}}{N}\right)$. Then,

$$\left|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\left(\beta_n^{(l)} - \sum_{k=1}^{m}\frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,l}\right)\right| \le \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^{m}\frac{A_{n-1}^{N,k}}{\|A_{n-1}^N\|}\lambda_{kl}\frac{\mathbb{1}_{\{k\neq l\}}}{N} \le \frac{\lambda}{N}. \tag{3.26}$$

Take the limit as $N \to +\infty$, we have that

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=1}^{m}\frac{A_{n-1}^{(k)}}{\|A_{n-1}\|}\mu_n^{k,l} = \int_0^t\sum_{k=1}^{m}\frac{a_s^{(k)}}{\|a_s\|}\mu_s^{k,l}\mathrm{d}s.$$

Further, the $\mathcal{F}_t^N-$martingale $\frac{1}{N}\sum_{n=1}^{\lfloor N\cdot\rfloor}\zeta_n^{(l)}$ converges in $\mathbb{L}^2$ to 0 uniformly in $t\in[0,1]$. Thus, (3.12) is proved.

For the proof of (3.13), by the definition of I_n as in (2.1), we have

$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[I_n^{(l)}|\mathcal{F}_{n-1}] = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\frac{A_{n-1}^{(l)}}{\|A_{n-1}\|} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\frac{A_{n-1}^{(l)}/N}{\|A_{n-1}\|/N}.$$

Take the limit as $N\to+\infty$, we obtain the limit in the right hand side of (3.13).

The preceding steps allow to conclude the proof of Proposition 3.4. □

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 (1.2). If it is the case, then the process $(X^N)_N$ admits a unique limiting value and thus converges to x.

Assume that there exist two solutions x^1 and x^2 to ODEs (1.2) on the interval $[0,t_0']$, where

$$t_0' = \inf\{t\in[0,1] : a_{t_0'}^1 = 0 \text{ or } a_{t_0'}^2 = 0\}.$$

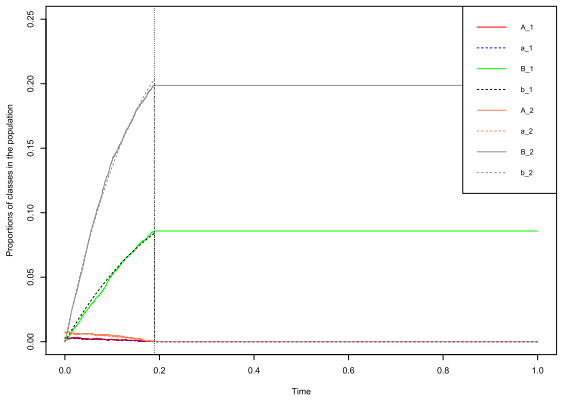
Then using the intermediate value theorem, there exist $\xi_{ij}(s)\in[x_{ij}^1(s),x_{ij}^2(s)]$ such that

$$\begin{aligned}\|x_t^1 - x_t^2\| = \left\|\int_0^t (f(x_s^1)-f(x_s^2))\mathrm{d}s\right\| &\le \int_0^t\sum_{i=1}^{3}\sum_{j=1}^{m}\left|\frac{\partial f}{\partial x_{ij}}(\xi_{ij}(s))\right|\left|x_{ij}^1(s)-x_{ij}^2(s)\right|\mathrm{d}s\\ &\le \int_0^t L(s)\|x_s^1 - x_s^2\|\mathrm{d}s,\end{aligned}$$

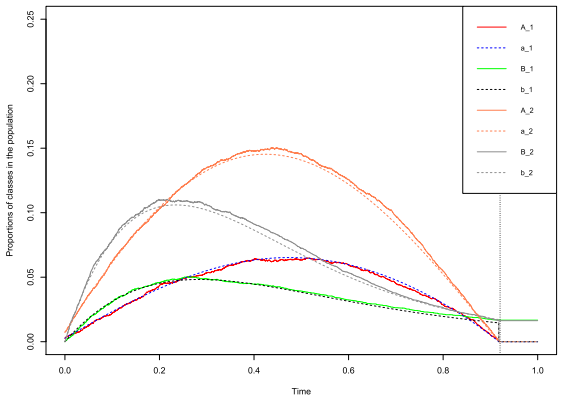
where $x_s^k = (x_{ij}(s))_{\substack{1\le i\le 3\\ 1\le j\le m}}$ $(k\in\{1,2\})$ and $L(s) = \sum_{i=1}^{3}\sum_{j=1}^{m}\max\left|\frac{\partial f}{\partial x_{ij}}(x_s)\right|$, of which the maximum is over $x(s) = (x_{ij}(s))_{ij}\in[0,1]^{3m}$ such that $\forall i,j : x_{ij}\in[x_{ij}^1,x_{ij}^2]$, 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

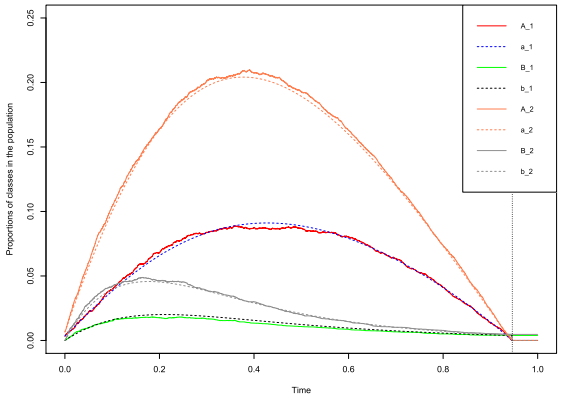
$$\|x_t^1 - x_t^2\| \le \|x_0^1 - x_0^2\|\exp\left(\int_0^t L(s)\mathrm{d}s\right) = 0.$$



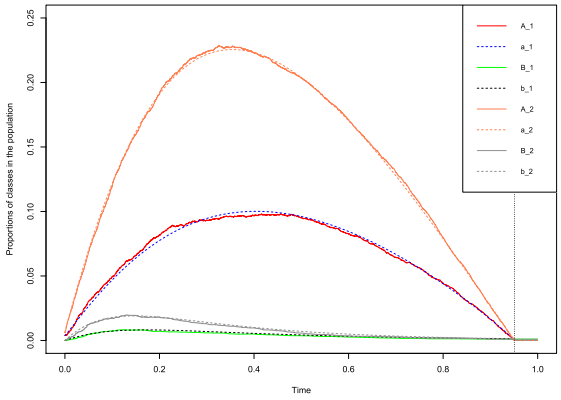
(A) $c = 1$



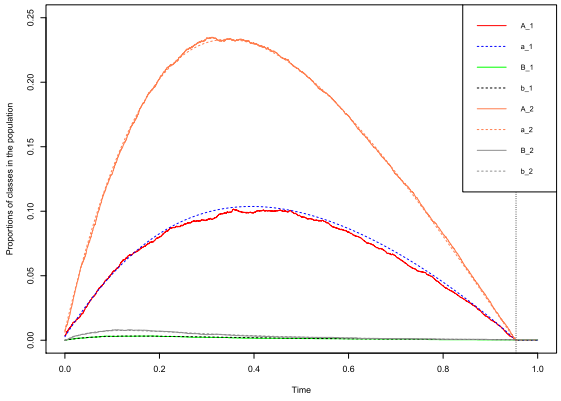
(B) $c = 2$



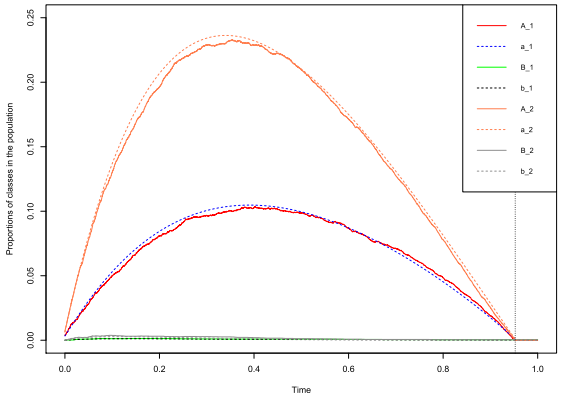
(C) $c = 3$



(D) $c = 4$



(E) $c = 5$



(F) $c = 6$

FIGURE 2. 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 $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$.

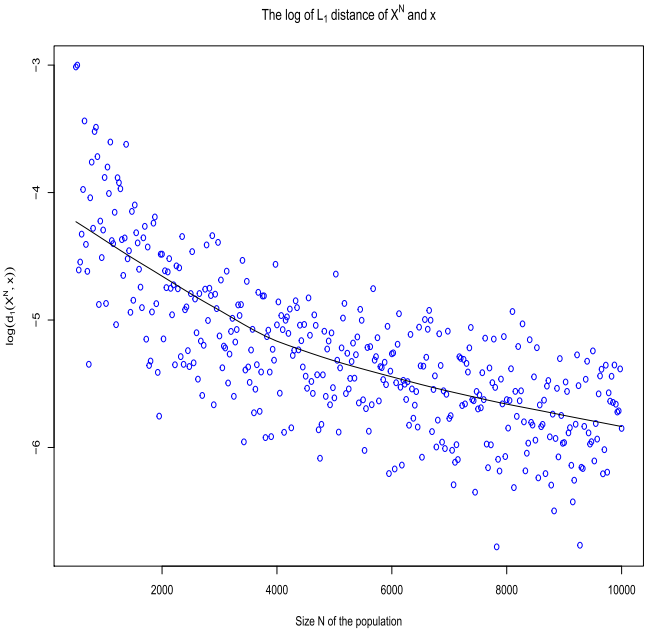


FIGURE 3. Scatter plot of $\ln d_1(X^N, x)$ along with the smoothing line suggesting the linear relationship between $\ln d_1(X^N, x)$ and N. The plot is done for the case $c = 3$, the number of initial individuals are 1% of the population and the size N varies from 500 to 10000. All other parameters are fixed: $\pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$.

This shows that $x_t^1 \equiv x_t^2$ for all $t \in [0, t_0']$. It also follows $t_0' = t_0$.

4. Simulation

The simulations show that the deterministic solution of the system of ODEs (1.2) fits well with our stochastic process, see Figure 2. The sequence of stochastic process $(X_\cdot^N)_N$ 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 $(X_\cdot^N)_N$ is asymptotically a deterministic function, which is a solution of a system of (1.2). 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:

$$\int_0^1 (\|A_t^N - a_t\| + \|B_t^N - b_t\| + \|U_t^N - u_t\|)\mathrm{d}t,$$

Figure 3. The speed of convergence has been studied in the case of Erdös-Rényi graphs in the Ph.D. thesis, by establishing a central limit theorem [28].

By studying the solution of (1.2), 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, 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 1.

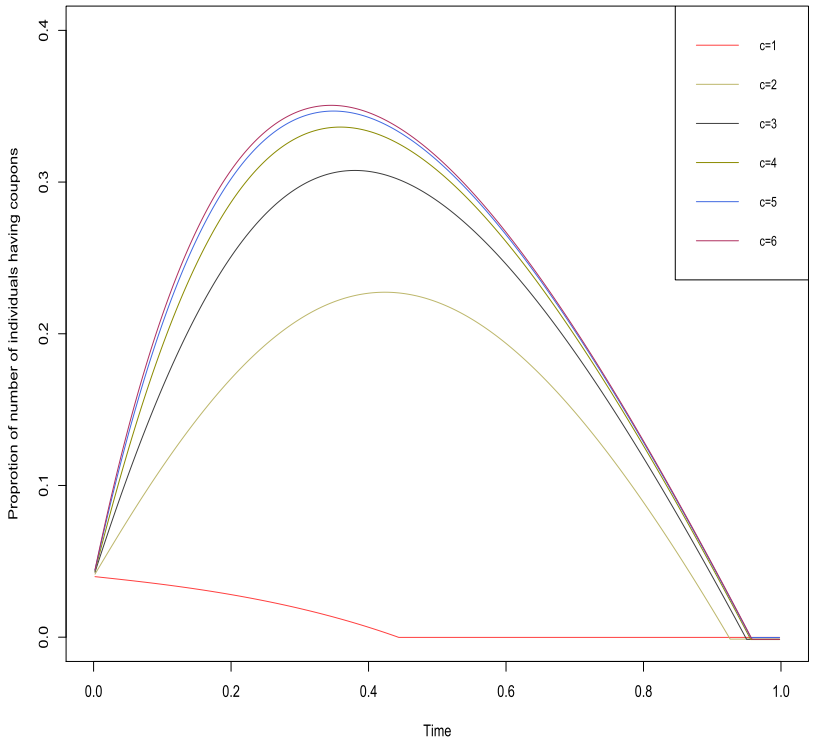


FIGURE 4. Plot the function $\|a\|$ for 6 cases: c takes values from 1 to 6. All other parameters are fixed: $\|a_0\| = 0.05, \pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{12} = 3, \lambda_{22} = 4$. The values $\|a_t\|$ represents the proportion of individuals having coupons at time t.

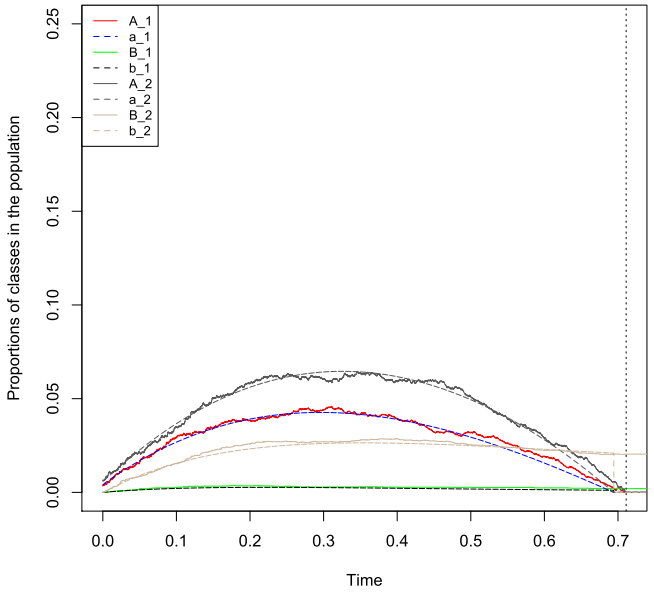


FIGURE 5. Plot the proportion of classes in the case $c = 3, N = 1000, A_0 = 10, \pi = (1/3, 2/3)$ and the graph is bipartite $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 4$.

TABLE 1. Numerical computation of t_0 for varying parameters c.

c	1	2	3	4	5	6	...
t_0	0.18	0.91	0.94	0.95	0.95	0.95	...

TABLE 2. Numerical computation of $\|A_t^N\| + \|B_t^N\|$ for varying parameters $c \in \{1, \dots, 6\}$ at time $t = 0.2$ and $N = 1000, A_0 = 10, \pi = (1/3, 2/3)$, $\lambda_{11} = 2, \lambda_{22} = 4, \lambda_{12} = 3$.

c	1	2	3	4	5	6
$\|A_{0.2}^{1000}\| + \|B_{0.2}^{1000}\|$	0.213	0.308	0.268	0.308	0.310	0.260

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 2a. The random walk stops when it encounters a node of degree 1 and cannot 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 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: $\|A_{0.2}^N\| + \|B_{0.2}^N\|$ for each c varying from 1 to 6 is given in Table 2.

Changing the parameters λ_{kl} 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 5.

REFERENCES

[1] E. Abbe, Community detection and stochastic block models. *Found. Trends Commun. Inform. Theory* **14** (2018) 1–162.
[2] K.B. Athreya and P. Jagers (Eds.), Classical and Modern Branching Processes, Vol. 84. Springer Science & Business Media (2012).
[3] S. Athreya and A. Röllin, Respondent driven sampling and sparse graph convergence. *Electron. Commun. Probab.* **23** (2018) 3.
[4] A. Bagheri and M. Saadati, Exploring the effectiveness of chain referral methods in sampling hidden populations. *Indian J. Sci. Technol.* **8** (2015) 1–8.
[5] A.D. Barbour, L. Holst and S. Janson, Poisson Approximation, Vol. 2 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York (1992).
[6] A. Barbour and G. Reiner, Approximating the epidemic curve. *Electron. J. Probab.* **18** (2013) 54.
[7] B. Bollobás and O. Riordan, Asymptotic normality of the size of the giant component via a random walk. *J. Combinat. Theory Ser. B* **102** (2012) 53–61.
[8] P. Billingsley, Convergence of Probability Measures. John Wiley & Sons, New York (1968).
[9] T. Britton and E. Pardoux, Stochastic Epidemic Models with Inference. Springer (2019).
[10] A. Cousien, J.S. Dhersin, V.C. Tran and T.P.T. Vo, Respondent driven sampling on sparse Erdös-rényi graphs. In preparation (2019).
[11] R. Durrett, Random Graph Dynamics, Vol. 200, No. 7. Cambridge University Press, Cambridge (2007).
[12] N. Enriquez, G. Faraud and L. Ménard, Limiting shape of the depth first search tree in an Erdös-Rényi graph. *Random Struct. Algorith.* **56** (2020) 501–516.

[13] S.N. Ethier and T.G. Kurtz, Markov Processes, Characterization and Convergence. John Wiley & Sons, New York (1986).
[14] A. Gadde, E.E. Gad, S. Avestimehr and A. Ortega, Active learning for community detection in stochastic block models. In *2016 IEEE International Symposium on Information Theory (ISIT)* (2016) 1889–1893.
[15] M. Girvan and M.E.J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **99** (2002) 7821–7826.
[16] L.A. Goodman, Snowball sampling. *Ann. Math. Statist.* (1961) 148–170.
[17] D.D. Heckathorn, Respondent-driven sampling: a new approach to the study of hidden populations. *Social Probl.* **44** (1997) 74–99.
[18] D.D. Heckathorn, Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Probl.* **49** (2002) 11–34.
[19] P.W. Holland, K.B. Laskey and S. Leinhardt, Stochastic blockmodels: first steps. *Social Netw.* **5** (1983) 109–137.
[20] A. Jakubowski, On the Skorokhod topology. *Ann. Inst. Henri Poincaré* **22** (1986) 263–285.
[21] S. Janson, M. Luczak and P. Windridge, Law of large numbers for the SIR epidemic on a random graph with given degrees. *Random Struct. Algorith.* **45** (2014) 726–763.
[22] A. Joffe and M. Métivier, Weak convergence of sequences of semimartingales with applications to multitype branching processes. *Adv. Appl. Probab* **18** (1986) 20–65.
[23] P. Barbillon, S. Donnet, E. Lazega and A. Bar-Hen, Stochastic Block Models for Multiplex networks: an application to networks of researchers. Preprint arXiv:1501.06444 (2015).
[24] M. Métivier, Semimartingales: A Course on Stochastic Processes, Vol. 2. Walter de Gruyter (2011).
[25] A. Shaghaghi, R.S. Bhopal and A. Sheikh, Approaches to recruiting 'hard-to-reach' populations into research: a review of the literature. *Health Promotion Perspect.* **1** (2011) 86.
[26] V.C. Tran, P. Moyal, L. Decreusefond and J.S. Dhersin, Limite en grand graphe d'un processus SIR décrivant la propagation d'une épidemie sur un réseau. *Journées MAS et Journée en l'honneur de Jacques Neveu, August 2010, Talence, France* (2010).
[27] R. Van Der Hofstad, Random Graphs and Complex Networks. Vol. 1. Cambridge University Press (2016).
[28] T.P.T. Vo, *Exploration of Random Graphs by the Respondent Driven Sampling method.* Ph.D. thesis, University Paris 13 (2020).