Respondent-driven sampling on sparse Erdös-Rényi graphs

Anthony Cousien, Jean-Stéphane Dhersin, Viet Chi Tran, Thi Phuong Thuy Vo

Respondent-driven sampling on sparse Erdös-Rényi graphs

Anthony Cousien[*] Jean-Stéphane Dhersin[†] Viet Chi Tran[‡] Thi Phuong Thuy Vo[§]

June 7, 2021

Abstract

We study the exploration of an Erdös-Rényi random graph by a respondent-driven sampling method, where discovered vertices reveal their neighbours. Some of them receive coupons to reveal in their turn their own neighbourhood. This leads to the study of a Markov chain on the random graph that we study. For sparse Erdös-Rényi graphs of large sizes, this process correctly renormalized converges to the solution of a deterministic curve, solution of a system of ODEs absorbed on the abscissa axis. The associated fluctuation process is also studied, providing a functional central limit theorem, with a Gaussian limiting process. Simulations and numerical computation illustrate the study.

Keywords: random graph; random walk exploration; respondent driven sampling; chain-referral survey.
AMS Classification: 62D05; 05C81; 05C80; 60F17; 60J20

Acknowledgements: This work was partially funded by the French Agence Nationale de Recherche sur le Sida et les Hépatites virales (ANRS, http://www.anrs.fr), grant number 95146. V.C.T. and T.P.T.V. have been supported by the GdR GeoSto 3477, ANR Econet (ANR-18-CE02-0010) and by the Chair "Modélisation Mathématique et Biodiversité" of Veolia Environnement-Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X. V.C.T. and T.P.T.V. acknowledge support from Labex Bézout (ANR-10-LABX-58). The authors would like to thank the working group previously involved in the development of the model for HCV transmission among PWID: Sylvie Deuffic-Burban, Marie Jauffret-Roustide and Yazdan Yazdanpanah.

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 public health issues for instance. We refer to [15, 2, 7, 29, 28] for AIDS or to [9, 8, 20] for HCV, for example. To achieve this in cases where the populations are hidden, it is possible to use respondent-driven sampling methods, where respondents recruit their peers [17, 19, 25]. 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 [19, 34, 16, 18, 23, 10]), an initial set of individuals are recruited in the population (with

[*]Université de Paris, IAME, INSERM, F-75018 Paris, France; E-mail: anthony.cousien@inserm.fr
[†]Jean-Stéphane Dhersin, Univ. Paris 13, CNRS, UMR 7539 - LAGA, 99 avenue J.-B. Clément, F-93430 Villetaneuse, France; E-mail: dhersin@math.univ-paris13.fr
[‡]Tran Viet Chi, LAMA, Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, F-77454 Marne-la-Vallée, France; E-mail: chi.tran@univ-eiffel.fr
[§]Vo Thi Phuong Thuy, Univ. Paris 13, CNRS, UMR 7539 - LAGA, 99 avenue J.-B. Clément, F-93430 Villetaneuse, France; E-mail: phuongthuywz@gmail.com

possible rules) and each of them is given a certain number of coupons. The coupons are distributed by recruited individuals to their contacts. The latter 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. [30, 31] 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 [14].

We consider here 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. Distributing coupons allows to control the growth of the number of people that are to be interviewed. 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 their 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 they have 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:

$$X_n = (A_n, B_n) \in \{0, \dots N\}^2, \quad n \in \mathbb{N}. \tag{1}$$

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 $A_n = 0$, then either the process stops, or a new individual is chosen from the unexplored population and recruited, no coupon is distributed, and we continue the survey. In this case, the process stops at $n = N$, when all vertices in the population have been explored. Thus,

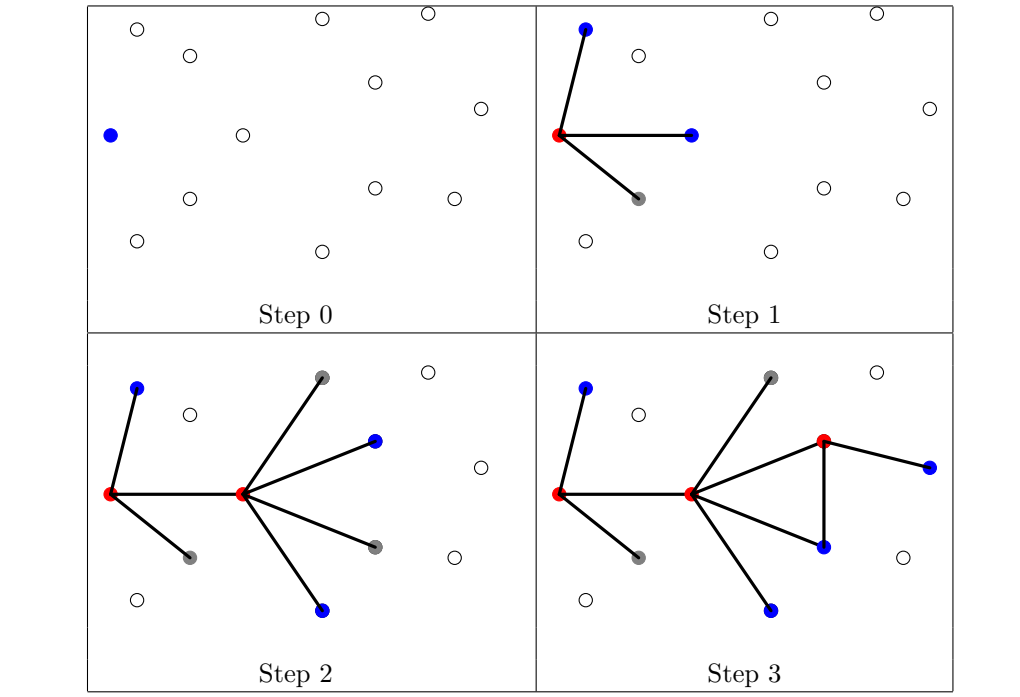

Figure 1: Description of how the chain-referral sampling works. In our model, the random network and the RDS are constructed simultaneously. For example at step 3, an edge between two vertices who are already known at step 2 is revealed.

$$\begin{aligned} A_{n+1} =& A_n - \mathbf{1}_{\{A_n \geq 1\}} + Y_{n+1} \wedge c, \\ B_{n+1} =& B_n + H_{n+1} - (H_{n+1} + K_{n+1}) \wedge c \end{aligned} \tag{2}$$

where Y_{n+1} is the number of new neighbors 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).

Case of Erdös-Rényi graphs If the graph that we explore is an Erdös-Rényi graph [5, 11], 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

$$Y_n \overset{(d)}{=} \mathcal{B}in\big(N-n-A_{n-1}, \frac{\lambda}{N}\big) \tag{3}$$

$$H_n \overset{(d)}{=} \mathcal{B}in\big(N-n-A_{n-1}-B_{n-1}, \frac{\lambda}{N}\big) \tag{4}$$

$$K_n \overset{(d)}{=} \mathcal{B}in\big(B_{n-1}, \frac{\lambda}{N}\big). \tag{5}$$

As we can notice in the presentation of the RDS process $(X_n)_{n\in\mathbb{N}}$, the exploration of the Erdös-Rényi random graph is done by visiting it with non-intersecting branching random walks. This idea is not new (see e.g. [6, 12, 27]).

Plan of the paper In Section 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 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 4.
This work is part of the PhD thesis of Vo Thi Phuong Thuy [33]. The law of large numbers (Theorem 2) can be seen as a particular case of the result of one of her other paper [32] where the considered graph is a Stochastic Block Model (see e.g. [1]). In the present work, the result is stated more clearly in this simplified setting (Erdös-Rényi graphs being seen as Stochastic Block Models with a single class) and is completed with the computation of the fluctuations (Section 4). We also considered the computation of several quantities of interest in Section 2.2 using the properties of Markov chains.

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

2 Study of the discrete-time RDS process on an Erdös-Rényi graph

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 (3). It is then possible to compute the transitions of this process that depend on the time $n\in\{0,\dots N\}$.

Proposition 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)-(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: $\mathbb{P}(X_n=(a',b')\mid X_{n-1}=(a,b))=P_n((a,b),(a',b'))$.*

$$P_n((a,b),(a',b')) = \sum_{(h,k)} \binom{b}{k}\binom{N-n-a-b}{h} p^{h+k}(1-p)^{N-n-a-h-k}, \tag{6}$$

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 (2) and (3). The fact that this probability depends only on X_n shows the Markov property and provides the transition probability (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:

$$A_n + n \leq N \qquad \Leftrightarrow \qquad A_n \leq N - n. \tag{7}$$

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,

$$\ell - (n-m) \leq A_n \leq \ell + (n-m) \times (c-1). \tag{8}$$

Thus, the points (n, A_n), for $n \geq m$, belong to the grey area on Fig. 2. Let us denote by S this grey region defined by (7) and (8).

$$S = \left\{ (n,a) \in \{m, ..., N\} \times \{0, ..., N-\ell\} \;\middle|\; \max\{\ell-(n-m), 0\} \leq a \leq \min\{\ell+(n-m)\times(c-1), N-n\} \right\}. \tag{9}$$

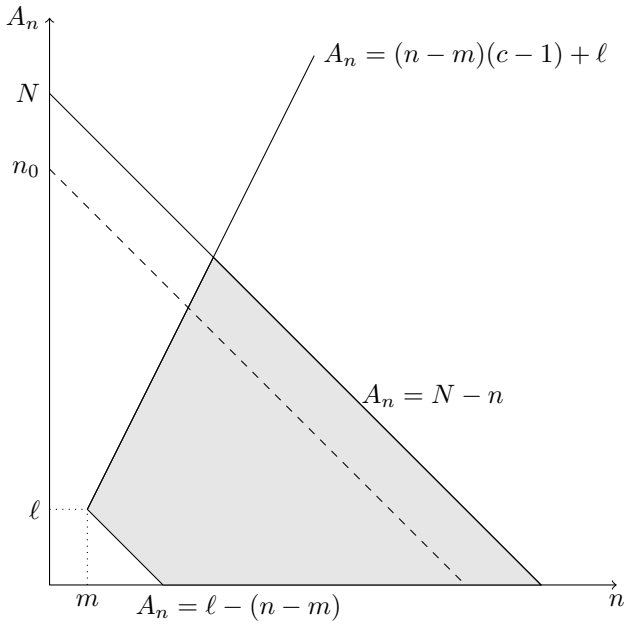

Figure 2: *Grey area S: Set of states susceptible to be reach from the process (A_n) started at time m with $A_m = \ell$, as defined by the constraints* (7) *and* (8)*. The process (A_n) can be stopped upon touching the abscissa axis, which corresponds to the state when the interviews stop because there 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.*

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

$$\tau := \inf\{n \geq 0, A_n = 0\} \tag{10}$$

the first time where the RDS process touches the abcissa 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:

$$u_{n_0}(m, \ell) = \mathbb{P}\big(\tau = n_0 \mid A_m = \ell\big). \tag{11}$$

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:

$$u_{n_0}(n_0, 0) = 1, \qquad \forall n \neq n_0,\ u_{n_0}(n, 0) = 0, \tag{12}$$

$$u_{n_0}(n, a) = \sum_{a' \ | \ (n+1,a') \in S} P_n\big(a, a'\big) u_{n_0}(n+1, a'), \quad n \leq n_0 - 1,\ 1 \leq a \leq N, \tag{13}$$

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 $n_0 < N$:

$$S_{n_0} = \left\{ (n, a) \in \{m, ..., N\} \times \{0, ..., N - \ell\} \ \middle| \ \max\{\ell - (n-m), 0\} \leq a \leq \min\{\ell + (n-m) \times (c-1), n_0 - n\} \right\} \tag{14}$$

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. 2).

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

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

Let us define for $n \geq 0$:

$$\mathbf{U}_{n_0}^{(n)} := \begin{pmatrix} u_{n_0}(n, 1) \\ \vdots \\ u_{n_0}(n, a) \\ \vdots \\ u_{n_0}(n, n_0) \end{pmatrix} \tag{15}$$

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 solution of the system (13) can be solved recursively with the boundary conditions (12) and:

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

We can compute solve the above equations, as represented in Fig. 3.

Starting from 1 coupon at time 0, we can also compute the probabilities that $\mathbb{P}(\tau > n_0 \mid A_0 = 1)$ as seen in Fig. 4

3 Limit of the normalized RDS process

In this section, we consider the RDS process stopped at the time τ_0^N defined for every $N \in \mathbb{N}^*$ as

$$\tau_0^N := \inf\{t > 0, A_t^N = 0\}. \tag{16}$$

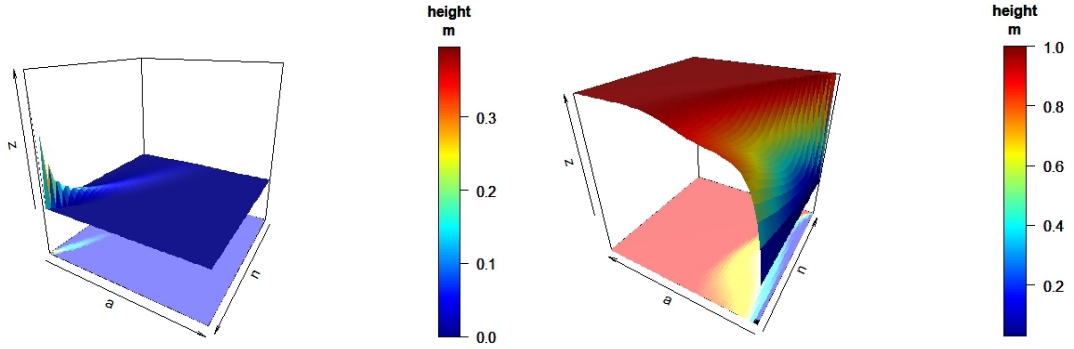

Figure 3: *Left: Numerical computation of* $u_{n_0}(n,a) = \mathbb{P}(\tau = n_0 \mid A_n = a)$*, with* $n_0 = 50$ *and for* a *and* n *varying between 0 and 50. Right: Numerical computation of the probability* $\mathbb{P}(\tau > n_0 \mid A_n = a)$*, with* $n_0 = 50$*. We can see that if* $a \geq n$ *then, this probability is equal to 1. We can use these numerical results for* $n = 0$*: this provides the probability, given the number of seeds, to reach a sample of size at least* n_0*.*

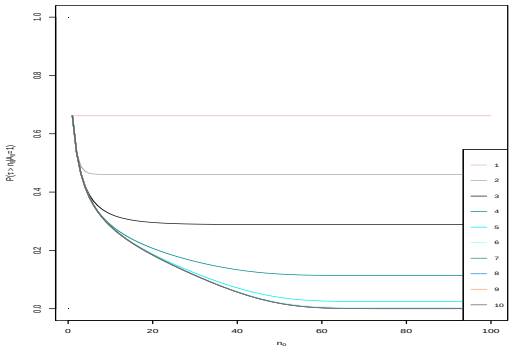

Figure 4: *Numerical computation of the probability* $\mathbb{P}(\tau > n_0 \mid A_0 = 1)$ *of obtaining a sample of size at least* n_0 *starting from 1 coupon at time 0 (ordinate), with* c *varying from 1 to 10 (colours) and* n_0 *varying between 1 and 100 (abscissa).*

This time corresponds to the proportion of population explored by the RDS.

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

$$X_t^N = \left(A_t^N, B_t^N\right) := \frac{1}{N} X_{\lfloor Nt \rfloor} = \left(\frac{A_{\lfloor Nt \rfloor}}{N}, \frac{B_{\lfloor Nt \rfloor}}{N}\right) \in [0,1]^2, \quad t \in [0,1]. \tag{17}$$

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 [4, 22]. 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. *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 2. *Under the assumption 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*

$$x_t = x_0 + \int_0^t f(s, x_s)ds, \tag{18}$$

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

$$f_1(t,x_t) = f_1(t,a_t,b_t) = c - \sum_{k=0}^{c-1}(c-k)p_k(t+a_t) - \mathbf{1}_{a_t>0} \tag{19}$$

$$f_2(t,x_t) = f_2(t,a_t,b_t) = (1-t-a_t-b_t)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t+a_t) - c, \tag{20}$$

with

$$p_k(z) := \frac{\lambda^k(1-z)^k}{k!}e^{-\lambda(1-z)}, \quad k \in \{0,...,c\}, \tag{21}$$

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

Remark 3. *Since the limiting process $x \in \mathcal{C}([0,1],[0,1]^2)$ is deterministic, the convergence in distribution of Theorem 2 is in fact a convergence in probability. The limiting shape is illustrated in Fig. 5.*
It can be seen graphically that after the time $t_0 = \inf\{t > 0,\ |a_t| = 0\}$, the solution of the ODE remains constant (see Prop 12) and that this time is approximated by $\tau_0^N = \inf\{t > 0,\ A_t^N = 0\}$. This time corresponds to the proportion of population explored by the RDS. We have simulated the RDS on the graph of size $N = 1000$, $\lambda = 2$ and for various values of c: $c = 1,2,3,4$, and have computed τ_0^N. We obtain the approximated values of t_0 in the table below:

c	*1*	*2*	*3*	*4*	*5*	*6*
t_0	*0.426*	*0.775*	*0.818*	*0.827*	*0.829*	*0.829*

The proof of Theorem 2 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 (Section 3.1). Then, we identify the equation satisfied by the limiting values of $(X^N)_{N\geq 1}$ (Section 3.2), and show that the latter has a unique solution (Section 3.3). A difficulty lies in the fact that when a_t touches zero, the solution stops and remains constant. Indeed, the solution of (18) describes how the number of coupons distributed in the population evolves. When $a_t = 0$, this gives the size of the cluster reached by the RDS without the introduction of additional seeds.

As explained, the proof of Theorem 2 relies on the Doob's decomposition of $(X^N)_{N\geq 1}$ as follows.

Lemma 4. *The process X^N, for $N \in \mathbb{N}^*$, admits the following Doob decomposition: $X_t^N = X_0^N + \Delta_t^N + M_t^N$, or in the vectorial form*

$$\begin{pmatrix} X^{N,1} \\ X^{N,2} \end{pmatrix} = \begin{pmatrix} A_0^N \\ B_0^N \end{pmatrix} + \begin{pmatrix} \Delta^{N,1} \\ \Delta^{N,2} \end{pmatrix} + \begin{pmatrix} M^{N,1} \\ M^{N,2} \end{pmatrix}. \tag{22}$$

The predictable process Δ^N is:

$$\begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt \rfloor} \begin{pmatrix} \mathbb{E}[Y_n \wedge c \mid \mathcal{F}_{n-1}] - \mathbf{1}_{A_{n-1}\geq 1} \\ \mathbb{E}[H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}] \end{pmatrix} \tag{23}$$

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

$$\langle M^N \rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt \rfloor} \begin{pmatrix} Var\big(Y_n \wedge c \mid \mathcal{F}_{n-1}\big) & Cov\big(Y_n \wedge c, H_n - Y_n \wedge c\big) \\ Cov\big(Y_n \wedge c, H_n - Y_n \wedge c\big) & Var\big(H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}\big) \end{pmatrix}. \tag{24}$$

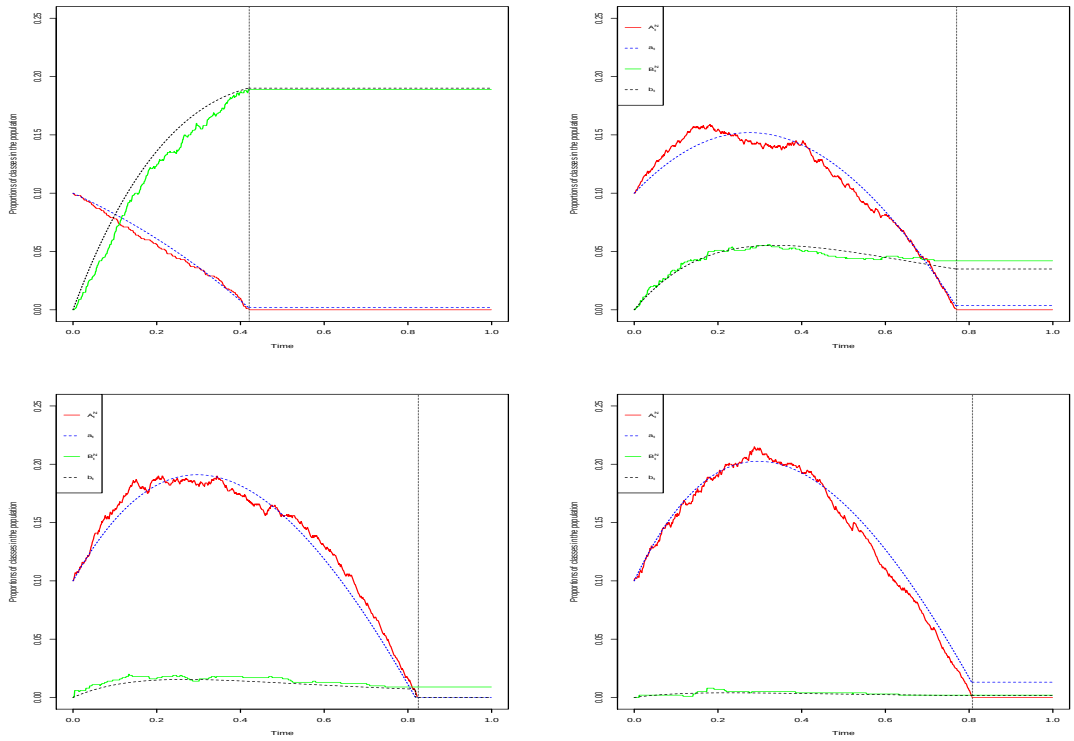

Figure 5: *Simulations of the process* (A^N, B^N) *(red/green lines) compared to solution* (a, b) *of the ODE's system (dashed lines) for the graph of size* $N = 1000$*, the value of* λ *is fixed:* $\lambda = 2$ *and various values of* c*:* $c = 1, 2, 3, 4$*.*

Notice that the quantities in (23) and (24) can be computed as functions of $A^N_{t_{n-1}} = A_{n-1}/N$ and $B^N_{t_{n-1}} = B_{n-1}/N$ for $n \in \{1, ..., N\}$ with the results of the following lemma:

Lemma 5. *We have the following expressions for the expectations and variances appearing in* (23) *and* (24)*.*
(i) For the moments of Y_n*:*

$$\mathbb{E}[Y_n \wedge c \mid \mathcal{F}_{n-1}] = c - \sum_{k=0}^{c-1}(c-k)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) \tag{25}$$

$$\mathbb{E}\left[(Y_n \wedge c)^2 \,\middle|\mathcal{F}_{n-1}\right] = c^2 + \sum_{k=0}^{c}(k^2 - c^2)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}); \tag{26}$$

where

$$\begin{aligned}\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) =& \frac{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1)!}{(N - Nt_{n-1} - NA^N_{t_{n-1}} - 1 - k)!N^k}\frac{\lambda^k}{k!} \\ &\times \left(1 - \frac{\lambda}{N}\right)^{N(1-t_{n-1}-A^N_{t_{n-1}})}\left(1 - \frac{\lambda}{N}\right)^{-k-1}.\end{aligned} \tag{27}$$

(ii) For the moments of H_n:

$$\mathbb{E}[H_n|\mathcal{F}_{n-1}] = \lambda\left(1 - t_n - A^N_{t_{n-1}} - B^N_{t_{n-1}}\right) \tag{28}$$

$$Var(H_n|\mathcal{F}_{n-1}) = \lambda\left(1 - t_n - A^N_{t_{n-1}} - B^N_{t_{n-1}}\right)\left(1 - \frac{\lambda}{N}\right). \tag{29}$$

(iii) And for the cross moments:

$$\begin{aligned}&\mathbb{E}[H_n(Y_n \wedge c)|\mathcal{F}_{n-1}]\\ &\quad= \left(1 - \frac{B^N_{t_{n-1}}}{1 - t_n - A^N_{t_{n-1}}}\right)\left[\sum_{k=0}^{c}(k^2 - ck)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\lambda\left(1 - t_n - A^N_{t_{n-1}}\right)\right].\end{aligned} \tag{30}$$

Proof of Lemma 5. Most of the computation comes straightforward from (3). For the conditional expectation of $Y_n \wedge c$, we have:

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

where

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

which yields (27).

Let us detail the proof of (30).

$$\begin{aligned}\mathbb{E}[H_n(Y_n \wedge c)|\mathcal{F}_{n-1}] =& \sum_{k=0}^{N-n-A_{n-1}} (k \wedge c)\mathbb{E}(H_n|Y_n = k)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1})\\ =& \frac{N - n - A_{n-1} - B_{n-1}}{N - n - A_{n-1}}\left[\sum_{k=0}^{c} k^2\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + \sum_{k=c+1}^{N-n-A_{n-1}} ck\mathbb{P}(Y_n = k|\mathcal{F}_{n-1})\right]\\ =& \left(1 - \frac{B_{n-1}}{N - n - A_{n-1}}\right)\left[\sum_{k=0}^{c}(k^2 - ck)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\mathbb{E}[Y_n|\mathcal{F}_{n-1}]\right]\\ =& \left(1 - \frac{B_{n-1}}{N - n - A_{n-1}}\right)\left[\sum_{k=0}^{c}(k^2 - ck)\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) + c\lambda\left(1 - \frac{n}{N} - \frac{A_{n-1}}{N}\right)\right].\end{aligned} \tag{31}$$

□

With the expressions obtained in Lemma 5, we can now prove Lemma 4.

Proof of Lemma 4. 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 $(\Delta_t^N)_{0\le t\le 1}$. 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 (2), as $Y_n \wedge c \le c$ and $\mathbb{E}[H_n \mid \mathcal{F}_{n-1}] \le \lambda$.

Furthermore, using (2), we have that for the first component:

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

and we can recover the expression (23) of Δ^N announced in the lemma with the results of Lemma 5.

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

$$|M_t^N| = \left|X_t^N - X_0^N - \Delta_t^N\right| \le 2 + (2c+\lambda)t \le 2 + 2c + \lambda < +\infty.$$

For all $s < t$,

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

Then M_t^N is an $(\mathcal{F}_t^N)$-martingale.

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

$$\langle M^N \rangle_t = \begin{bmatrix} \langle M^{N,1}, M^{N,1} \rangle_t & \langle M^{N,1}, M^{N,2} \rangle_t \\ \langle M^{N,2}, M^{N,1} \rangle_t & \langle M^{N,2}, M^{N,2} \rangle_t \end{bmatrix}, \tag{32}$$

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

$$\begin{aligned} \langle M^{N,k}, M^{N,\ell} \rangle_t = \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \big\{ & \mathbb{E}\left[(X_n^k - X_{n-1}^k)(X_n^\ell - X_{n-1}^\ell) \big| \mathcal{F}_{n-1} \right] \\ & - \mathbb{E}\left[(X_n^k - X_{n-1}^k) | \mathcal{F}_{n-1} \right] \mathbb{E}\left[(X_n^\ell - X_{n-1}^\ell) | \mathcal{F}_{n-1} \right] \big\}. \end{aligned} \tag{33}$$

Using (2), we have:

$$\begin{aligned} \langle M^{N,1} \rangle_t &= \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \mathbb{E}\left[(A_n - A_{n-1} - \mathbb{E}[A_n - A_{n-1} | \mathcal{F}_{n-1}])^2 \, \big| \mathcal{F}_{n-1} \right] \\ &= \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \mathrm{Var}(Y_n \wedge c | \mathcal{F}_{n-1}) \leq \frac{c^2}{N}. \end{aligned} \tag{34}$$

Proceeding similarly for the other terms, we obtain

$$\begin{aligned} \langle M^{N,2} \rangle_t &= \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \mathrm{Var}(H_n - Y_n \wedge c | \mathcal{F}_{n-1}) \leq \frac{\lambda}{N}, \\ \langle M^{N,1}, M^{N,2} \rangle_t &= \frac{1}{N^2} \sum_{n=1}^{\lfloor Nt \rfloor} \mathrm{Cov}(Y_n \wedge c, H_n - Y_n \wedge c \mid \mathcal{F}_{n-1}) \leq \frac{c\sqrt{\lambda}}{N}. \end{aligned} \tag{35}$$

This finishes the proof of the Lemma. □

3.1 Tightness of the renormalized process

Lemma 6. *The sequence $(X^N)_{N \geq 1}$ is tight in $\mathcal{D}([0,1], [0,1]^2)$.*

Proof. The proof of tightness is based on the classical criterion of Aldous-Rebolledo ([24, 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 $t \in [0,1]$, $|A_t^N| + |B_t^N| \leq 2$, implying that (A_t^N, B_t^N) is tight for every $t \in [0,1]$.
Let $0 \leq s, t \leq 1$,

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

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

$$\mathbb{P}\left(\sup_{\substack{|t-s| \leq \delta \\ 0 \leq s,t \leq 1}} \left\| \Delta_t^N - \Delta_s^N \right\| > \eta \right) \leq \frac{1}{\eta} \mathbb{E}\left[\sup_{\substack{|t-s| \leq \delta \\ 0 \leq s,t \leq 1}} \left\| \Delta_t^N - \Delta_s^N \right\| \right] \leq \frac{(2c+\lambda)\delta}{\eta} \leq \varepsilon, \quad \forall N \geq 1. \tag{36}$$

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

Similarly, for the quadratic variations of the martingale parts, using (34) and (35), we have for all $0 \le s < t \le 1$,

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

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

$$\begin{aligned}
\sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}} \|\langle M^N\rangle_t - \langle M^N\rangle_s\| &\le \sup_{\substack{|t-s|\le\delta\\0\le s,t\le 1}} \Big(\left|\langle M^{N,1}\rangle_t - \langle M^{N,1}\rangle_s\right| + \left|\langle M^{N,2}\rangle_t - \langle M^{N,2}\rangle_s\right|\\
&\qquad + 2\left|\langle M^{N,1}, M^{N,2}\rangle_t - \langle M^{N,1}, M^{N,2}\rangle_s\right|\Big)\\
&\le \frac{c^2 + 4(\lambda+c^2) + c(\sqrt{\lambda}+c)}{N}\delta. \qquad (37)
\end{aligned}$$

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. □

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.* [3]). In the whole section, we will denote this sequence again by $(X^N)_{N\ge1} = (A^N, B^N)_{N\ge1}$. We now want to identify that limiting value.

3.2.1 Convergence of the martingale and predictable process

Proposition 7. *The sequence of martingales $(M^N)_{N\ge1}$ converges uniformly to 0 in probability when $N \to \infty$.*

Proof. With a computation similar the one leading to (37), we get

$$\|\langle M\rangle_t\| \le |\langle M^{N,1}\rangle_t| + |\langle M^{N,2}\rangle_t| + 2|\langle M^{N,1}\rangle_t|^{1/2}|\langle M^{N,2}\rangle_t|^{1/2} \le \frac{(6c^2+4\lambda)t}{N} \tag{38}$$

By Doob's inequality,

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

For every $\varepsilon > 0$,

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

□

The remaining work is figuring out the limit of finite variation part Δ^N. Recall the function f defined in (19)-(20).

Proposition 8. *There exists a constant $C = C(\lambda, c) > 0$ such that for all $N \ge 1$,*

$$\sup_{t\in[0,1]} \left\|\Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\Big(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N}\Big)\right\| \le \frac{C}{N} \tag{39}$$

Proof. Recall the equations for Δ^N in (23) and (27). Using (28), we have that:

$$\begin{aligned}
&\Big\|\Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\big(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N}\big)\Big\| \\
\le&\Big|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\Big(c - \sum_{k=0}^{c}(c-k)\mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - \mathbf{1}_{A_{n-1}\ge 1}\Big) \\
&\qquad\qquad - \Big(c - \sum_{k=0}^{c}(c-k)p_k\big(\frac{n-1}{N} + \frac{A_{n-1}}{N}\big) - \mathbf{1}_{\frac{A_{n-1}}{N}>0}\Big)\Big| \\
&+ \Big|\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\Big(\mathbb{E}[H_n \mid \mathcal{F}_{n-1}] + \sum_{k=0}^{c}(c-k)\mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - c\Big) \\
&\qquad\qquad - \Big(\lambda\big(1 - \frac{n-1}{N} - \frac{A_{n-1}}{N} - \frac{B_{n-1}}{N}\big) - \sum_{k=0}^{c}(c-k)p_k\big(\frac{n-1}{N} + \frac{A_{n-1}}{N}\big)\Big)\Big| \\
\le&\frac{2}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\sum_{k=0}^{c}(c-k)\left|\mathbb{P}\big(Y_n = k \mid \mathcal{F}_{n-1}\big) - p_k\left(\frac{n-1}{N} + \frac{A_{n-1}}{N}\right)\right|.
\end{aligned} \tag{40}$$

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:

$$\begin{aligned}&\left|\frac{(N - Nt_i - NA^N_{t_i} - 1)!}{(N - Nt_i - NA^N_{t_i} - k - 1)!N^k} - (1 - t_i - A^N_{t_i})^k\right| \\ \leq &\sum_{j=0}^{k-1}\left(\frac{k-1}{N}\right)^{k-j} \leq \frac{\sum_{\ell=1}^{k}(k-1)^\ell}{N}. \end{aligned} \tag{41}$$

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:

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

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 $C_0 = C_0(\lambda)$,

$$0 \leq e^{-\lambda(1-t_n-A^N_{t_n})} - \Big(1 - \frac{\lambda}{N}\Big)^{N(1-t_n-A^N_{t_n})} \leq \frac{C_0}{N}. \tag{42}$$

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

$$1 \leq \Big(1 - \frac{\lambda}{N}\Big)^{-(k+1)} \leq 1 + \frac{C_1}{N}. \tag{43}$$

Gathering (27), (41), (42) and (43), there thus exists a constant $C_2 = C_2(c, \lambda)$ such that

$$\Big|\mathbb{P}(Y_n = k|\mathcal{F}_{n-1}) - p_k(t_{n-1} + A^N_{t_{n-1}})\Big| \leq \frac{C_2(\lambda, c)}{N}. \tag{44}$$

As a result, from (40) and (44) 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. □

3.2.2 Equations satisfied by the limiting values

First, let us start with a useful lemma.

Lemma 9. *The limiting value $(\bar{a}, \bar{b})$ is continuous.*

Proof. The jumps of the component A^N are uniformly bounded by c/N. However the component B^N may have jumps of order 1. To obtain the result, we will use the Proposition 3.26(iii) in [13] and prove that these big jumps tend to zero in probability when $N \to +\infty$. We will use (3). The random variables H_n, for $n \in \{1, \dots N\}$ do not have the same distribution, but we can couple them with i.i.d. dominating random variables $\widetilde{H}_n \rightsquigarrow \mathcal{B}in(N, \lambda/N)$. Then, for $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\Big(\sup_{t\in[0,1]} |B_t^N - B_{t_-}^N| > \varepsilon\Big) \leq & \mathbb{P}\Big(\sup_{n\in\{1,\dots N\}} |H_n| > N\varepsilon\Big) \\
\leq & \mathbb{P}\Big(\sup_{n\in\{1,\dots N\}} |\widetilde{H}_n| > N\varepsilon\Big) = 1 - \mathbb{P}\Big(\sup_{n\in\{1,\dots N\}} |\widetilde{H}_n| \leq N\varepsilon\Big) \\
= & 1 - \mathbb{P}\Big(|\widetilde{H}_1| \leq \varepsilon N\Big)^N = 1 - \Big(1 - \mathbb{P}\Big(|\widetilde{H}_1| > \varepsilon N\Big)\Big)^N \\
= & 1 - \left(1 - \mathbb{P}\Big(\widetilde{H}_1 - \lambda > \varepsilon N - \lambda\Big)\right)^N. \qquad (45)
\end{aligned}$$

By Hoeffding's inequality,

$$\mathbb{P}\big(\widetilde{H}_1 - \lambda \geq \varepsilon N - \lambda\big) \leq \exp\left(\frac{-2(\varepsilon N - \lambda)^2}{N}\right) = \exp\left(-2\varepsilon^2 N + 4\varepsilon\lambda - \frac{2\varepsilon\lambda^2}{N}\right) \leq \exp\left(-2\varepsilon^2 N + 4\varepsilon\lambda\right).$$

The right hand side of (45) is less or equal to

$$1 - \left(1 - \exp\big(-2\varepsilon^2 N + 4\varepsilon\lambda\big)\right)^N = 1 - \exp\left(N \ln\left(1 - e^{-2\varepsilon^2 N + 4\varepsilon\lambda}\right)\right).$$

Thus,

$$\lim_{N\to+\infty} \mathbb{P}\Big(\sup_{t\in[0,1]} |B_t^N - B_{t_-}^N| > \varepsilon\Big) \leq \lim_{N\to+\infty} \left(1 - \exp\left(N \ln\left(1 - e^{-2\varepsilon^2 N + 4\varepsilon\lambda}\right)\right)\right) = 0.$$

□

With the above results, we can now identify the equations satisfied by the limiting values of (X^N), and we will prove in the rest of this section that

Proposition 10. *The limiting values of $(X^N)_{N\geq 1}$ are solutions of* (18).

A difficulty in the proof comes from the change of the dynamics when the component A^N touches zero. Because of the indicator in (19), f_1 is not continuous when its second argument touches zero. We separate the proof into several steps.

Step 1: Recall the stopping time τ_0^N defined in (16). This time corresponds to the proportion of the graph discovered by the RDS when starting from the initial seeds at time 0 and without introduction of new coupons (see Remark 3). Also, we define for the limiting value:

$$t_0 := \inf\{t \in [0,1] : a_t = 0\}. \qquad (46)$$

Lemma 11. *We have that*

$$\lim_{N\to+\infty} \mathbb{P}(\tau_0^N \geq t_0) = 1. \qquad (47)$$

Proof of Lemma 11. For $\varepsilon > 0$, let

$$\tau_\varepsilon^N := \inf\{t > 0, A_t^N \leq \varepsilon\} \tag{48}$$

and

$$t_\varepsilon := \inf\{t > 0, a_t \leq \varepsilon\}. \tag{49}$$

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

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

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

$$\lim_{N\to\infty} \mathbb{P}(\tau_0^N \geq t_0) = 1. \tag{50}$$

□

Step 2: we now prove that on $[0, t_0)$, $(\bar{a}, \bar{b})$ satisfies the ODEs (18). From (22), Propositions 7 and 8, 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 $N \to +\infty$. A difficulty is that f is discontinuous when $A^{(N)}$ touches zero, but this happen with large probability after time t_0 when $N \to +\infty$ (see Lemma 11). Hence, by continuity, for $t < t_0$:

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

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 $(X^N)_{N\geq 1}$, $(\bar{a}, \bar{b})$ is necessarily solution of (18) on $[0, t_0)$.

In the sequel, we denote by (a, b) the solution of (18) and will prove that $(\bar{a}, \bar{b}) = (a, b)$.

Step 3: We now study the time t_0 at which the first component a of the solution of (18) touches zero. We first prove that:

Proposition 12. *For all $t \geq t_0$, we have $a_t = 0$.*

Before proving the proposition, let us start with a lemma.

Lemma 13. *Denote*

$$\phi(z) := c - \sum_{k=0}^{c-1}(c-k)\frac{[\lambda(1-z)]^k}{k!}e^{-\lambda(1-z)}, \quad c \geq 2, \lambda > 1. \tag{51}$$

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 $z \in [0,1]$,

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

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

Proof of Proposition 12. For $c = 1$, (19)-(20) 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 (19)-(20) 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

$$\phi(z) := c - \sum_{k=0}^{c-1}(c-k)p_k(z) = c - \sum_{k=0}^{c-1}(c-k)\frac{\lambda^k(1-z)^k}{k!}e^{-\lambda(1-z)}. \tag{52}$$

By Lemma 13, ϕ is a positive function on $(0,1)$ and 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. □

Step 2, Prop 12 and Lemma 11 ensure that $(X^N)_{N\geq 1}$ converges uniformly to $(\bar{a}, \bar{b})$ on $[0; t_0)$.

Step 4: We now precise the result of Lemma 11 by showing that τ_0^N converges to t_0 in probability. This results from Lemma 11 and:

Lemma 14. *For all $\sigma > 0$,*

$$\lim_{N\to+\infty} \mathbb{P}(\tau_0^N \leq t_0 + \sigma) = 1. \tag{53}$$

Proof. Let $\sigma > 0$ and $\delta > 0$ be small positive numbers,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{P}(\tau_0^N > t_0 + \sigma) &= \lim_{N\to\infty} \mathbb{P}\big(\inf_{t\leq t_0+\sigma} A_t^N > 0\big) \\
&= \lim_{N\to\infty} \mathbb{P}\big(\inf_{t\leq t_0+\sigma} A_t^N > 0 \mid A_{t_0}^N \leq \delta\big)\mathbb{P}\big(A_{t_0}^N \leq \delta\big) \\
&\leq \lim_{N\to\infty} \mathbb{P}\big(A_{t_0+\sigma}^N > 0 \mid A_{t_0}^N \leq \delta\big),
\end{aligned}$$

since $A_{t_0}^N$ converges to $a_{t_0} = 0$ in probability, in the second line. Using Lemma 4,

$$A_{t_0+\sigma}^N - A_{t_0}^N = \frac{1}{N} \sum_{n:\ t_0\leq \frac{n}{N}\leq t_0+\sigma} \big(\mathbb{E}\big(Y_n \wedge c \mid \mathcal{F}_{n-1}\big) - 1\big) + M_{t_0+\sigma}^{N,1} - M_{t_0}^{N,1}.$$

For n such that $t_0 \leq \frac{n}{N} \leq t_0 + \sigma$, there exists $\epsilon > 0$ such that

$$\mathbb{E}\big(Y_n \wedge c \mid \mathcal{F}_{n-1}\big) - 1 < -\epsilon,$$

since $\mathbb{E}\big(Y_n \wedge c \mid \mathcal{F}_{n-1}\big) - 1 = \phi\big(\frac{n-1}{N} + \frac{A_{n-1}}{N}\big) - 1 + o\big(\frac{1}{N}\big)$ and since $\phi(t) - 1 < 0$ for all $t \geq t_c$, where $t_c \in [0, t_0]$ is defined in the proof of Lemma 11. Thus,

$$\begin{aligned}
\mathbb{P}\big(A_{t_0+\sigma}^N > 0 \mid A_{t_0}^N \leq \delta\big) &\leq \mathbb{P}\bigg(M_{t_0+\sigma}^{N,1} - M_{t_0}^{N,1} + \delta - \frac{\lfloor N\sigma\rfloor}{N}\epsilon > 0\bigg) \\
&= \mathbb{P}\bigg(M_{t_0+\sigma}^{N,1} - M_{t_0}^{N,1} > \frac{\lfloor N\sigma\rfloor}{N}\epsilon - \delta\bigg) \\
&\leq \mathbb{P}\bigg(M_{t_0+\sigma}^{N,1} - M_{t_0}^{N,1} > \sigma\epsilon - \frac{\epsilon}{N} - \delta\bigg).
\end{aligned}$$

For $\delta < \sigma\epsilon$, there exists N_0 such that $\sigma\epsilon - \delta - \frac{\epsilon}{N_0} > 0$. Thus, using Markov's inequality and then using (38), we have that for all $N \geq N_0$,

$$\mathbb{P}\bigg(M_{t_0+\sigma}^{N,1} - M_{t_0}^{N,1} > \sigma\epsilon - \frac{\epsilon}{N} - \delta\bigg) \leq \frac{\mathbb{E}\bigg(\Big|\langle M\rangle_{t_0+\sigma}^{N,1} - \langle M\rangle_{t_0}^{N,1}\Big|\bigg)}{\bigg(\sigma\epsilon - \delta - \frac{\epsilon}{N_0}\bigg)^2} \leq \frac{2(6c^2+4\lambda)}{N\bigg(\sigma\epsilon - \delta - \frac{\epsilon}{N_0}\bigg)^2},$$

which tends to zero when $N \to +\infty$. □

Step 5: Conclusion. We have proved that X^N converges uniformly to the solution $(\bar{a}, \bar{b})$ of (19)-(20) on $[0, t_0]$. The stochastic process is frozen at the time τ_0^N that converges to t_0 in probability. After that time τ_0^N, X^N remains constant by construction with $A_t^N = 0$. So, A^N converges uniformly on $[0, 1]$ to the solution $\bar{a}$ of (19). The continuity of the process A^N yields the continuity of the process B^N that hence also converges uniformly to the solution $\bar{b}$ of (20) on the whole time interval $[0, 1]$. This concludes the proof of Proposition 10.

3.3 Uniqueness of the ODE solutions

To prove Theorem 2, it remains to prove the uniqueness of the limiting value, *i.e.* that:

Proposition 15. *The system of differential equations* (19)-(20) *admits a unique solution.*

Proof. Suppose that (19)-(20) have two solutions (a^1, b^1) and (a^2, b^2), then for all $t \in [0,1]$,

$$|a_t^1 - a_t^2| \leq \int_0^t |g(s, a_s^1) - g(s, a_s^2)| ds + \int_0^t \left| \mathbf{1}_{\{a_s^1 > 0\}} - \mathbf{1}_{\{a_s^2 > 0\}} \right| ds, \tag{54}$$

where

$$g(t, a_t, b_t) := c - \sum_{k=0}^{c-1} (c-k) p_k(t + a_t). \tag{55}$$

In the first term of the right hand side of (54), we have

$$|g(s, a_s^1) - g(s, a_s^2)| \leq |\partial_a g(s, \xi_s)| |a_s^1 - a_s^2|, \tag{56}$$

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 $t \in [0,1]$,

$$\int_0^t \left| \mathbf{1}_{a_s^1 > 0} - \mathbf{1}_{a_s^2 > 0} \right| ds = 0. \tag{57}$$

In order to do so, we first prove that all the solutions of (19) touch zero at the same point and that after touching zero, they stay at zero. Consider the equation:

$$\frac{d\bar{a}_t}{dt} = g(t, \bar{a}_t) - 1. \tag{58}$$

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 (19)' has unique solution $\bar{a}_t$ for t in $[0,1]$. Let us define

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

Since the two equations (19) and (58) 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 (57), 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 12.

Consequently, from (56) and (57), we have

$$|a_t^1 - a_t^2| \leq \int_0^t |\partial_a g(s, \xi_s)| |a_s^1 - a_s^2| ds. \tag{59}$$

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

$$|b_t^1 - b_t^2| \leq \int_0^t \max_{a \in [0,1]} |\partial_b f_2(s, a, \zeta_s)| |b_s^1 - b_s^2| ds, \tag{60}$$

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 (19)-(20) have at most one solution. □

Every subsequence $(X^{\ell_N})_{N\geq 1} \subset (X_N)_{N\geq 1}$ converges in distribution to a solution of the differential equations (19)-(20). And because of the uniqueness of the solution of (19)-(20), 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.

4 The central limit theorem

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 ([11]) 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. 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, there is a positive fraction of seeds belonging to the giant component of $ER(N, \lambda/N)$ with a probability converging to 1.

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 lemma 11, we see that the Markov process $(A_t^N)_{N\geq 1}$ absorbs just 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 hence consider the following process, for $t \in [0, t_0]$:

$$W_t^N := \frac{X_{\lfloor Nt \rfloor} - N(a_t, b_t)}{\sqrt{N}} = \sqrt{N}(X_t^N - x_t), t \in [0, t_0], N \in \mathbb{N}^*. \tag{61}$$

Assumption 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 16. *Under Assumption 2, the process $(W^N)_{N\geq 1}$ converges in distribution in $\mathcal{D}([0, t_0], \mathbb{R}^2)$ to Y, which satisfies*

$$W_t = W_0 + \int_0^t G(s, a_s, b_s, W_s)ds + M(t, a_t, b_t) \tag{62}$$

where

$$G(t, a, b, w) := \begin{pmatrix} \phi'(t+a)w^1 \\ -\lambda(w^1 + w^2) - \phi'(t+a)w^1 \end{pmatrix}; \tag{63}$$

$$\phi(z) := c - \sum_{k=0}^{c-1}(c-k)\frac{\lambda^k(1-z)^k}{k!}e^{-\lambda(1-z)}, \tag{64}$$

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

$$\langle M(\cdot, a., b.)\rangle_t := \left(\int_0^t m_{ij}(s, a_s, b_s)ds\right)_{i,j\in\{1,2\}}, \tag{65}$$

in which

$$m_{11}(t,a,b) := \sum_{k=0}^{c}(c-k)^2 p_k(t+a) - \left(\sum_{k=0}^{c}(c-k)p_k(t+a)\right)^2; \tag{66}$$

$$m_{22}(t,a,b) := \lambda(1-t-a-b) + 2\lambda(1-t-a-b) \times \left(c(\lambda-1) + \sum_{k=0}^{c} p_k(t+a)\right) + m_{11}(t,a,b); \tag{67}$$

$$m_{12}(t,a,b) : = \lambda(1-t-a-b)\left(c(\lambda-1) + \sum_{k=0}^{c} p_k(t+a)\right) - m_{11}(t,a,b). \tag{68}$$

The performance of fluctuation process $\sqrt{N}(A^N - a)$ is illustrated in Fig. 6.

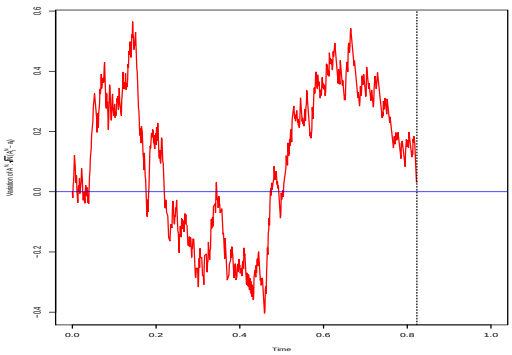

Figure 6: Fluctuation process $\sqrt{N}(A_t^N - a_t), t \in [0; t_0]$ for $N = 1000$, $\lambda = 2$ and $c = 3$.

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 4 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

$$\langle M^N\rangle_t = \begin{bmatrix} \langle M^{N,1}, M^{N,1}\rangle_t & \langle M^{N,1}, M^{N,2}\rangle_t \\ \langle M^{N,2}, M^{N,1}\rangle_t & \langle M^{N,2}, M^{N,2}\rangle_t \end{bmatrix}. \tag{69}$$

From the proof of Lemma 4, we recall the equation (39):

$$\left\| \Delta_t^N - \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} f\big(\frac{n-1}{N}, \frac{A_{n-1}}{N}, \frac{B_{n-1}}{N}\big)\right\| \le \frac{C}{N}, \tag{70}$$

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

$$f_1(t,a) := c - \sum_{k=0}^{c-1}(c-k)p_k(t+a) - 1$$
$$f_2(t,a,b) := (1-t-a-b)\lambda + \sum_{k=0}^{c-1}(c-k)p_k(t+a) - c.$$

and recall the components of the quadratic variation $\langle M^N\rangle_t$ given by (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 $\{\tau_0^N \ge t_0\}$ and that all processes are defined in the time interval $[0,t_0]$, thus all the terms $\mathbf{1}_{A_{i-1}\ge 1}, 1 \le i \le \lfloor Nt_0\rfloor$, $\mathbf{1}_{A_t^N>0}$, $\mathbf{1}_{a_t>0}$ 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.

4.1 Tightness of the process $(W^N)_{N\ge 1}$

Proposition 17. *The sequence $(W^N)_{N\ge1}$ is tight in $\mathcal{D}([0,t_0],\mathbb{R}^2)$.*

Proof. To prove that the distributions of the semi-martingales $(W^N)_{N\ge1}$ form a tight family, we use the Aldous-Rebolledo criterion as in Lemma 6. To achieve this, we start with establishing some moment estimates that will be useful.

Step 1: moment estimates

From (38), we have

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

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

$$\begin{aligned}
|\widetilde{\Delta}_t^{N,1}| \leq \sqrt{N} \left| \Delta_t^{N,1} - \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} \left\{ c - \sum_{k=0}^{c} (c-k) p_k \left(\frac{i-1}{N} + \frac{A_{i-1}}{N} \right) - 1 \right\} \right| \\
+ \sqrt{N} \left| \frac{1}{N} \sum_{i=1}^{\lfloor Nt \rfloor} \left\{ c - \sum_{k=0}^{c} (c-k) p_k \left(\frac{i-1}{N} + \frac{A_{i-1}}{N} \right) - 1 \right\} \right. \\
\left. - \sum_{i=1}^{\lfloor Nt \rfloor} \int_{(i-1)/N}^{i/N} \left(c - \sum_{k=0}^{c} (c-k) p_k \left(s + a_s \right) - 1 \right) ds \right| \\
+ \sqrt{N} \left| \int_{\lfloor Nt \rfloor / N}^{t} \left(c - \sum_{k=0}^{c} (c-k) p_k \left(s + a_s \right) - 1 \right) ds \right|. \qquad (71)
\end{aligned}$$

Thanks to (70), we have that

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

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 (71) is upper bounded by $\frac{\max_{(t,a,b)\in[0,1]^3} |f_1(t,a,b)|}{\sqrt{N}}$.

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

$$\begin{aligned}
\left| p_k(s + a_s) - p_k \left(\frac{i-1}{N} + \frac{A_{i-1}}{N} \right) \right| &\leq \left(\left| s - \frac{i-1}{N} \right| + \left| a_s - A^N_{\frac{i-1}{N}} \right| \right) \sup_{z\in[0,1]} \left| p_k'(z) \right| \qquad (72) \\
&\leq \left(\frac{1}{N} + \left| \frac{W_s^{N,1}}{\sqrt{N}} \right| \right) \sup_{z\in[0,1]} \left| p_k'(z) \right|. \qquad (73)
\end{aligned}$$

The second term in the r.h.s. of (71) is bounded by

$$\begin{aligned}
&\sqrt{N} \sum_{i=1}^{\lfloor Nt \rfloor} \sum_{k=0}^{c} (c-k) \int_{(i-1)/N}^{i/N} \left| p_k(s + a_s) - p_k \left(\frac{i-1}{N} + \frac{A_{i-1}}{N} \right) \right| ds \\
&\qquad \leq \sup_{z\in[0,1]} \left| p_k'(z) \right| \frac{c(c-1)}{2} \left(\frac{1}{\sqrt{N}} + \int_0^t |W_s^{N,1}| ds \right).
\end{aligned}$$

Thus,

$$|\widetilde{\Delta}_t^{N,1}| \leq \frac{C + \max_{(t,a,b)\in[0,1]^3} |f_1(t,a,b)| + \sup_{z\in[0,1]} \left|p_k'(z)\right| \frac{c(c-1)}{2}}{\sqrt{N}} + \sup_{z\in[0,1]} \left|p_k'(z)\right| \frac{c(c-1)}{2} \int_0^t |W_s^{N,1}| ds.$$

Using the similar argument, we have that

$$|\widetilde{\Delta}_t^{N,2}| \leq \frac{C + \sup_{(t,a,b)\in[0,1]^3} |f_2(t,a,b)| + \sup_{z\in[0,1]} \left|p_k'(z)\right| \frac{c(c-1)}{2} + \lambda}{\sqrt{N}} + \left(\sup_{z\in[0,1]} \left|p_k'(z)\right| \frac{c(c-1)}{2} + \lambda \right) \int_0^t |W_s^{N,1}| ds + \lambda \int_0^t |W_s^{N,2}| ds.$$

Hence,

$$\|\widetilde{\Delta}_t^N\| \leq \frac{C'(\lambda,c)}{\sqrt{N}} + C''(\lambda,c) \int_0^t \|W_s^N\| ds \tag{74}$$

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

$$\sup_{t\in[0,t_0]} \mathbb{E}[\|W_t^N\|] \leq (6c^2 + 4\lambda + C'(\lambda,c)) e^{C''(\lambda,c)} = C''', \quad \forall N \geq 1. \tag{75}$$

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 $\delta < \eta\varepsilon (C'(\lambda,c) + 6c^2 + 4\lambda + C''(\lambda,c) C''')^{-1}$,

$$\mathbb{P}\left(\sup_{\substack{|t-s|<\delta \\ 0\leq s<t\leq 1}} \|W_t^N - W_s^N\| > \eta \right) \leq \eta^{-1} \mathbb{E}\left[\sup_{\substack{|t-s|<\delta \\ 0\leq s<t\leq 1}} \|W_t^N - W_s^N\| \right] < \varepsilon. \tag{76}$$

By (75) and (76), we can conclude that $(W^N)_{N\geq 1}$ is tight in $\mathcal{D}([0,t_0], \mathbb{R}^2)$. □

Proposition 18. *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 (33), we have

$$\langle \widetilde{M}^{N,1} \rangle_t = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \left\{ \mathbb{E}\left[(Y_n \wedge c)^2 \left| \mathcal{F}_{n-1} \right.\right] - \left(\mathbb{E}\left[Y_n \wedge c \middle| \mathcal{F}_{n-1} \right] \right)^2 \right\}; \tag{77}$$

$$\begin{aligned} \langle \widetilde{M}^{N,2} \rangle_t = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \Big\{ & \mathrm{Var}(H_n | \mathcal{F}_{n-1}) - 2\Big(\mathbb{E}[H_n (Y_n \wedge c) | \mathcal{F}_{n-1}] \\ & - \mathbb{E}[H_n | \mathcal{F}_{n-1}] \mathbb{E}[Y_n \wedge c | \mathcal{F}_{n-1}] \Big) \Big\} + \langle \widetilde{M}^{N,1} \rangle_t; \end{aligned} \tag{78}$$

$$\begin{aligned} \langle \widetilde{M}^{N,1}, \widetilde{M}^{N,2} \rangle_t = \frac{1}{N} \sum_{n=1}^{\lfloor Nt \rfloor} \Big\{ & \mathbb{E}\left[H_n (Y_n \wedge c) \middle| \mathcal{F}_{n-1} \right] \\ & - \mathbb{E}\left[H_n | \mathcal{F}_{n-1} \right] \mathbb{E}\left[Y_n \wedge c | \mathcal{F}_{n-1} \right] \Big\} - \langle \widetilde{M}^{N,1} \rangle_t \end{aligned} \tag{79}$$

From (77), (26) and (44),

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

From (78), (29), (30) and (44),

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

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 (79), (30),

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

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 $\langle \widetilde{M}^N \rangle_t$ 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 [26], 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 19. *The finite variation* $\left(\widetilde{\Delta}_t^N, t \in [0,t_0] \right)_{N\geq 1}$ *converges in distribution to the process* $(\Delta_t, t \in [0,t_0])$, *which is the unique solution of the stochastic differential*

$$\Delta_t = \int_0^t G(s, a_s, b_s, W_s) dt \tag{80}$$

Proof.

$$\begin{aligned}\widetilde{\Delta}_t^N &= \sqrt{N}\left(\Delta_t^N - \frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor} f\left(\frac{i-1}{N}, A^N_{\frac{i-1}{N}}, B^N_{\frac{i-1}{N}}\right)\right)\\ &\quad + \left(\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}\sqrt{N} f\left(\frac{i-1}{N}, A^N_{\frac{i-1}{N}}, B^N_{\frac{i-1}{N}}\right) - \int_0^t \sqrt{N} f(s,a_s,b_s)ds\right)\\ &= D_t^N + E_t^N, \end{aligned} \tag{81}$$

where

$$\begin{aligned} f(t,a,b) &:= \begin{pmatrix} c - \sum_{k=0}^{c-1}(c-k)\frac{\lambda^k}{k!}(1-t-a)^k e^{-\lambda(1-t-a)} - 1 \\ (1-t-a-b)\lambda - c + \sum_{k=0}^{c}(c-k)\frac{\lambda^k}{k!}(1-t-a)^k e^{-\lambda(1-t-a)} \end{pmatrix}\\ &= \begin{pmatrix} f_1(t,a,b) \\ f_2(t,a,b)\end{pmatrix} \end{aligned} \tag{82}$$

From (70), we have

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

We need to find the limit of E_t^N.

$$E_t^N = \sum_{i=1}^{\lfloor Nt\rfloor}\sqrt{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}}\left(f(\frac{i-1}{N}, A^N_{\frac{i-1}{N}}, B^N_{\frac{i-1}{N}}) - f(s,a_s,b_s)\right)ds - \sqrt{N}\int_{\frac{\lfloor Nt\rfloor}{N}}^{t} f(s,a_s,b_s)ds \tag{83}$$

Because f is continuous function, defined in the compact set $[0,1]^3$, the second term in the r.h.s. of (83) 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 $\phi(z) = c - \sum_{k=0}^{c-1}(c-k)\frac{[\lambda(1-z)]^k}{k!}e^{-\lambda(1-z)}$ and $\psi(z) = \lambda(1-z)$. Then

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

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

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

So the first term in the right hand side of (83) can be written as

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{\lfloor Nt\rfloor}&\begin{pmatrix} W^{N,1}_{\frac{i-1}{N}}\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)\\ -\lambda\left(W^{N,1}_{\frac{i-1}{N}}+W^{N,2}_{\frac{i-1}{N}}\right)-\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)W^{N,1}_{\frac{i-1}{N}}\end{pmatrix}\\
&+\sum_{i=1}^{\lfloor Nt\rfloor}\begin{pmatrix} \int_{\frac{i-1}{N}}^{\frac{i}{N}}\left\{\sqrt{N}(\frac{i-1}{N}-s)\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)\right\}ds\\ -\int_{\frac{i-1}{N}}^{\frac{i}{N}}\left\{\sqrt{N}(\frac{i-1}{N}-s)\left(1+\phi'\left(\frac{i-1}{N}+A^N_{\frac{i-1}{N}}\right)\right)\right\}ds\end{pmatrix}\\
&+\sum_{i=1}^{\lfloor Nt\rfloor}\begin{pmatrix} -\int_{\frac{i-1}{N}}^{\frac{i}{N}}\sqrt{N}\left\{\left((\frac{i-1}{N}-s)+\frac{W^{N,1}_s}{\sqrt{N}}\right)^2\phi''(\xi_{i,s})\right\}ds\\ \int_{\frac{i-1}{N}}^{\frac{i}{N}}\sqrt{N}\left\{\left((\frac{i-1}{N}-s)+\frac{W^{N,1}_s}{\sqrt{N}}\right)^2\phi''(\xi_{i,s})\right\}ds\end{pmatrix}
\end{aligned}\tag{84}$$

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 (84) 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

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

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

$$\widetilde{\Delta}_t=\int_0^t\begin{pmatrix}\phi'(s+a_s)W^1_s\\ -\lambda(W^1_s+W^2_s)-\phi'(s+a_s)W^1_s\end{pmatrix}ds\tag{85}$$

□

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 (62) is proved if the criteria in Theorem 3.1 of [21, 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(62).

References

[1] E. Abbe. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.

[2] F. Ball, T. Britton, C. Larédo, E. Pardoux, D. Sirl, and V. Tran. *Stochastic epidemic models with inference*. Math-Biosciences. Springer, 2019.

[3] P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1968.

[4] P. Billingsley. *Probability and Measure*. John Wiley and Sons, New York, 3 edition, 1995.

[5] B. Bollobás. *Random graphs*. Cambridge University Press, 2 edition, 2001.

[6] B. Bollobás and O. Riordan. Asymptotic normality of the size of the giant component via a random walk. *Journal of Combinatorial Theory Serie B*, 102(1):53–61, Jan. 2012.

[7] S. Clémençon, H. D. Arazoza, F. Rossi, and V. Tran. A statistical network analysis of the hiv/aids epidemics in cuba. *Social Network Analysis and Mining*, 5:Art.58, 2015.

[8] A. Cousien, V. Tran, S. Deuffic-Burban, M. Jauffret-Roustide, J. Dhersin, and Y. Yazdanpanah. Hepatitis C treatment as prevention of viral transmission and level-related morbidity in persons who inject drugs. *Hepatology*, 63(4):1090–1101, 2016.

[9] A. Cousien, V. Tran, M. Jauffret-Roustide, J. Dhersin, S. Deuffic-Burban, and Y. Yazdanpanah. Dynamic modelling of hcv transmission among people who inject drugs: a methodological review. *Journal of Viral Hepatitis*, 22(3):213–229, 2015.

[10] F. W. Crawford, J. Wu, and R. Heimer. Hidden population size estimation from respondent-driven sampling: a network approach. *Journal of the American Statistical Association*, 113(522):755–766, 2018.

[11] R. V. der Hofstad. *Random Graphs and Complex Networks*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

[12] N. Enriquez, G. Faraud, and L. Ménard. Limiting shape of the depth first search tree in an erdős-rényi graph. *Random Structures & Algorithms*, 56(2):501–516, 2020.

[13] J. Jacod and A.N. Shiryaev Limit Theorems for Stochastic Processes. 1987. Springer-Verlag, Berlin.

[14] M. J.-R. et al. Inferring the social network of pwid in paris with Respondent Driven Sampling. 2020. Personnal communication.

[15] D. M. Frost, J. T. Parsons, and J. E. Nanin. Stigma, concealment and symptoms of depression as explanations for sexually transmitted infections among gay men. *Journal of health psychology*, 12(4):636–640, 2007.

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

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

[18] M. Handcock, K. Gile, and C. Mar. Estimating hidden population size using Respondent-Driven Sampling data. *Electronic Journal of Statistics*, 8(1):1491–1521, 2014.

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

[20] M. Hellard, D. A. Rolls, R. Sacks-Davis, G. Robins, P. Pattison, P. Higgs, C. Aitken, and E. McBryde. The impact of injecting networks on hepatitis C transmission and treatment in people who inject drugs. *Hepatology*, 60(6):1861–1870, 2014.

[21] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*, volume 24. North-Holland Publishing Company, 1989. Second Edition.

[22] A. Jakubowski. On the Skorokhod topology. *Annales de l'Institut Henri Poincaré*, 22(3):263–285, 1986.

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

[24] M. Métivier. *Semimartingales: a course on stochastic processes*. de Gruyter, Berlin, New-York, 1982.

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

[26] R. Rebolledo. *La méthode des martingales appliquée à l'étude de la convergence en loi de processus*. Number 62 in Mémoires de la Société Mathématique de France. Société mathématique de France, 1979.

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

[28] O. Robineau, M. Gomes, C. Kendall, L. Kerr, A. Périssé, and P.-Y. Boëlle. Model-based respondent driven sampling analysis for HIV prevalence in brazilian MSM. *Scientific Reports*, 10:2646, 2020.

[29] O. Robineau, A. Velter, F. Barin, and P.-Y. Boelle. HIV transmission and pre-exposure prophylaxis in a high risk MSM population: A simulation study of location-based selection of sexual partners. *PLoS ONE*, 12(11):e0189002, 2017.

[30] D. A. Rolls, R. Sacks-Davis, R. Jenkinson, E. McBryde, P. Pattison, G. Robins, and M. Hellard. Hepatitis c transmission and treatment in contact networks of people who inject drugs. *PLOS ONE*, 8(11):1–15, 11 2013.

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

[32] T. Vo. Chain-referral sampling on Stochastic Block Models. *ESAIM: PS*, 24:718–738, 2020.

[33] T. Vo. *Exploration of random graphs by the Respondent Driven Sampling method*. PhD thesis, Université Sorbonne Paris Nord, Paris, France, 2020.

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