Exploring random graphs by the Respondent-Driven Sampling method

Thuy Vo - Université Paris 13

Septembre 27, 2018

Les rencontres de la Chaire Modélisation Mathémathique et Biodiversité

Motivation

The motivation of the work is modeling an epidemic in a population, which is structured as a network.

Motivation

The motivation of the work is modeling an epidemic in a population, which is structured as a network.

There are populations where the membership involves stigmatized or illegal behaviors, such as a group of drug users, MSM,...

Motivation

The motivation of the work is modeling an epidemic in a population, which is structured as a network.

There are populations where the membership involves stigmatized or illegal behaviors, such as a group of drug users, MSM,...

$\longrightarrow$ It is difficult to access this kind of ”hidden population”.

Motivation

The motivation of the work is modeling an epidemic in a population, which is structured as a network.

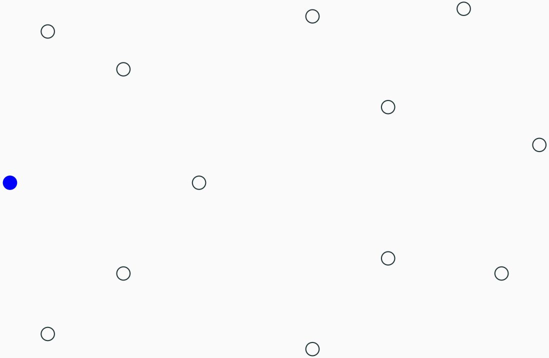
There are populations where the membership involves stigmatized or illegal behaviors, such as a group of drug users, MSM,...

$\longrightarrow$ It is difficult to access this kind of ”hidden population”.

In a research program of AIDS prevention intervention in 1997, Heckathorn [4] introduced the Respondent-Driven Sampling (RDS) method, an efficient way to study hidden population.

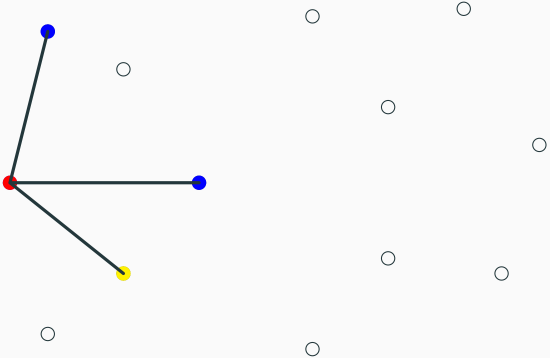
RDS is a peer-to-peer chain, each respondent is asked to name their social contacts and researchers keep track on who refers whom as in network-based samples.

Respondent-Driven Sampling (RDS) method



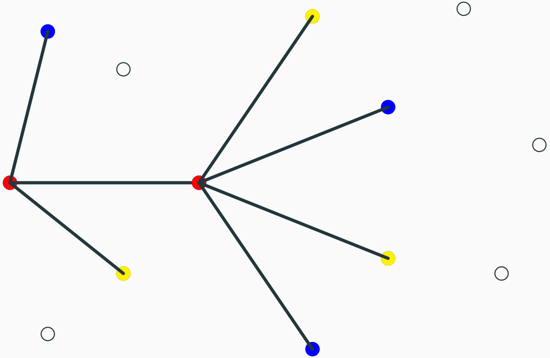
● who has coupon but has not been interviewed

Respondent-Driven Sampling (RDS) method



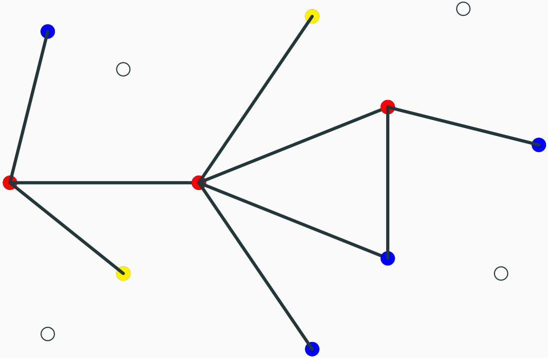
- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

Respondent-Driven Sampling (RDS) method



- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

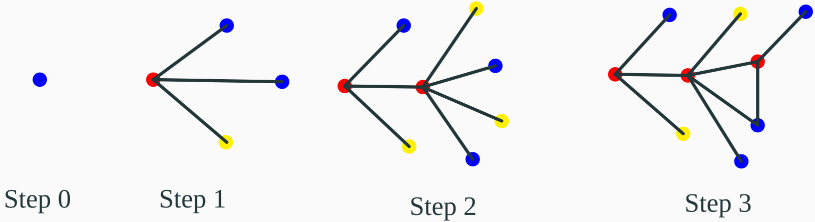
Respondent-Driven Sampling (RDS) method



- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

The Respondent-Driven Sampling (RDS) method

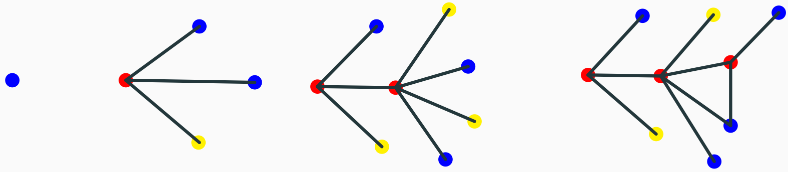
RDS with number of maximum coupons to be delivered is $c = 2$



- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

The Respondent-Driven Sampling (RDS) method

RDS with number of maximum coupons to be delivered is $c = 2$



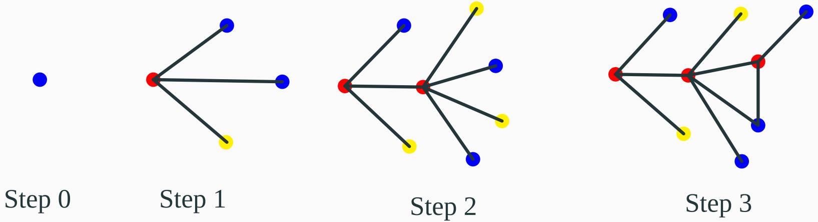
Step 0 Step 1 Step 2 Step 3

- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

The network is progressively discovered when the RDS explores it.

The Respondent-Driven Sampling (RDS) method

RDS with number of maximum coupons to be delivered is $c = 2$



- who has been interviewed
- who has coupon but has not been interviewed
- who has been named but did not receive coupons

The network is progressively discovered when the RDS explores it.

Question: The number of individuals explored by the RDS?

Random networks

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

Examples:

- Erdös-Rényi graphs $G(n, p), 0 < p < 1$ [1]: each edge is included in the graph with probability p independently from every other edge.
- Stochastic block model (SBM) [5]: the set of n vertices is partitioned into m blocks $\{B_1, ..., B_m\}$; for every couple of vertices $u \in B_l$ and $v \in B_k$, the probability of connecting these two points is p_{lk} $(0 < p_{lk} < 1)$.

The RDS method on random graph

Suppose that the population is of size N and is structured by a random graph.

We can associate to the RDS a stochastic process in discrete time.

The RDS method on random graph

Suppose that the population is of size N and is structured by a random graph.

We can associate to the RDS a stochastic process in discrete time.

At the step n:

- A_n = # individuals who have received the coupons but have not been interviewed yet;
- B_n = # individuals who are already explored but have not received any coupon;
- U_n = # individuals who are interviewed up to step n;

Let us consider the process $X_n := (A_n, B_n, U_n)$ in discrete time n. Then how process X_n evolves in time when we let N tends to infinity?

The RDS method on random graph

Suppose that the population is of size N and is structured by a random graph.

We can associate to the RDS a stochastic process in discrete time.

At the step n:

- A_n = # individuals who have received the coupons but have not been interviewed yet;
- B_n = # individuals who are already explored but have not received any coupon;
- U_n = # individuals who are interviewed up to step n;

Let us consider the process $X_n := (A_n, B_n, U_n)$ in discrete time n. Then how process X_n evolves in time when we let N tends to infinity? The normalized process

$$X_t^N := \frac{X_{\lfloor Nt \rfloor}}{N}, \quad t \in [0, 1] \tag{1}$$

The RDS on sparse Erdös-Rényi graph

Assume that the random network we consider is an Erdös-Rényi graph $G(N, \lambda/N)$.

A famous result (in [1]) of Erdös-Rényi graph says that: The Erdös-Rényi graph $G(N, p_N)$ is asymptotically almost surely connected if $p \geq (\log N + \lambda)/N$. Then it is significant to consider $p_N = \lambda/N$ with $\lambda > 1$.

Theorem 1

When N tends to infinity, the process $(X^N)_N = (A^N, B^N)_N$ converges in distribution to a deterministic function in $\mathcal{C}([0,1], \mathbb{R}_+^2)$, which is the unique solution of the differential equations

$$da_t = \left\{ c - \sum_{k=0}^{c-1} (c-k) \frac{[\lambda(1-t-a_t)]^k}{k!} e^{-\lambda(1-t-a_t)} - \mathbb{1}_{a_t>0} \right\} dt \qquad (2)$$

$$db_t = \left\{ (1-t-a_t-b_t) + \sum_{k=0}^{c-1} (c-k) \frac{[\lambda(1-t-b_t)]^k}{k!} e^{-\lambda(1-t-a_t)} \right\} dt \qquad (3)$$

Respondent-Driven Sampling on the Stochastic block model (SBM)

Stochastic block model (SBM) is a more realistic model. It has many applications in community detection in Statistic, network sciences (e.g. [3], [2], [6],...).

Suppose that the network is structured as an SBM with the size is N, the partition of vertices into m blocks is with proportions $\pi = (\pi_1, ..., \pi_m)$ and the probability of connecting vertices between pairs of blocks is defined by the block-matrix $P = (\lambda_{lk}/N)_{l,k\in 1,\dots,m}, (\lambda_{lk} > 0)$.

Process $(X^N)_N$ is written in a $3 \times m$-dimensional form as

$$X_t^N = \begin{pmatrix} A_t^N \\ B_t^N \\ U_t^N \end{pmatrix} = \begin{pmatrix} (A_t^{N,1}, ..., A_t^{N,m}) \\ (B_t^{N,1}, ..., B_t^{N,m}) \\ (U_t^{N,1}, ..., U_t^{N,m}) \end{pmatrix}, \quad t \in [0,1]. \tag{4}$$

Theorem 3

When N tends to infinity, the process $(X_{\cdot}^{N})_N$ converges in distribution to a deterministic vectorial function $x = (x_{\cdot}^{(l)})_{1\leq l\leq m} = (a_{\cdot}^{(l)}, b_{\cdot}^{(l)}, u_{\cdot}^{(l)})_{1\leq l\leq m}$ in $\mathcal{C}([0,1],[0,1]^{3\times m})$, which is the unique solution of the differential equations

$$x_t = \int_0^t f(s, x_s)ds \qquad (5)$$

where $f(s, x_s) := (f_{il}(s, x_s))_{\substack{1\leq i\leq 3\\ 1\leq l\leq m}}$ has the explicit formula

$$f_{1l}(s,x_s) = \sum_{k=1}^{m} \frac{a_s^{(k)}}{|a_s|} \frac{\lambda_s^{k,l}}{\Lambda_s^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right) - \frac{a_s^{(l)}}{|a_s|} \tag{6}$$

$$f_{2l}(s,x_s) = \sum_{k=1}^{m} \frac{a_s^{(k)}}{|a_s|} \mu_s^{k,l} - \sum_{k=1}^{m} \frac{a_s^{(k)}}{|a_s|} \frac{\lambda_s^{k,l}}{\Lambda_s^k} \left(c - \sum_{h=0}^{c} (c-h) \frac{(\Lambda_s^k)^h}{h!} e^{-\Lambda_s^k} \right) \tag{7}$$

$$f_{3l}(s,x_s) = \frac{a_s^{(l)}}{|a_s|} \tag{8}$$

with

$$\lambda_s^{k,l} := \lambda_{kl} \left(\pi_l - a_s^{(l)} - u_s^{(l)} \right); \quad \Lambda_s^k := \sum_{l=1}^{m} \lambda_s^{k,l} \tag{9}$$

$$\text{and} \quad \mu_s^{k,l} := \lambda_{kl}(\pi_l - a_s^{(l)} - b_s^{(l)} - u_s^{(l)}) \tag{10}$$

Remark: When $m = 1$, this result coincides with the Erdös-Rényi case.

Ideas of the proof

- Write the Doob's decomposition of X_t^N;
- Check the tightness of sequence $(X^N)_N$;
- Determine the limiting values of $(X^N)_N$;
- Prove that the ODEs has unique solution.

Sketch of the proof

Define the canonical filtration associated to $(X^N)_N$
$(\mathcal{F}_t^N)_{t\in[0,1]} := (\mathcal{F}_{\lfloor Nt\rfloor})_{t\in[0,1]}$, where $\mathcal{F}_n := \sigma(\{X_i, i \le n\})$.

$$X_t^N = X_0^N + \Delta_t^N + M_t^N,$$

where

$$\Delta_t^N = \begin{pmatrix} \Delta_t^{N,1} \\ \Delta_t^{N,2} \\ \Delta_t^{N,3} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor} \begin{pmatrix} \mathbb{E}[A_n - A_{n-1}|\mathcal{F}_{n-1}] \\ \mathbb{E}[B_n - B_{n-1}|\mathcal{F}_{n-1}] \\ \mathbb{E}[U_n - U_{n-1}|\mathcal{F}_{n-1}] \end{pmatrix}; \tag{11}$$

the square integrable centered martingale $(M_t^N)_t$ has the quadratic variation process $\langle M^N\rangle_t$ given as follow: for every $(l,k) \in \{1,...,m\}^2$,

$$\langle M^{(l),N}, M^{(k),N}\rangle_t = \frac{1}{N^2}\sum_{n=1}^{\lfloor Nt\rfloor} \mathbb{E}\Big[\Big(X_n^{(l)} - \mathbb{E}[X_n^{(l)}|\mathcal{F}_{n-1}]\Big) \times \Big(X_n^{(k)} - \mathbb{E}[X_n^{(k)}|\mathcal{F}_{n-1}]\Big)^T|\mathcal{F}_{n-1}\Big]$$

is a 3×3-matrix, where X is a column vector and X^T is its transpose.

Sketch of the proof

$$A_n = A_{n-1} - I_n + C_n, \tag{12}$$

where

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

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

Z_n is the number of candidates, who are able to be given coupons at step n; $C'_n = (C_n'^{(1)}, ..., C_n'^{(m)})$ having the multivariate hypergeometric distribution with parameters $(m; c, (Z_n^{(1)}, ..., Z_n^{(m)}))$.

Let $(X^N)_N = (A^N, B^N, U^N)_N$ converge to a limiting value $x = (a, b, c)$, we get

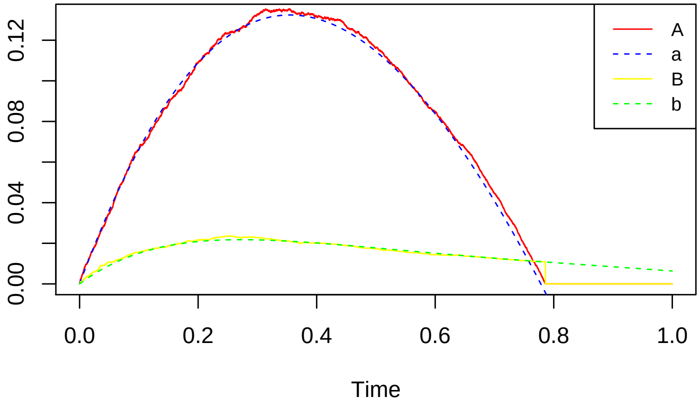
$$\frac{1}{N}\sum_{n=1}^{\lfloor Nt\rfloor}\mathbb{E}[C_n^{(l)}|\mathcal{F}_{n-1}] \xrightarrow{(d)} \int_0^t \left\{\sum_{k=1}^{m}\frac{a_s^{(k)}}{|a_s|}\frac{\lambda_s^{k,l}}{\Lambda_s^k}\left(c-\sum_{h=0}^{c}(c-h)\frac{(\Lambda_s^k)^h}{h!}e^{-\Lambda_s^k}\right)\right\}ds$$

and

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

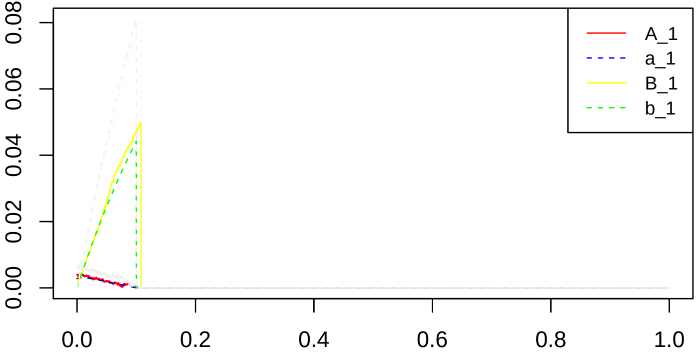
Simulation results

Simulation comparing process $(X^N)_N = (A^N, B^N)_N$ with the solution of ODEs in the case $N = 1000, m = 1, \lambda = 2, c = 3$



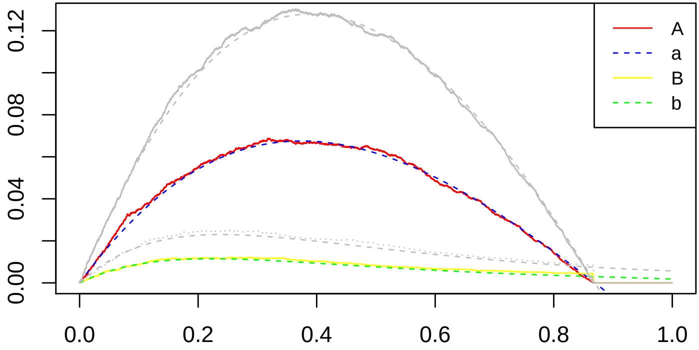
Simulation results

Simulation comparing process $(X^N)_N = (A^N, B^N)_N$ with the solution of ODEs in the case $N = 1000, m = 2, \lambda = (2, 3), \pi = (1/3, 2/3), c = 1$



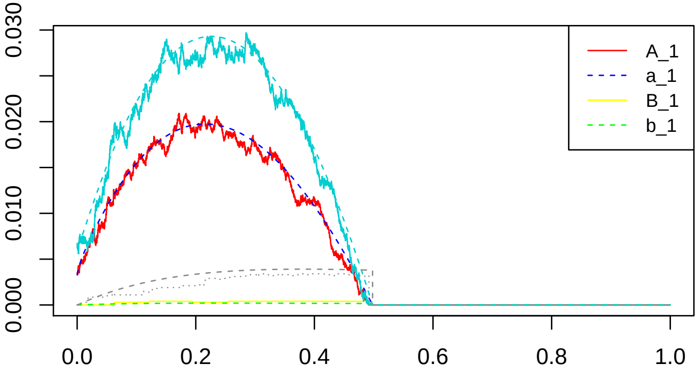
Simulation results

Simulation comparing process $(A^N, B^N)_N$ with the solution of ODEs in the case $N = 1000, m = 2, c = 3, \lambda_{11} = 2, \lambda_{12} = 3, \pi = (1/3, 2/3)$.



Simulation results

Simulation comparing process $(A^N, B^N)_N$ with the solution of ODEs in the case $N = 1000, m = 2, c = 4, \lambda_{11} = 0, \lambda_{12} = 3, \pi = (1/3, 2/3)$.



References i

A. Erdös, P. Rényi.
On random graphs.
Publicationes Mathematicae, (6):290–297, 1959.

S. A. A. O. Gadde, E. E. Gad.
Active learning for community detection in stochastic block models.
arXiv:1605.02372, 2016.

M. Girvan and M. E. J. Newman.
Community structure in social and biological networks.
PNAS, 99(12):7821–7826, June 2002.

References ii

D. D. Heckathorn.
Respondent-driven Sampling: a new approach to the study of hidden populations.
Social Problems, 44:74–99, 1997.

P. W. Holland.
Stochastic block models: First steps.
Social network 5, North-Holland, pages 109–137, 1983.

E. L. A. B.-H. P. Barbillon, S. Donnet.
Stochastic block models for multiplex networks: an application to networks of researchers.
arXiv:1501.06444.