Respondent driven sampling for estimating hard to reach population

Giorgio Alleva, Piero Falorsi, Francesca Petrarca, Paolo Righi

Giornate della Ricerca MEMOTEF, June 2023

Outline

1. Background
2. Respondent Driven Sampling (RDS) data collection strategy
3. The indirect sampling mechanism
4. Unbiased estimators for RDS
    5.1. First estimator (initial random sample)
    5.2. Second estimator (initial non-random sample)
    5.3. CregG estimator accounting for non-coverage

Background

Data gap for SDG indicators on various hard-to-reach or stigmatised populations, such as

- Indigenous populations or ethnic minorities (e.g. the Rohingya in Myanmar)
- Refugees
- HIV's

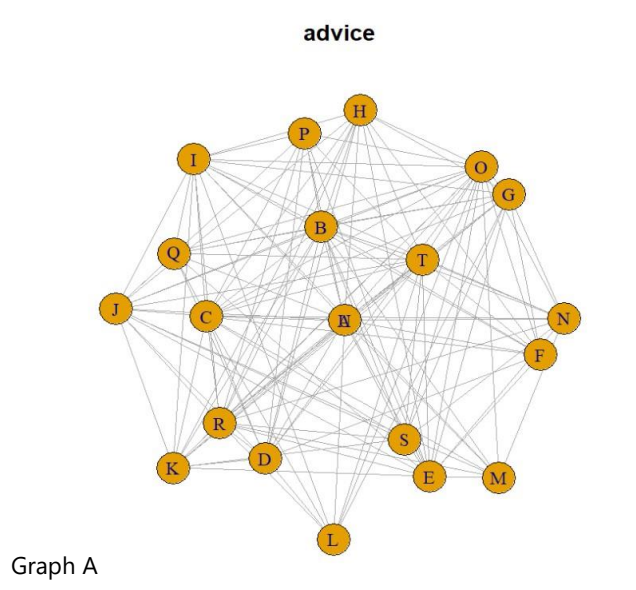
Due to several critical issues that are difficult to overcome in the current context of official statistics in different countries.

It is very complex (and often impossible) to estimate the totals of variables related to those populations through models as in other situations.

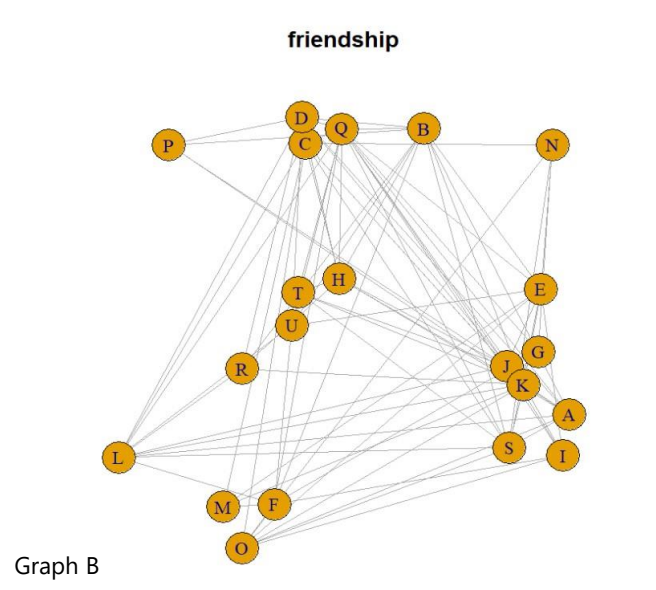
Therefore, defining and implementing sampling strategies that can quickly improve this situation becomes necessary.

The respondent-driven sampling (RDS) method, exploiting existing connections among individuals of the target population, can be a helpful sampling tool to survey these populations.

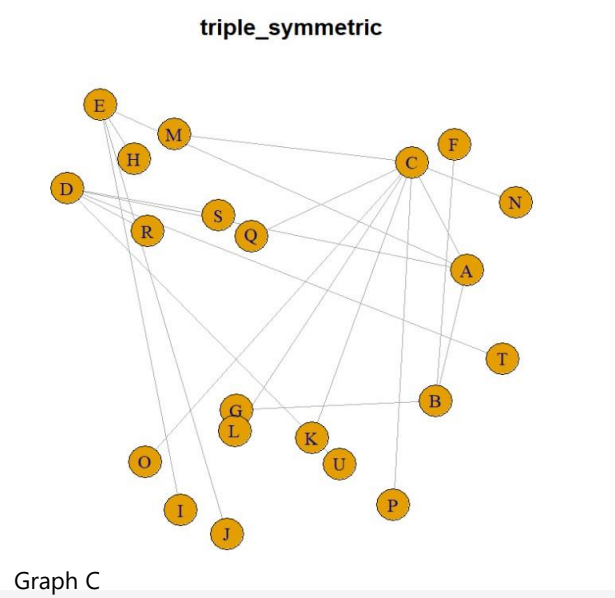
Examples of graphs describing the connections among the people in the population of interest



Graph A



Graph B



Graph C

	Graph A Advice	Graph B Friendship	Graph B Who reports whom
Order: # Nodes (g)	21	21	21
Degree Sum (D)	220	146	40
Size: # Edges (L=D/2)	110	73	20
Maximum Size: g(g-1)/2	210	210	210
Density (2L/g(g-1)	0,524	0,348	0,095
Average arcs per node	10,5	7,0	1,9

The Respondent Driven Sampling (RDS) method (Heckathorn, 1997) is a network-based sampling technique (it includes the Snowball sampling).

Since its establishment, RDS has been employed in countless investigations of such populations across many nations (White et al., 2015).

It starts with a small sample of participants with which the researchers are familiar.
Each participant identifiers their contacts in the target population, enrolling them in the study and increasing the sample size until the sample includes the desired number of respondents.
And so on the sample evolves (adapts) with the progress of the interviews.

While the first selection is generally non-random, the selection of subsequent contacts is by random choice.

Objective of this presentation

The RDS method suffers lack of an estimation methodology that is sufficiently robust concerning varying conditions under which it is applied.

While it is advantageous when estimating mean and proportion values, the accuracy of the total estimates depends on several features, including the nature of the network connecting the individuals in the population (elenco piero).

Below, we address the estimation problem and propose three sample-design unbiased estimation methods by approaching the **RDS method as a particular indirect sampling technique** (Lavallé, 2007).

The **first** method assumes a random sampling of the initial participants.

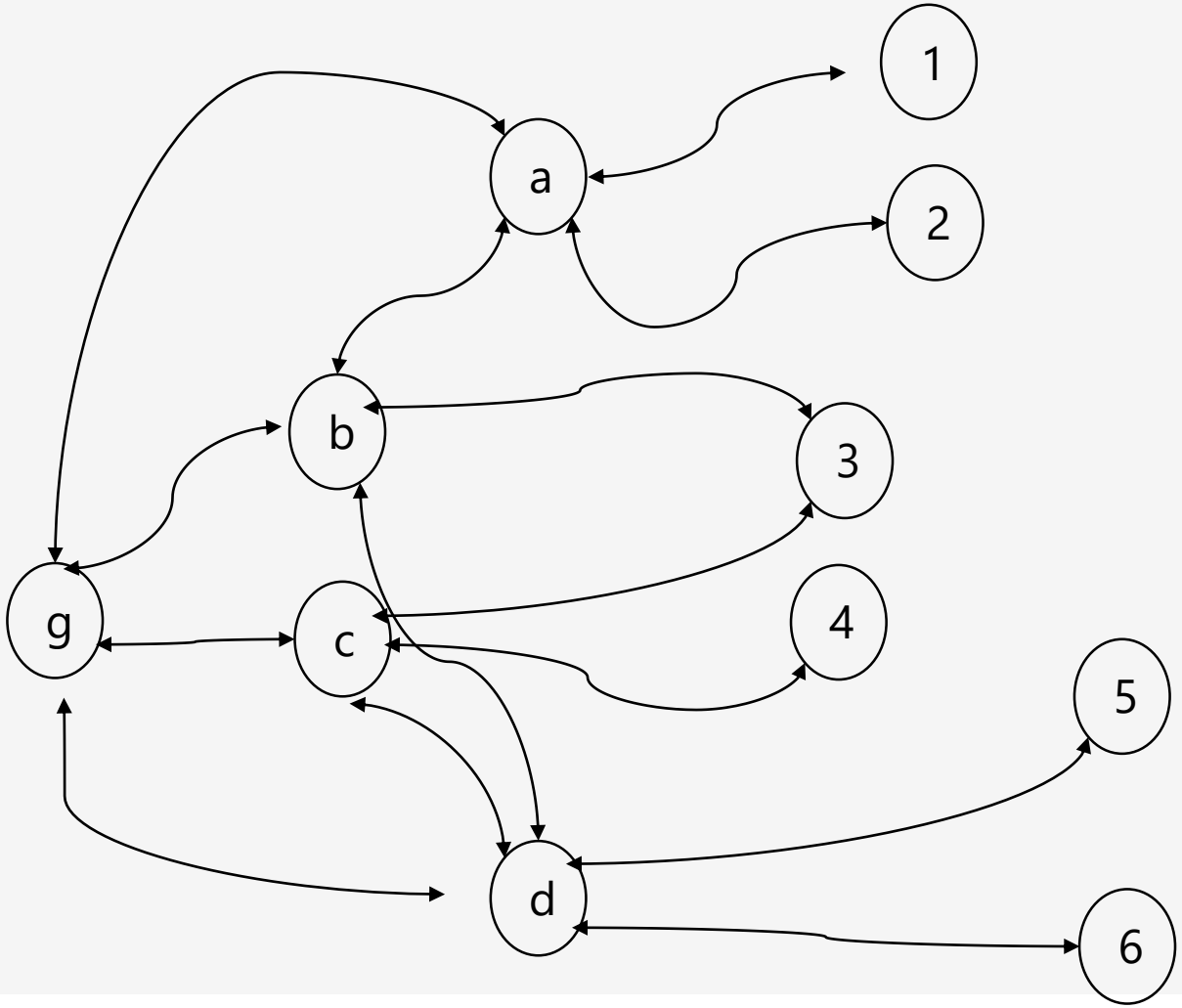
The **second** method, which considers a non random sample selection of the initial participants, as in the original proposal, gives a unbiased estimation of the total number of people connected directly or indirectly to the initial selection.

The **third** method, leveraging the Generalized Capture-Recapture estimation approach (Lavallé and Rivest, 2012), proposes an estimator that accounts for the no-coverage of the two above estimators.

Pierre Lavalle', who joined this research activity, suggests that the two basic samples are non-random but with a different mechanism of undercoverage of the two respondent groups.

RDS data collection

Consider the following graph representing the relationships connecting the units j, a, b, c, d, 1, 2, 3, 4, 5, 6



The relationship between two participants can be **direct** or **indirect**.

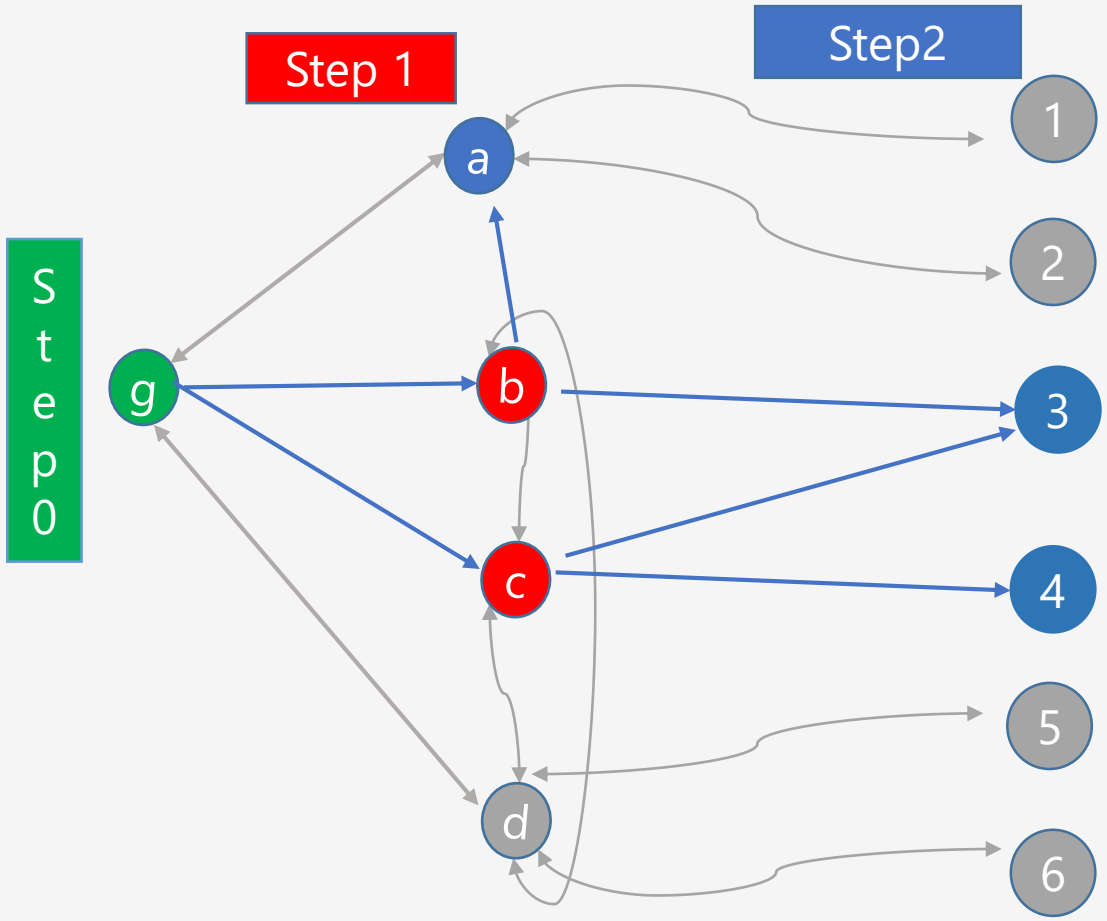
(For example, participant **g** is directly connected to participant **a**; while participant **g** is indirectly related to participant 1 via participant **a**)

Direct relations can be **unidirectional or birectional** (not oriented graph).

(In the sense that if participant **g** knows participant **a**, participant **a** also knows participant **g**).

Data collection through RDS: an example of the mechanism

Example of a network sampling process



At Step 0 we have an initial participant **g**.

At Step 1 two participants are then randomly selecte from the contacts of those progressively included in the sample (**b** and **c**).

At Step 2 two participants are then randomly selected from those connected with **b** (**a** and **3**) and with **c** (**3** and **4**) and included in the sample.

→ Links observed in the sample.

↔ Links not observed in the sample.

Up to and including step 2, participants **g , b , c , a , 3 , 4** are kept in the sample. Participants **d, 1, 2, 5, 6** are not observed.

Stop rule. The RDS process stops either when in the selection process, we encounter only units already identified in the previous steps or at a predetermined step.

The indirect sampling mechanism

In indirect sampling, we have a $\boldsymbol{U^A}$ population of $\boldsymbol{N^A}$ units from which the research starts, and a $\boldsymbol{U^B}$ population of $\boldsymbol{N^B}$ units that constitute the study's target population.

The target parameter

$$\boldsymbol{Y = \sum_{k \in U^B} y_k.}$$

may be viewed as the total

$$\boldsymbol{Y = \sum_{j \in U^A} \bar{y}_j^A} \quad (1)$$

of the population $\boldsymbol{U^A}$ of the variables $\boldsymbol{\bar{y}_j^A}$ where

$$\boldsymbol{\bar{y}_j^A = \sum_{k \in U^B} \frac{\lambda_{j,k}}{L_k^B} y_k}$$

being

$$L_k^B = \sum_{j \in U^A} \lambda_{j,k}$$

the total of direct links ($\lambda_{j,k}$) of Unit $k \in U^B$ with Unit $j \in U^A$.

First estimator: random selection of the initial sample

In each step of the RDS mechanism, an unbiased estimator of the total *Y* can be obtained.

Let $\boldsymbol{S_0}$ be the initial sample of n_0 units selected to start the RDS search, and let $\boldsymbol{S_{0T}}$ be the subset of S_0 including the participants belonging to the target population.

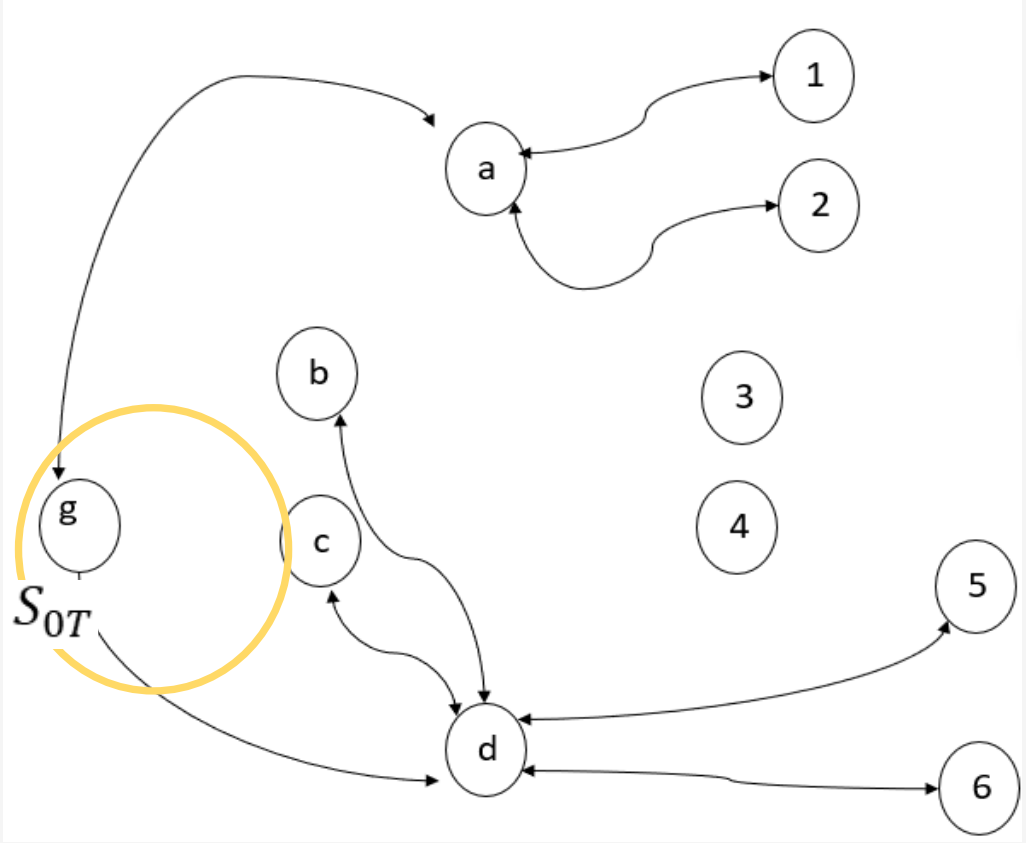
The total *Y* for the target people may be estimated with the standard Horvitz-Thompson estimator:

$$\hat{Y}_0 = \sum_{j \in S_{0T}} y_j \frac{1}{\pi_j}$$

$$= \sum_{k \in S_{0T}} y_k w_k$$

HT sample weight at step 0

where π_j is the inclusion probability.

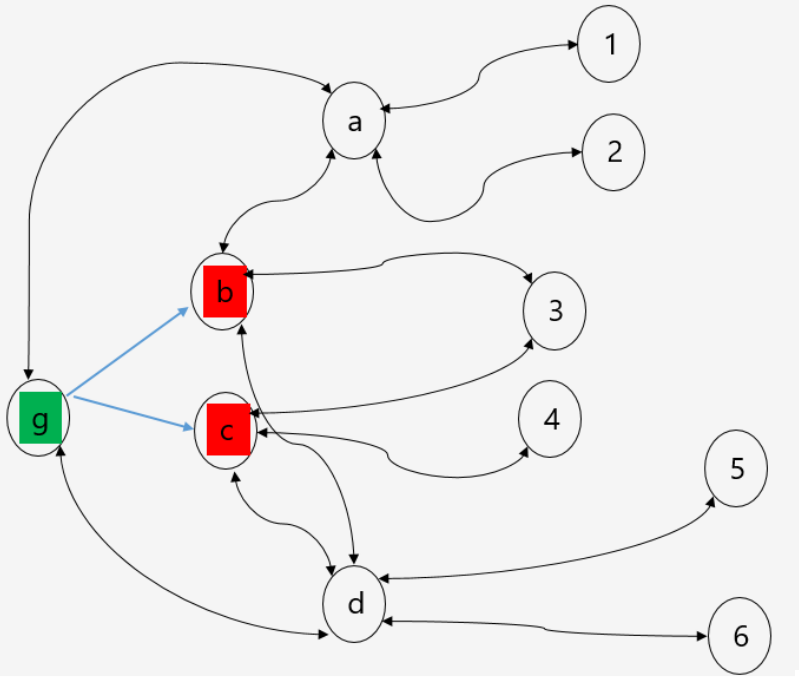


First estimator: step 1

Sample S_1 is formed by taking all the participants of sample S_{0T} plus the set S_1^+ including the participants randomly selected from the links of S_{0T}.

$$\boldsymbol{S_1 = S_{0T} + S_1^+}$$

S_1^+ is formed selecting, independently, $\bar{m}$ units (e.g. 2 or 3) for each unit in S_{0T} from the L_j^A units that are their direct contacts.



The unbiased estimator of *Y* based on the units selected in $\boldsymbol{S_1}$ through the RDS process can be expressed in the standard *weighted form:*

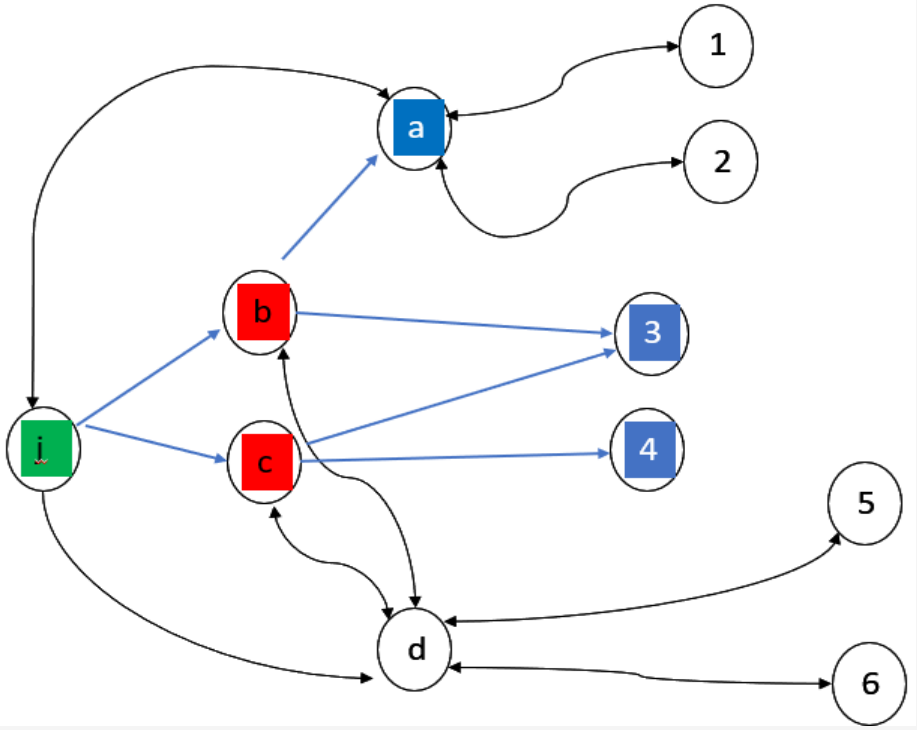
$$\hat{Y}_1 = \sum_{k \in S_1} y_k w_k \text{ , where } w_k = \sum_{j \in S_{0T}} \frac{\lambda_{j,k}}{L_k^B} \left(\frac{1}{\pi_j} \frac{1}{\tau_{k|j \in S_{0T}}} \right) \quad (3)$$

$$\text{and} \quad \tau_{k|j \in S_{0T}} = \begin{cases} 1 & \text{if } j = k \\ \frac{\bar{m}}{L_j^A} & \text{otherwise} \end{cases}$$

First estimator: step 2

We form sample S_2 by taking all the participants of sample S_1, to which we add the sample S_2^+ including the participants randomly selected from the links of S_1^+.

$$S_2 = S_{0T} \cup S_1^+ \cup S_2^+$$



The unbiased estimator in $\boldsymbol{S_2}$ is $\hat{Y}_2 = \sum_{k \in S_2} y_k w_k$

where $w_k = \sum_{j \in S_{0T}} \sum_{j_1 \in S_1} \frac{\lambda_{j,j_1}}{L_{j_1}^B} \frac{\lambda_{j_1,k}}{L_k^B} \left(\frac{1}{\pi_j} \frac{1}{\tau_{j|j \in S_{0T}}} \frac{1}{\tau_{k|j_1 \in S_1}} \right)$

and

$$\tau_{k|j_1 \in S_1} = \begin{cases} 1 & \text{if } j_1 = k \\ \dfrac{\bar{m}}{L_{j_1}^A} & \text{otherwise} \end{cases}$$

First estimator: step r

Continuing the above illustrated process recursively, in the rth step, we form the sample S_r by taking all the participants of sample S_{r-1}, to which we add the participants randomly selected from the links of S_{r-1}^{+}.

The **conditional** probability that unit k is selected in sample S_r, given $j_{r-1} \in S_{r-1}$ is:

$$\tau_{k|j_{r-1}\in S_{r-1}} = \begin{cases} 1 & \text{if } k = j_{r-1} \\ \dfrac{\bar{m}}{L_{j_1}^{A}} & \text{otherwise} \end{cases}$$

The unbiased estimator of Y in $\boldsymbol{S_r}$ is:

$$\hat{Y}_r = \sum_{k\in S_r} y_k w_k$$

where

$$w_k = \sum_{j\in S_{0T}} \dots \sum_{j_{r-1}\in S_{r-1}} \frac{\lambda_{j,j_1}}{L_{j_1}^{B}} \times \dots \times \frac{\lambda_{j_{r-1},k}}{L_k^{B}} \left(\frac{1}{\pi_j} \frac{1}{\tau_{j_1|S_0}} \times \dots \times \frac{1}{\tau_{j_{r-1}|S_{r-2}}} \frac{1}{\tau_{k|S_{r-1}}} \right)$$

Design of sample S_0

The sampling design should maximize the number of observed individuals of the target population in the sample S_0 by adopting proper choices.

The sample design should tend to oversample:

- areas where the researchers have some *a priori* information of a high concentration of the target population;
- on auxiliary variables predictive of membership in the target population.

Second estimator: non-random selection of the initial sample

The S_0 sample is selected in a non-random mode:

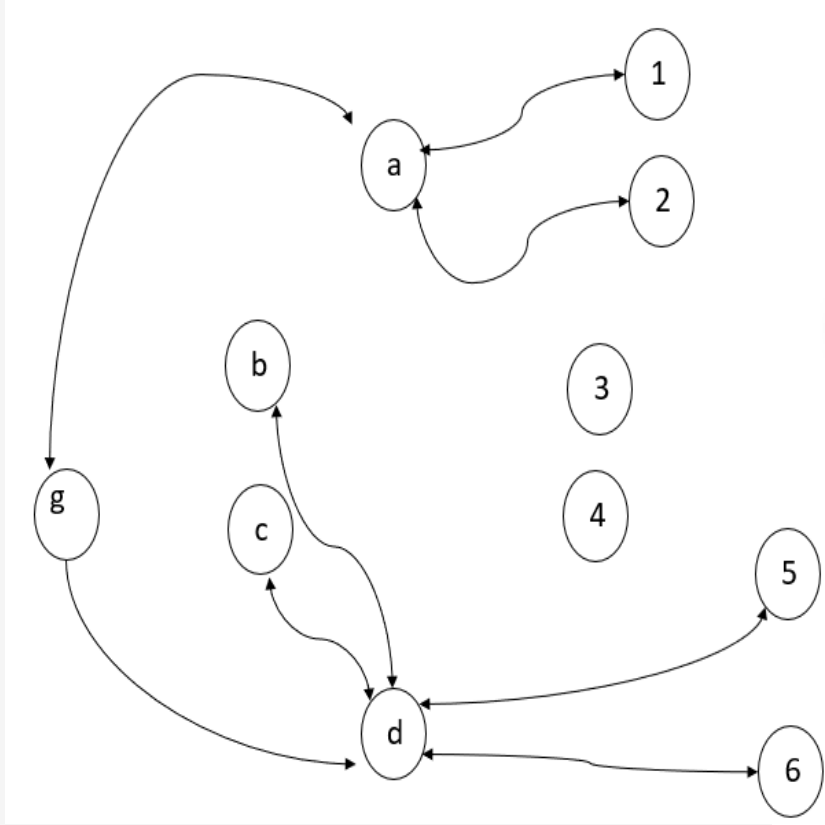
S_0 coincides with S_{0T}.

In this case, we can only obtain a correct estimate of the set of units directly or indirectly connected with the participants of S_0.

We denote this total as $Y_{S_0\rightarrow}$

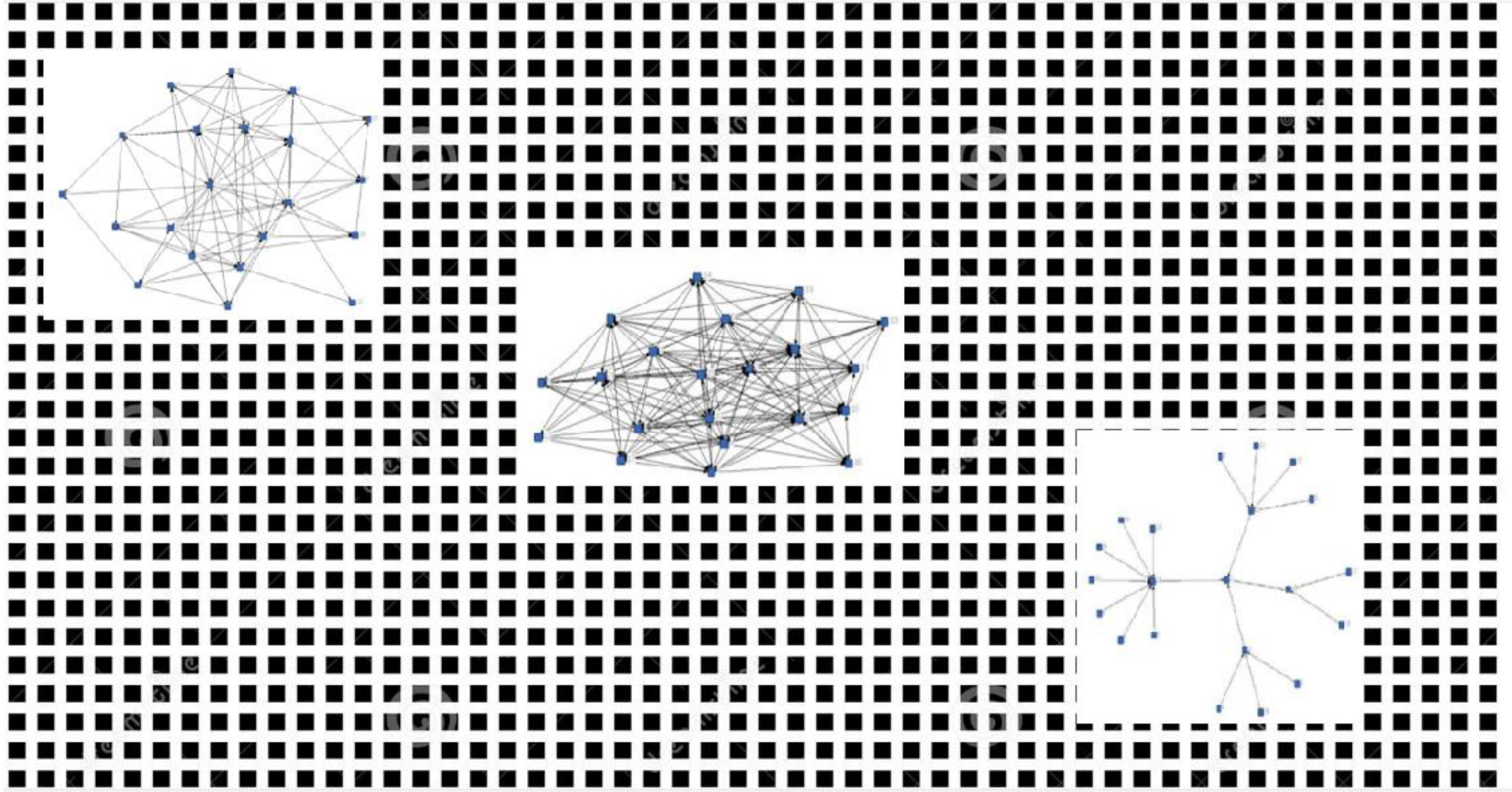
In the example we are considering $Y_{S_0\rightarrow}$ as the sum of the variable y of the units g, a, b, c, d, 1, 2, 5, 6. (excluding 3 and 4).

- If there are clusters that include people of the target population *unconnected* with those in S_0, we have $Y_{S_0\rightarrow} < Y$.
- If the participants of S_0 fall into all *disjointed* clusters in which the population of interest is organised, $Y_{S_0\rightarrow}$ coincides with the total Y.



Example of three groups of separate units

$Y_{S_0\rightarrow} < Y$ if S_0 does not cover all the following three groups



Let r be the step where the RDS process stops.

The unbiased estimator $\hat{Y}_{(S_0)r}$ of $Y_{(S_0)}$ can be obtained as:

$$\hat{Y}_{(S_0)r} = \sum_{k \in S_r} y_k w_{(S_0)k}$$

where $w_{(S_0)k} = \sum_{j \in S_0} \cdots \sum_{j_{r-1} \in S_{r-1}} \frac{\lambda_{j,j_1}}{L_{j_1}^B} \times \cdots \times \frac{\lambda_{j_{r-1},k}}{L_k^B} \left(\frac{1}{\tau_{j_1|j \in S_0}} \times \cdots \times \frac{1}{\tau_{j_{r-1}|j_{r-2} \in S_{r-2}}} \frac{1}{\tau_{k|j_{r-1} \in S_{r-1}}} \right)$

<u>Note:</u> The estimator $\hat{Y}_{(S_0)r}$ is unbiased for $Y_{S_0 \rightarrow}$ if r is greater than the maximum of the shortest paths between any pair of nodes in each cluster of the units of S_0.

Third estimator for dealing under-coverage

Even if the S_0 sample is randomly selected, the *first estimator* $\hat{Y}_r$ may be biased:

under-coverage may occur if respondents do not trust the interviewers and tend to hide their status.

Likewise, if the S_0 sample is non-randomly chosen, the *second estimator* can be affected by under-coverage if total $Y_{S_0\rightarrow}$ does not coincide with Y.

The Generalised Capture-Recature estimator (CReG) (Lavallé and Rivest, 2012), allows us to overcome both of the above mentioned forms of under coverage leveraging on a capture-recapture perspective

$$\hat{Y}_{CReG} = \frac{\hat{Y}_r \times \hat{Y}_{(S_0)r}}{\hat{Y}_{intersect}}$$

where

$$\hat{Y}_{intersect} = \sum_{k \in S_{intersect}} w_k \, w_{(S_0)k} \, y_k$$

where $S_{intersect}$ is the sample that includes the common units in the random and non-random samples.

Pierre Lavalle‘ suggests that the two basic samples are non-random but with a different mechanism of under-coverage of the two respondent groups.

Conclusions

- The disaggregation of data for SDG indicators on hard-to-reach populations presents several critical issues that are difficult to overcome in the current context of official statistics in different countries.
- Therefore, defining and implementing a sampling strategy that can quickly improve this situation becomes necessary. It is helpful to consider sampling designs which maximise the number of observed individuals of the target population.
- The respondent-driven sampling (RDS) method, based on existing connections among individuals of the target population, can be a helpful sampling tool to survey these populations.

Conclusions

In this presentation, we reviewed the RDS method and proposed three sampling unbiased estimators, overcoming the defects of the traditional RDS technique.

What we have presented here represents ongoing research, the initial results of which are encouraging.

The research team is currently running experiments on simulated data and the empirical results will be presented in the final version of our paper.

References

Gile K, Handcock MS. 2015. Network model-assisted inference from respondent-driven sampling data. J. R. Stat. Soc. A 178(3): 619–39.

Gile K, Johnston LG, Salganik MJ. 2015. Diagnostics for respondent-driven sampling. J. R. Stat. Soc. A 178(1):241–69.

Gile K, Beaudry. I S., Handcock M.S. and Miles Q. 2018. Methods for Inference from Respondent-Driven Sampling Data. Annu. Rev. Stat. Appl. 2018. 5:4.1–429

Goodman LA. 1961. Snowball sampling. Ann. Math. Stat. 32:148–70.

Handcock MS, Gile KJ. 2011. Comment: on the concept of snowball sampling. Sociol. Methodol. 41(1):367–71.

Hansen MH, Hurwitz WN. 1943. On the theory of sampling from finite populations. Ann. Math. Stat. 14(4):333–62.

Heckathorn DD. And Cameron, J. 2017. Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling. Annual Review of Sociology · August 2017.

Heckathorn DD. and Jeffri. 2001. Finding the Beat. Using respondent-driven sampling to study jazz musicians. Poetics. 28 2001. P. 307-329. Elsevier.

Heckathorn DD. 2008. Assumptions of RDS: analytic versus functional assumptions. Presented at CDC Consult. Anal. Data Collect. Respond.-Driven Sampl., Atlanta, GA.

Heckathorn DD. 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc. Probl. 49:11–34.

Heckathorn DD. 1997. Respondent driven sampling: a new approach to the study of hidden samples. Soc. Probl. 44(2):174–99.

Johnston LG. 2013. Introduction to HIV/AIDS and sexually transmitted infection surveillance. Module 4. Introduction to respondent-driven sampling. World Health Organ., Geneva. http://www.lisagjohnston. com/respondent-driven-sampling/respondent-driven-sampling.

Johnston LG. 2007. Conducting respondent driven sampling studies in diverse settings: a manual for planning RDS studies. Cent. Dis. Control Prev., Atlanta, GA.

Lavallée P. 2007. Indirect Sampling. Springer. New York.

Lavallée P., Rivest L. P., 2012. Capture–Recapture Sampling and Indirect Sampling. Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 1–27.

Matthew J. Salganik; Douglas D. Heckathorn. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology, Vol. 34. (2004), pp. 193-239.

Salehi M.M. and Seber G. A. F. 2002. Unbiased Estimators for Restricted Adaptive Cluster Sampling. Australian and New Zealand Journal of Statistics. 44. 63-74.

Salehi M.M. and Seber G. A. F 2001. A New Proof of Murthy Estimator which applies to Sequential Sampling. Australian and New Zealand Journal of Statistics. 43. 281-286.

Verdery AM, Merli MG, Moody J, Smith J, Fisher JC. 2015a. Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. Epidemiology 26:661.

Volz E, Heckathorn DD. 2008. Probability based estimation theory for respondent driven sampling. J. Official Statistics. 24:79.

Demonstrations

Demonstration of Formula 1

$$Y = \sum_{j \in U^A} \bar{y}_j^A = \sum_{j \in U^A} \sum_{k \in U^B} \frac{\lambda_{j,k}}{L_k^B} y_k = \sum_{k \in U^B} y_k \sum_{j \in U^A} \frac{\lambda_{j,k}}{L_k^B} = \sum_{k \in U^B} y_k.$$

Demonstration of Formula 2

If unit $k \in S_0$, it is included in S_1 with certainty.

If unit $k \neq j \notin S_0$, it is selected in S_1 independently from every unit $j \in S_0$ with conditional probability

$$\tau_{j|k \in S_0} = \frac{\bar{m}}{L_j^A}.$$

Demonstrations

Demonstration of Formula 3

$$E(\hat{Y}_1) = \sum_{j \in U^A} \frac{1}{\pi_j} E[\delta_j(S_0)] \sum_{k \in U} y_k \frac{\lambda_{j,k}}{L_k^B} \left(\frac{1}{\tau_{k|j \in S_0}} \right) E[\delta_k(S_1) | j \in S_0]$$

$$= \sum_{j \in U^A} \frac{\pi_j}{\pi_j} \sum_{k \in U} y_k \frac{\lambda_{j,k}}{L_k^B} \left(\frac{1}{\tau_{k|j \in S_0}} \right) \tau_{k|j \in S_0}$$

$$= \sum_{j \in U^A} \sum_{k \in U} y_k \frac{\lambda_{j,k}}{L_k^B}$$

where

$\delta_k(A) = 1$ if unit k belongs to set *A* and $\delta_k(A) = 0$.

Demonstrations

Demonstration of Formula 4.

$$E(\hat{Y}_2) = \sum_{j\in U^A}\sum_{j_1\in U^A}\sum_{k\in U^B} y_k \frac{\lambda_{j,j_1}}{L^B_{j_1}}\frac{\lambda_{j_1,k}}{L^B_k}\left(\frac{E[\delta_j(S_0)]}{\pi_j}\frac{E[\delta_{j_1}(S_1|j\in S_0)]}{\tau_{j_1|j\in S_0}}\frac{E[\delta_k(S_2|j_1\in S_1]}{\tau_{k|j_1\in S_1}}\right)$$

$$= \sum_{j\in U^A}\sum_{j_1\in U^A}\sum_{k\in U^B} y_k \frac{\lambda_{j,j_1}}{L^B_{j_1}}\frac{\lambda_{j_1,k}}{L^B_k}\left(\frac{\pi_j}{\pi_j}\frac{\tau_{j_1|j\in S_0}}{\tau_{j_1|j\in S_0}}\frac{\tau_{k|j_1\in S_1}}{\tau_{k|j_1\in S_1}}\right)$$

$$= \sum_{j\in U^A}\sum_{j_1\in U^A}\sum_{k\in U^B} y_k \frac{\lambda_{j,j_1}}{L^B_{j_1}}\frac{\lambda_{j_1,k}}{L^B_k}$$

$$= \sum_{k\in U^B} y_k \sum_{j\in U^A}\frac{\lambda_{j,j_1}}{L^B_{j_1}}\sum_{j_1\in U^A}\frac{\lambda_{j_1,k}}{L^B_k}$$

$$= \sum_{k\in U^B} y_k \sum_{j\in U^A}\frac{\lambda_{j,j_1}}{L^B_{j_1}} 1 = \sum_{k\in U} y_k\, 1\times 1$$

$$= \sum_{k\in U^B} y_k$$