Respondent driven-sampling

Procedure to sample from hidden or hard-to-reach populations

Lucas Moschen

School of Applied Mathematics
Fundação Getulio Vargas

July 7, 2021

Table of Contents

Table of Contents

Hidden and hard-to-reach populations

- No sampling frame exists: unknown size and boundaries;
- Privacy concerns: stigmatized or illegal behavior;
- Fear of exposition or prosecution complicates the enumeration and learning about these populations;
- Examples: Heavy drug users, sex workers, homeless people, and men who have sex with men.

Existing sampling methods

- **Snowball** [Goodman, 1961]
  From starting individuals, each subject provides a list of names of known individuals from the target population. The researcher invites this person to participate, who can agree or deny it.

- **Key informant** [Deaux and Callaghan, 1985]
  Expert respondents are selected to answer about target population's behavior. For instance, social workers, drug abuse counselors, public health officials, etc.

- **Targeted** [Watters and Biernacki, 1989]
  Field researchers build an ethnographic mapping of the target population, and recruit a number of individuals at the identified site.

Existing sampling methods

- **Snowball** [Goodman, 1961]
  From starting individuals, each subject provides a list of names of known individuals from the target population. The researcher invites this person to participate, who can agree or deny it.

- **Key informant** [Deaux and Callaghan, 1985]
  Expert respondents are selected to answer about target population's behavior. For instance, social workers, drug abuse counselors, public health officials, etc.

- **Targeted** [Watters and Biernacki, 1989]
  Field researchers build an ethnographic mapping of the target population, and recruit a number of individuals at the identified site.

Existing sampling methods

- **Snowball** [Goodman, 1961]
  From starting individuals, each subject provides a list of names of known individuals from the target population. The researcher invites this person to participate, who can agree or deny it.

- **Key informant** [Deaux and Callaghan, 1985]
  Expert respondents are selected to answer about target population's behavior. For instance, social workers, drug abuse counselors, public health officials, etc.

- **Targeted** [Watters and Biernacki, 1989]
  Field researchers build an ethnographic mapping of the target population, and recruit a number of individuals at the identified site.

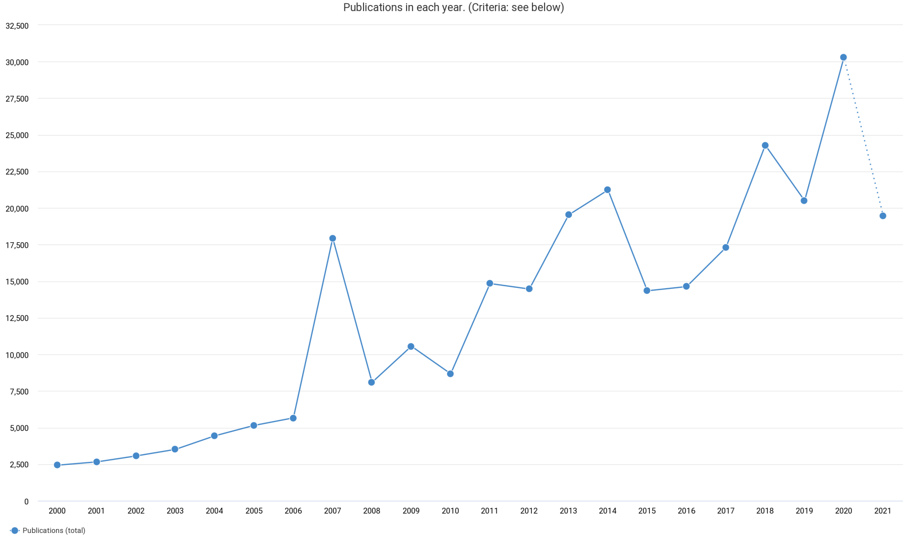
Problems with snowball sampling

- Inferences about the individuals depend on the initial sample. [Frank and Snijders, 1994] recommended beginning with ethnographic mapping;
- Bias towards individuals who are more cooperative (volunteerism);
- Bias by masking, that is, protecting friends by not referring them;
- Individuals with more links may be oversampled.

Respondent-driven sampling

- Proposed by [Heckathorn, 1997] as an approach to estimate proportions in a hard-to-reach population;
- Theory based on Markov chains;
- [Crawford, 2016] models as an interaction network and defines a probability distribution over the observed subgraph;
- The sampling is without replacement.

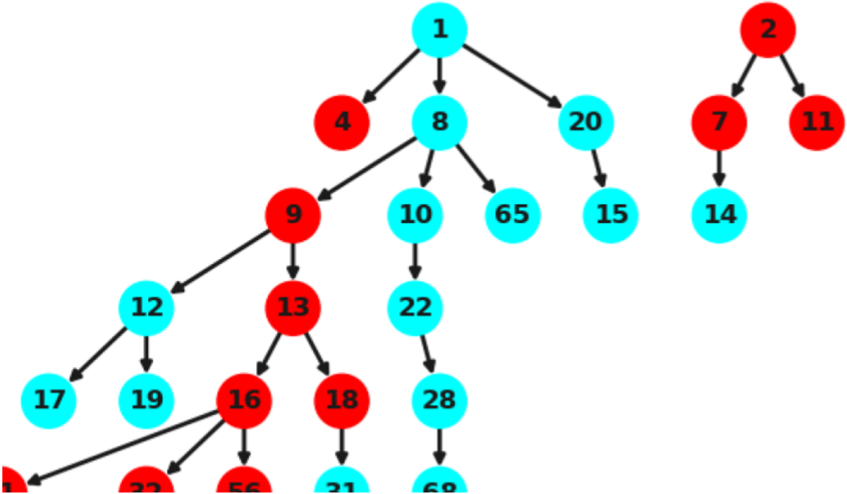
Respondent-driven sampling



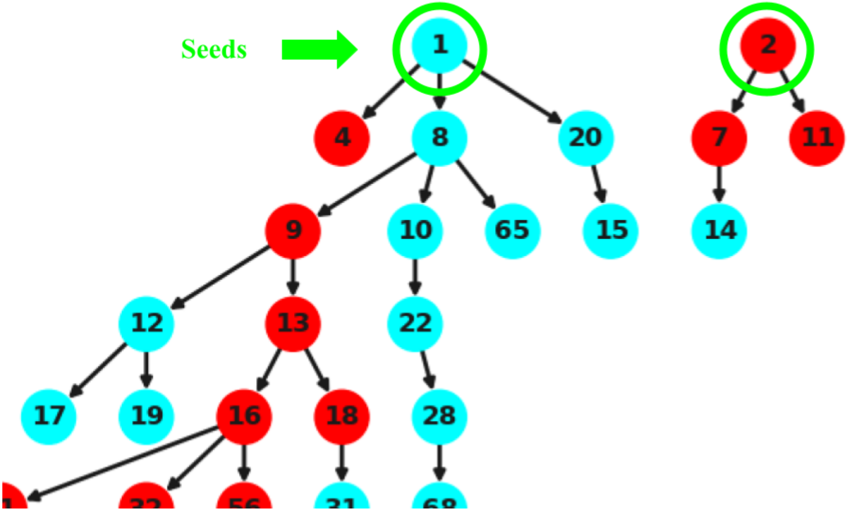
Source: https://app.dimensions.ai
Exported: July 07, 2021
Criteria: Text - 'respondent-driven sampling' in full data.

© 2021 Digital Science and Research Solutions Inc. All rights reserved. Non-commercial redistribution / external re-use of this work is permitted subject to appropriate acknowledgement. This work is sourced from Dimensions® at www.dimensions.ai.

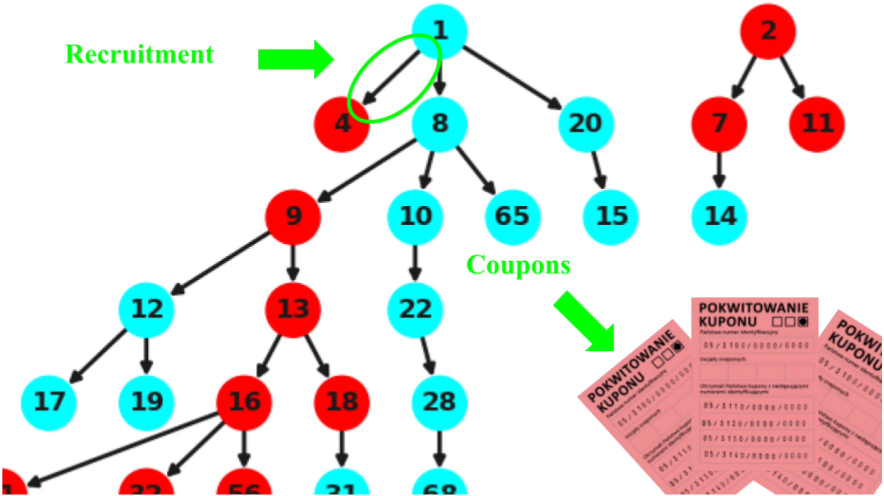
Respondent-driven sampling



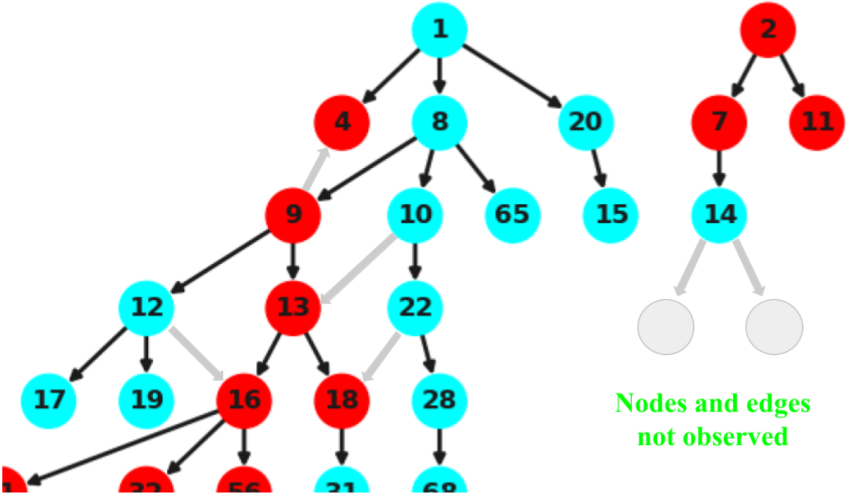
Respondent-driven sampling



Respondent-driven sampling



Respondent-driven sampling



Dual system of incentives

Two different sources of theoretical incentive (dual incentive system):

- **Individual-sanction based control:** reward for participating in the research.

- **Group-mediated social control:** reward for recruiting peers. When social approval is important, it's more efficient and cheaper. Material incentive can be transformed into symbolic incentive.

Dual system of incentives

Two different sources of theoretical incentive (dual incentive system):

- **Individual-sanction based control:** reward for participating in the research.
- **Group-mediated social control:** reward for recruiting peers. When social approval is important, it's more efficient and cheaper. Material incentive can be transformed into symbolic incentive.

Table of Contents

Formal model

The RDS can be built mathematically with different approaches:

- **Markov process** [Heckathorn, 1997]
  The recruiter's social characteristics affect the characteristics of the recruits. A limited number of states can be assumed, and the recruits are a function of the recruiter's characteristics.

- **Graphical structure** [Crawford, 2016]
  A hidden population is an undirected graph, and we observe it partially in the *recruitment graph*, as also the coupon matrix and recruitment times. The unobserved graph is treated as *missing data* and can be interpreted as an Exponential Random Graph Model.

Formal model

The RDS can be built mathematically with different approaches:

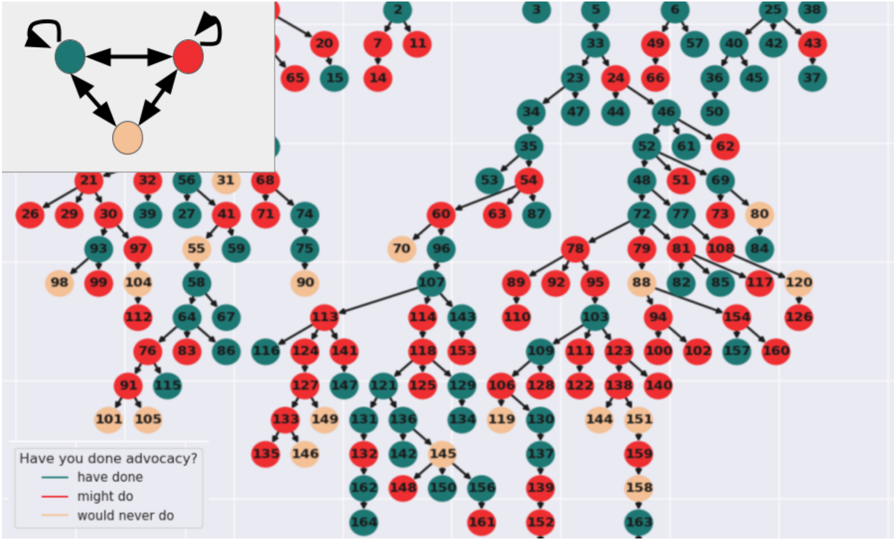
- **Markov process** [Heckathorn, 1997]
  The recruiter's social characteristics affect the characteristics of the recruits. A limited number of states can be assumed, and the recruits are a function of the recruiter's characteristics.

- **Graphical structure** [Crawford, 2016]
  A hidden population is an undirected graph, and we observe it partially in the *recruitment graph*, as also the coupon matrix and recruitment times. The unobserved graph is treated as *missing data* and can be interpreted as an Exponential Random Graph Model.

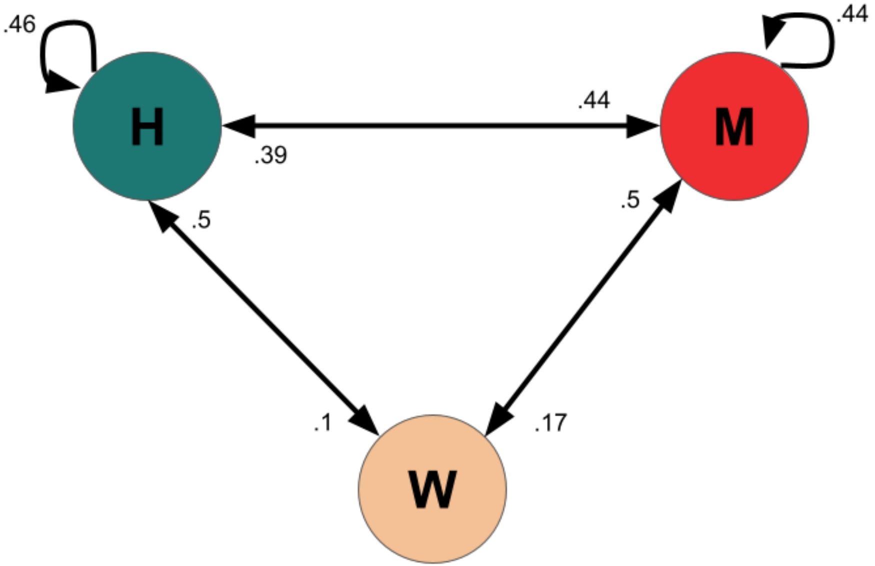
Markov chain model

- In a survey, questions create states describing the participant;
- Heckathorn concluded that the recruitment was a first-order memoryless process (first-order Markov process).
- The Markov chain indicates the most recent recruit's characteristic;
- The Markov chain must be ergodic.

Markov chain model



Markov chain model



Consequences of Markov chain theory

Theorem

An equilibrium mix of recruits will be attained when the number of waves goes to infinity, and it is independent from which recruitment began. The pooling approaches the equilibrium in a geometric rate.

Convergence analysis

Assessing bias in RDS

- Inbreeding bias event: there is a positive probability of the subject recruit from in-group with certainty. This is also called *homophily*.

- The paper's conclusion is that RDS produces unbiased samples if the inbreeding bias event is equal for all groups.

Evolution of RDS

1. [Heckathorn, 2002] extended the model considering symmetric relations, that is, if A relates to B, B relates to A. The model was called *reciprocity model*. Self-reported degrees were also added to the model.

2. Bootstrap was used to estimate standard deviation of the estimations. The ideia is to use the transition matrix to generate Bootstrap Markov chains. [Salganik, 2006] improved the variance estimator.

3. Under some regularity conditions, population estimates are asymptotically unbiased [Salganik and Heckathorn, 2004].

4. The RDS II estimator and an analytical variance estimation is presented [Volz and Heckathorn, 2008].

Evolution of RDS

1. [Heckathorn, 2002] extended the model considering symmetric relations, that is, if A relates to B, B relates to A. The model was called *reciprocity model*. Self-reported degrees were also added to the model.
2. Bootstrap was used to estimate standard deviation of the estimations. The ideia is to use the transition matrix to generate Bootstrap Markov chains. [Salganik, 2006] improved the variance estimator.
3. Under some regularity conditions, population estimates are asymptotically unbiased [Salganik and Heckathorn, 2004].
4. The RDS II estimator and an analytical variance estimation is presented [Volz and Heckathorn, 2008].

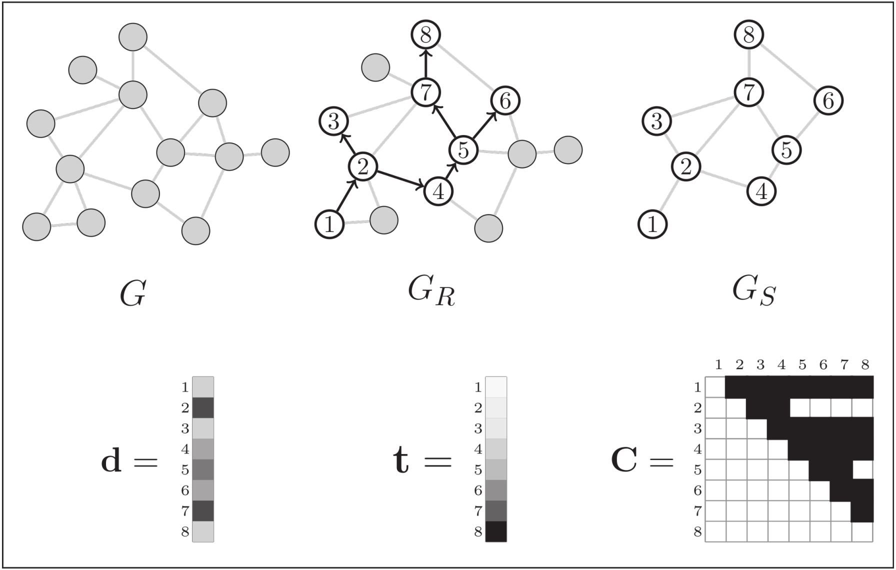
Evolution of RDS

1. [Heckathorn, 2002] extended the model considering symmetric relations, that is, if A relates to B, B relates to A. The model was called *reciprocity model*. Self-reported degrees were also added to the model.
2. Bootstrap was used to estimate standard deviation of the estimations. The ideia is to use the transition matrix to generate Bootstrap Markov chains. [Salganik, 2006] improved the variance estimator.
3. Under some regularity conditions, population estimates are asymptotically unbiased [Salganik and Heckathorn, 2004].
4. The RDS II estimator and an analytical variance estimation is presented [Volz and Heckathorn, 2008].

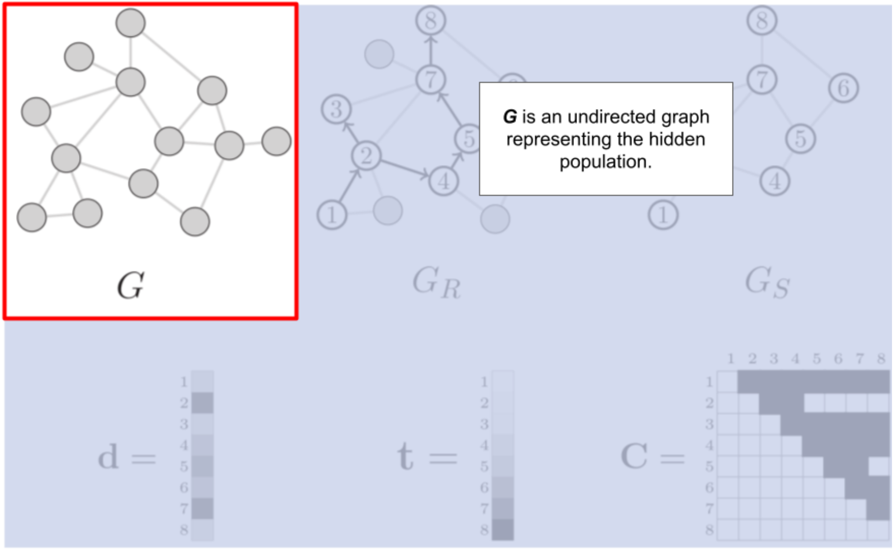
Evolution of RDS

1. [Heckathorn, 2002] extended the model considering symmetric relations, that is, if A relates to B, B relates to A. The model was called *reciprocity model*. Self-reported degrees were also added to the model.
2. Bootstrap was used to estimate standard deviation of the estimations. The ideia is to use the transition matrix to generate Bootstrap Markov chains. [Salganik, 2006] improved the variance estimator.
3. Under some regularity conditions, population estimates are asymptotically unbiased [Salganik and Heckathorn, 2004].
4. The RDS II estimator and an analytical variance estimation is presented [Volz and Heckathorn, 2008].

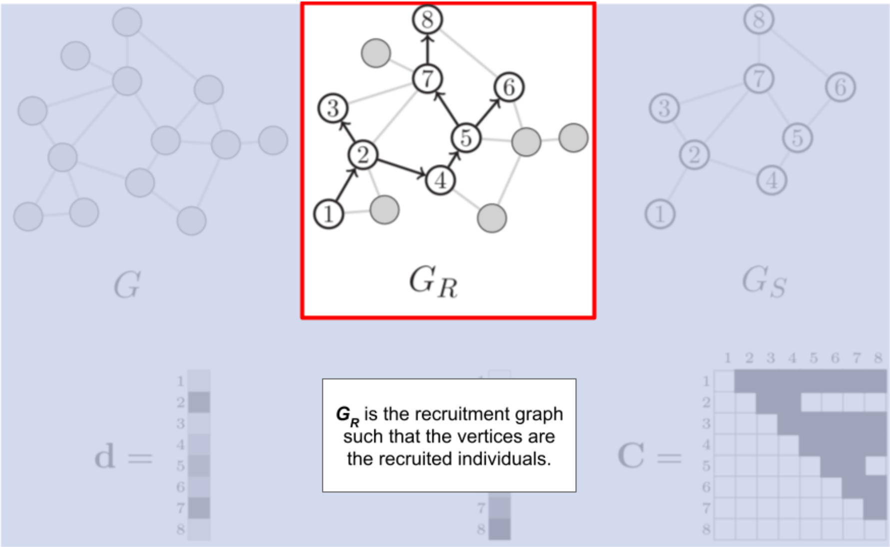
Network model



Network model

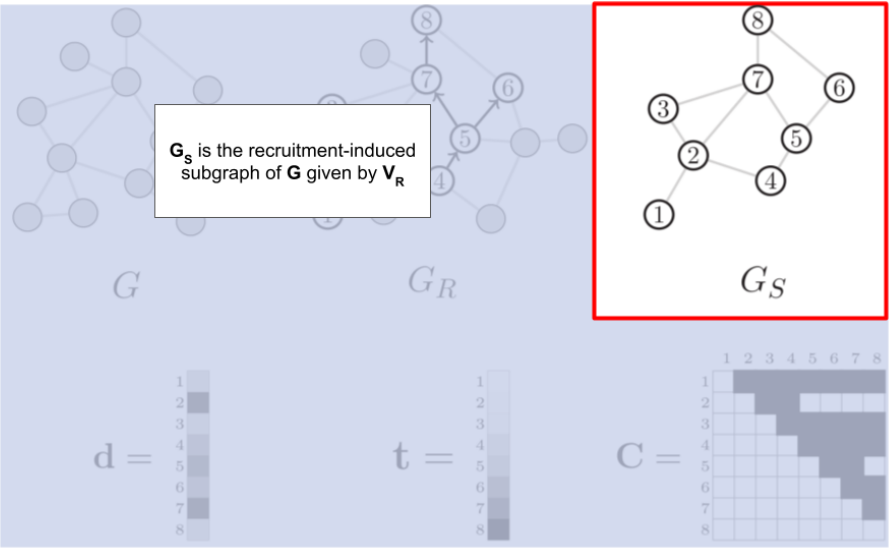


Network model

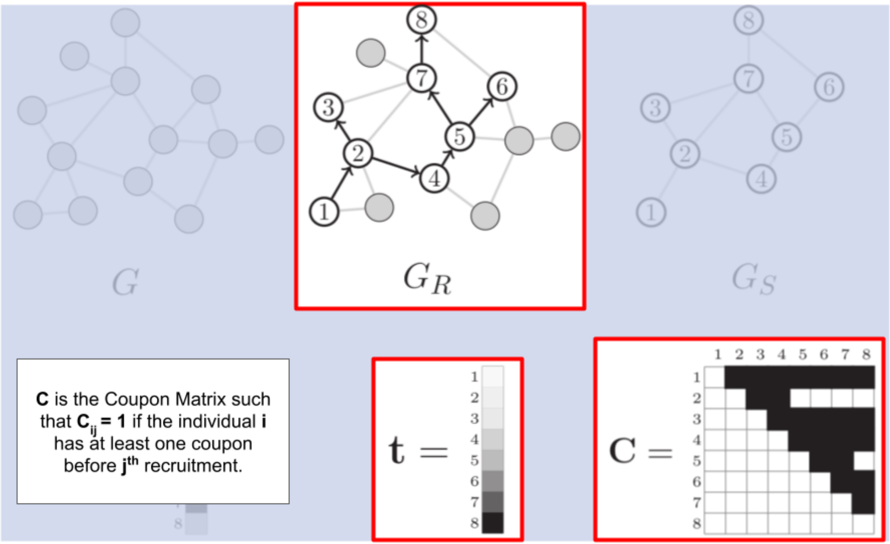


G_R is the recruitment graph such that the vertices are the recruited individuals.

Network model

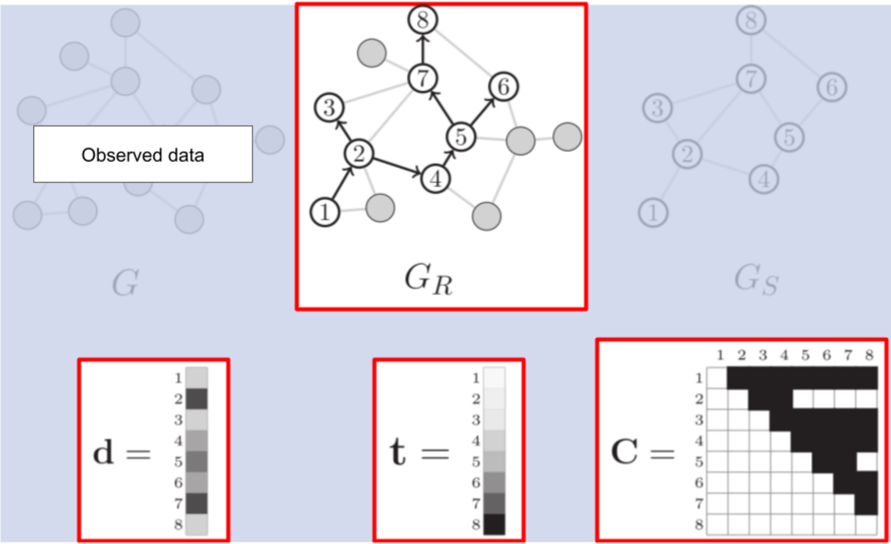


Network model



C is the Coupon Matrix such that $\mathbf{C}_{ij}$ **= 1** if the individual **i** has at least one coupon before **j**$^{\text{th}}$ recruitment.

Network model



Consequences

The time to recruitment along a *susceptible edge* has Exponential distribution, independent of the identity, neighbor, and all the other waiting times.

Theorem (Waiting time for a recruitment)

Let u be a recruiter and $v \in S_u$ a susceptible neighbor. The waiting time to u recruit v conditioned on the recruitment event has Exponential distribution with rate $\lambda |S_u|$. The probability of $v \in S_u$ to be the next recruited is uniform.

Theorem (Waiting time for some recruitment to occur)

The waiting time to the next recruitment is distributed as Exponential with rate $\lambda \sum_{u \in R} |S_u|$.

Likelihood of the recruitment time series

Compatibility

An estimated subgraph $\hat{G}_S = (\hat{V}_S, \hat{E}_S)$ is *compatible* with the data if:

1. The vertices in the estimated subgraph are the same as the observed vertices;
2. Each recruitment edge is an undirected edge of the estimated subgraph;
3. For all $v \in V_R, \sum_{u \in V_R/\{v\}} \mathbb{1}_{\hat{E}_S}(\{u, v\}) \leq d_v$.

Likelihood of the recruitment time series

Let A be the adjacency matrix of a compatible estimated subgraph, that is,

$$[A]_{ij} = 1 \text{ iff } \{i, j\} \in G_S.$$

Then

$$[AC]_{ij} = \sum_k [A]_{ik}[C]_{kj} = \sum_k \mathbb{1}(\{i,k\} \in G_S \text{ and } k \text{ can recruit in } t_j),$$

that is, the number of recruiters connected to i just before the j^{th} recruitment, when $j \leq i$. Let u_i be the number of edges linking the sampled node i with others not sampled. Then,

$$[C^T u]_i = \sum_k [C]_{ki} u_k = \sum_k \mathbb{1}(k \text{ can recruit in } t_i) \cdot \#\text{susceptible edges of } k$$

Likelihood of the recruitment time series

The likelihood of the recruitment time series $w = (0, t_2 - t_1, ..., t_n - t_{n-1})$ is

$$L(w|G_S, \lambda) = \left(\prod_{k \text{ isn't seed}} \lambda s_k\right) \exp(-\lambda s^T w),$$

where

$$s = \text{tril}(AC)^T 1 + C^T u$$

indicates the number of susceptible edges just before each recruitment.

Likelihood of the recruitment time series

Setting $T(A) = -\lambda s$ and $B(A) = \sum_{k \text{ isn't seed}} \log(\lambda s_k)$, the likelihood from above can be normalized to obtain the probability

$$P(A|w) \propto \exp\left[T(A)^T w + B(A)\right]$$

which can be interpreted as an Exponential Random Graph Model.

Reconstruction of the recruitment-induced subgraph

The ideia is to sample from

$$p(G_S, \lambda | G_R, C, d, t) \propto L(w | G_S, \lambda) P(G_S) \pi(\lambda),$$

where $P(G_S)$ and $\pi(\lambda)$ are the prior distributions. For instance, it can be taken uniformly over the compatible subgraphs. A Metropolis-within-Gibbs sampling scheme is used to draw pairs (G_S, λ).

Table of Contents

Applications

1. HIV prevalence estimation [Gile, 2011]: improved RDS II and applied to three different sites.

2. Sampling for understanding: Jazz musicians in New York and San Francisco. Comparison with data from American Federation of Musicians [Salganik and Heckathorn, 2004].

3. Hidden population size estimation [Crawford et al., 2018].

Hidden population size estimation

- Let $\mathcal{G}$ be a Erdős-Rényi random graph with probability p. Then the degree of vertex i is $d_i \sim \text{Bin}(N-1, p)$. Assume the hidden population has this distribution.

- The probability of recruitment depends only on the edges it shares with recruiters. This allows the construction of $L(N, p; G_S, Y)$.

- Consider the likelihood of the recruitment time series.

- Establishing priors $\pi(N), \pi(p), \pi(G_S)$, and $\pi(\lambda)$. Then, the paper obtains the marginal distribution of N given the data

$$P(N, p | G_S, G_R, C, d, t) \propto L(N, p; G_S, G_R, C, d, t)\pi(N)\pi(p).$$

Hidden population size estimation: algumas conclusões

- Erdos-Rényi model has proven to be empirically useful in a wide variety of population size estimation applications;
- RDS was not designed for population size estimation, and it should not be used for this purpose.

Table of Contents

Evaluation

- [Goel and Salganik, 2009] High-homophily breakpoint, when the probability of within-group recruitment is high. The convergence requires many more waves;
- [McCreesh et al., 2012] concluded that only a third of the RDS estimates were closer to the true proportions. The Bootstrap intervals were underestimated. This influenced [Baraff et al., 2016] to improve the method;
- The estimates validity depends on multiple assumptions that frequently do not hold in the field;
- [Shi et al., 2019] shows how the model can be adjusted to specific cases to reduce bias.

References I

Baraff, A. J., McCormick, T. H., and Raftery, A. E. (2016).
Estimating uncertainty in respondent-driven sampling using a tree bootstrap method.
Proceedings of the National Academy of Sciences, 113(51):14668–14673.

Crawford, F. W. (2016).
The graphical structure of respondent-driven sampling.
Sociological Methodology, 46(1):187–211.

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

References II

Deaux, E. and Callaghan, J. W. (1985).
Key informant versus self-report estimates of health-risk behavior.
Evaluation Review, 9(3):365–368.

Frank, O. and Snijders, T. (1994).
Estimating the size of hidden populations using snowball sampling.
JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-, 10:53–53.

Gile, K. J. (2011).
Improved inference for respondent-driven sampling data with application to hiv prevalence estimation.
Journal of the American Statistical Association, 106(493):135–146.

Goel, S. and Salganik, M. J. (2009).
Respondent-driven sampling as markov chain monte carlo.
Statistics in medicine, 28(17):2202–2229.

References III

Goodman, L. A. (1961).
Snowball Sampling.
The Annals of Mathematical Statistics, 32(1):148–170.

Heckathorn, D. D. (1997).
Respondent-driven sampling: A new approach to the study of hidden populations.
Social Problems, 44(2):174–199.

Heckathorn, D. D. (2002).
Respondent-driven sampling ii: deriving valid population estimates from chain-referral samples of hidden populations.
Social problems, 49(1):11–34.

References IV

McCreesh, N., Frost, S., Seeley, J., Katongole, J., Tarsh, M. N., Ndunguse, R., Jichi, F., Lunel, N. L., Maher, D., Johnston, L. G., et al. (2012).
Evaluation of respondent-driven sampling.
Epidemiology (Cambridge, Mass.), 23(1):138.

Salganik, M. J. (2006).
Variance estimation, design effects, and sample size calculations for respondent-driven sampling.
Journal of Urban Health, 83(1):98.

Salganik, M. J. and Heckathorn, D. D. (2004).
Sampling and estimation in hidden populations using respondent-driven sampling.
Sociological methodology, 34(1):193–240.

References V

Shi, Y., Cameron, C. J., and Heckathorn, D. D. (2019).
Model-based and design-based inference: reducing bias due to differential recruitment in respondent-driven sampling.
Sociological Methods & Research, 48(1):3–33.

Volz, E. and Heckathorn, D. D. (2008).
Probability based estimation theory for respondent driven sampling.
Journal of official statistics, 24(1):79.

Watters, J. K. and Biernacki, P. (1989).
Targeted sampling: Options for the study of hidden populations.
Social Problems, 36(4):416–430.