Network Size: Measurement and Errors in Respondent-Driven Sampling

JPSM/MPSDS Seminar

Ai Rene Ong, Yibo Wang

March 8, 2023

Study II

A Latent Variable Model for Individual Degree Estimation in Respondent-Driven Sampling

Motivation

Individual degree is a crucial factor in RDS analysis

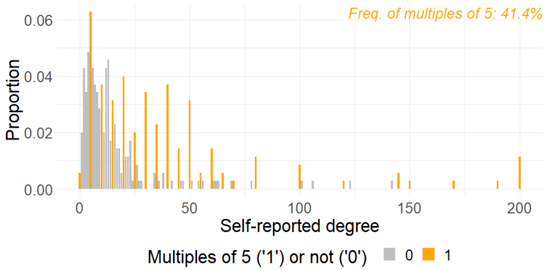
- network-based sampling $\Rightarrow$ a statistically invalid sample of broader coverage
- RDS provides a mathematical model of recruitment process then weights network-based samples to compensate for non-random recruitment patterns.
- Individual degree is used as a proxy for the sampling probability.

Self-reported degree

- is one commonly used estimation of degree
- has well-documented problems (Brewer, 2000)
- is frequently rounded to the nearest five or ten , known as **heaping** (Avery et al., 2021)
- can bias inferences when being used as sampling probability

Example from PATH Study (Lee, 2017)

- "How many males/females in Great Detroit Area do you know who inject and you have seen in the past 30 days?"



Goals

- explore the reporting behavior and establish reporting rules
- propose a new estimation of the individual degree
- quantify the extent to which using reported/estimated degree affects statistical inference

Existing method I (Bar and Lillard, 2012)

- analyze the reported data on smoking behavior
    - "How long ago (in years) did you quit smoking"
- assume respondents either report accurately or a heaped value
    - *other forms of reporting error? a random guess*
- propose a heaping rule: round the truth to the nearest multiples of 5 or 10
    - *reasonable for their research problem*
    - *a more flexible rule may be more suitable in our case*

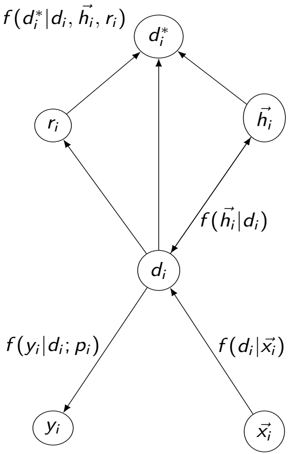
Existing method II (McCormick, Salganik, and Zheng, 2010)

- estimate personal network size by asking how many people they know in specific subpopulations
  - 12 subpopulations defined by the first name
  - external data on the population-level size proportion of each selected subpopulation in the nation.
- propose a latent nonrandom mixing model which is built on the scale-up method (Killworth et al., 1998) and resolves previously documented problems
- when applied to RDS:
  - most likely do not know target population-level size proportion of people with particular first names
  - use the nationwide information as a substitute

Our solution

- blend the analysis of reporting behaviors and information of subpopulation
- construct a latent variable model to make inferences about individual degree

Model Structure



- d_i (unobserved truth) vs d_i^* (self-reported degree)
- $\vec{h}_i = (h_{i,exact}, h_{i,heap}, h_{i,guess})$, reporting behavior indicator
- r_i: self-recruitment rate
- y_i: number of friends named Pat
- $\vec{x}_i$: variables associated with d_i

Individual's true degree (d_i) $\sim$ Covariates of interest ($\vec{x_i}$)

- $d_i \sim$ 0-truncated Poisson with mean on the log scale $= \vec{x_i}^T \vec{\alpha}$
- $\vec{x_i}$: demographic characteristics and characters associated with the target population

Number of Pat friends (y_i) $\sim$ Individual's true degree (d_i)

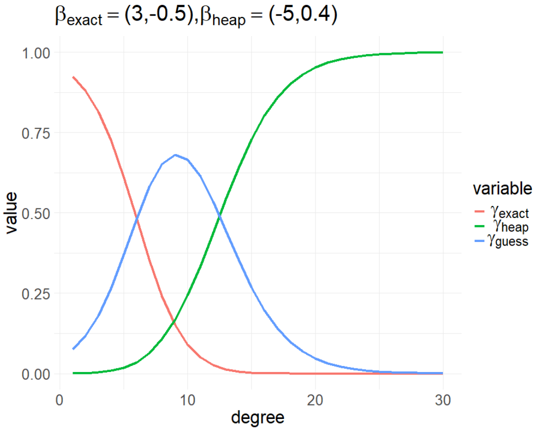
- $y_i \sim Binomial(d_i, p_i)$, p_i is known
- assume no or small reporting error in y_i (nationwide size proportion $\approx 1\%$)

Reporting behavior ($\vec{h_i}$) $\sim$ Individual's true degree (d_i)

- Assume 3 possible reporting behaviors:
  - reporting accurately: $d_i^* = d_i$
  - heaping, i.e., a multiple of 5: $d_i^* = 5n$
  - making a guess
- Intuitively, people are more likely to heap if d_i is large. Conversely, reporting an exact value if d_i is small. Otherwise, some guesses will be reported.

Reporting behavior ($\vec{h_i}$) ~ Individual degree (d_i) (Cont.)

- $\vec{h_i} = (h_{i,exact}, h_{i,heap}, h_{i,guess}) \sim Multinomial(\vec{\gamma}(d_i; \vec{\beta}))$, where $\vec{\gamma}$ is modeled via a spline model: $\begin{cases} log(\frac{\gamma_{i,exact}}{\gamma_{i,guess}}) = \beta_{exact,0} + \beta_{exact,1} d_i \\ log(\frac{\gamma_{i,heap}}{\gamma_{i,guess}}) = \beta_{heap,0} + \beta_{heap,1} d_i \end{cases}$



Reporting rules $f(d_i^*|d_i, \vec{h}_i, r_i)$

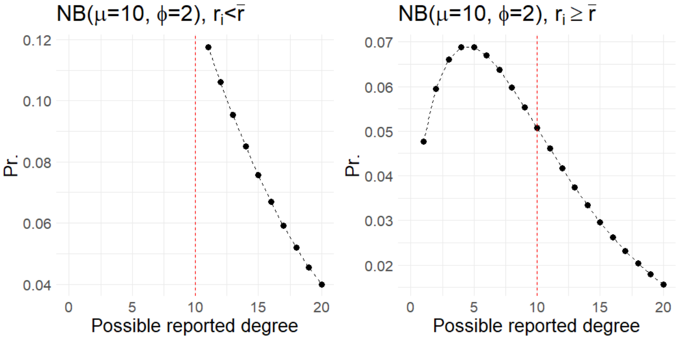
- if $h_{i,exact} = 1$, always report the truth: $Pr(d_i^*|d_i, h_{i,exact} = 1) = I(d_i^* = d_i)$
- for the other two cases, leverage the information provided by r_i:
    - if recruiting less than average, the participant is believed to overestimate his degree: $d_i^* > d_i$

Reporting rules $f(d_i^*|d_i, \vec{h}_i, r_i)$ (Cont.)

- if $h_{i,guess} = 1$, d_i^* is drawn from a truncated Negative Binomial distribution:

$$Pr(d_i^*|d_i, h_{i,guess} = 1, r_i) = \begin{cases} I(d_i^* > d_i)\frac{Pr(X=d_i^*)}{Pr(X>d_i)}, & \text{if } r_i < \bar{r} \\ I(d_i^* > 0)\frac{Pr(X=d_i^*)}{Pr(X>0)}, & \text{if } r_i \geq \bar{r} \end{cases}$$

where $X \sim \text{NegBin}(d_i, \phi)$, $E[X] = d_i$, $Var[X] = d_i + d_i^2/\phi$

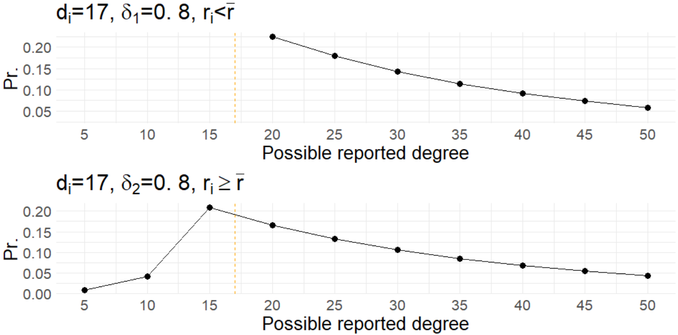


Reporting rules $f(d_i^*|d_i, \vec{h}_i, r_i)$ (Cont.)

- if $h_{i,heap} = 1$, d_i^* is a multiple of 5, drawn from

$$Pr(d_i^*|d_i, h_{i,heap}=1, r_i) = \begin{cases} \sum_{k\geq 1} \frac{\delta_1^k}{\sum_{n\geq 1}\delta_1^n} I(d_i^* = 5\lfloor d_i/5 \rfloor + 5k), & \text{if } r_i < \bar{r} \\ \begin{cases} Pr(d_i^* = 5\lfloor d_i/5\rfloor + 5k_1|d_i, h_{i,heap=1}) = \frac{\delta_2^{k_1}}{\sum_{n_1\geq 1}\delta_2^{n_1} + \sum_{n_2=0}^{K}(1-\delta_2)^{n_2}} \\ Pr(d_i^* = 5\lfloor d_i/5\rfloor - 5k_2|d_i, h_{i,heap=1}) = \frac{(1-\delta_2)^{k_2}}{\sum_{n_1\geq 1}\delta_2^{n_1} + \sum_{n_2=0}^{K}(1-\delta_2)^{n_2}} \end{cases}, & \text{if } r_i \geq \bar{r} \end{cases}$$

- Under this model
  - most likely, a heaped value around the truth will be reported.
  - appropriate values of (δ_1, δ_2) result in extremely large reported degree.



Computation Algorithm

Algorithm: Monte Carlo Expectation-Maximization algorithm

Result: Posterior mean of unobserved latent individual degree conditional on the observed data

Input : Observed data $Y_{obs} = \{$reported degree d_i^*, self-recruitment rate r_i, number of acquaintance in a subpopulation y_i, characteristics of interest $\vec{x_i}\}$

External information: size proportion of the subpopulation p

Step 1: Initialize unobserved latent variables:

$$Y_{mis} = \{\text{individual degree } d_i^{(0)}, \text{ reporting behavior indicator } \vec{h_i}^{(0)}\}$$

Initialize hyperparameters of interest $\Theta^{(0)} = \{\vec{\alpha}^{(0)}, \vec{\beta}^{(0)}, \vec{\delta}^{(0)}, \phi^{(0)}\}$

Step 2: Monte Carlo - Expectation step:

simulate a sample $\{Y_{mis,i}\}_{i=1}^{M}$ from $f(Y_{mis}; \Theta^{(t)})$

estimate the expectation of functions of data $g(Y_{mis})$:

$$E_{Y_{mis}} g(Y_{mis}) \approx \frac{\sum_{i=1}^{M} g(Y_{obs}, Y_{mis}) f(Y_{obs}|Y_{mis})}{\sum_{i=1}^{M} f(Y_{obs}|Y_{mis})}$$

Step 3: Maximization step:

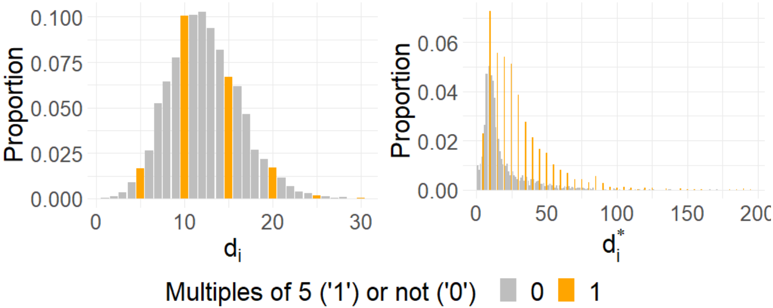
update the estimates $\Theta^{(t+1)}$ via a one-step Fisher scoring algorithm

Step 4: Iterate between Steps 2 and 3 until convergence

Simulation Study

Create a population of size 5000. For individual i, simulate

- multiple characteristics $\vec{x_i}$
- degree $d_i \sim Poisson(\mu = e^{\vec{x_i}^T \vec{\alpha}})$
- the number of Pat friends $y_i \sim Binomial(d_i, p)$
- self-recruitment rate r_i from $\{0, 1/3, 2/3, 1\}$ with probability $(0.4, 0.3, 0.2, 0.1)$
- reported degree d_i^* following the proposed reporting mechanism
- a binary trait correlated with d_i^* and d_i

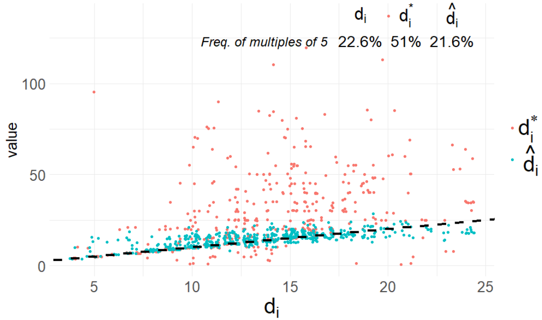


	d_i	d_i^*
Freq. of multiples of 5	20.36%	42.88%

Simulate 1000 RDS samples:

- build a social connection network based on simulated $\{d_i\}_{i=1}^{N}$
- recruit individuals following the standard RDS procedure:
  - start with 3 seeds, sampling w/o replacement w/. probability proportional to d_i
  - issue $\min(3, d_i^*)$ coupons to each participant
  - select subsequent participants w/o replacement and at random from among the contacts of the current recruiter
  - keep recruiting (and add seeds if necessary) until reaching 500

Degree estimation of a randomly chosen sample



Summary of 1000 samples

	$\hat{d}_i$	d_i^*
Ave.MSE	9.77	574.15
SD.MSE	1.91	93.14
Ave.Freq of Multiples of 5	20.52%	46.96%

Note: $\text{MSE}(\boldsymbol{x}) = \sum_{i=1}^{s}(d_i - x_i)^2/s$

Methods of estimating the prevalence of a binary trait

- all use degree as sampling weights in some form
- RDS_I (Salganik and Heckathorn, 2004): equate the number of network ties between every pair of subgroups with different trait responses, with a critical step to estimate average degree for people in different trait groups
- RDS_II (Volz and Heckathorn, 2008): generalize Horvitz-Thompson type point estimator by approximating the sampling probability as proportional to the individual's degree
- RDS_SS (Gile, 2011): advance RDS_II by incorporating successive sampling model to account for the sampling without replacement feature

Sample-based estimated prevalence of a binary trait

	RDS_I			RDS_II			RDS_SS		
Degree type	d_i	$\hat{d}_i$	d_i^*	d_i	$\hat{d}_i$	d_i^*	d_i	$\hat{d}_i$	d_i^*
Ave.Bias	0.001	-0.007	-0.227	0.001	-0.006	-0.227	0.004	-0.003	-0.214
Ave.SD	0.035	0.037	0.060	0.050	0.050	0.051	0.046	0.047	0.051
CI width[1]	0.139	0.145	0.237	0.195	0.196	0.201	0.182	0.183	0.200
Coverage rate	0.991	0.990	0.012	0.998	0.999	0.001	0.996	0.998	0.001

Notes: this trait has 70% true prevalence,
and its Spearsman's rank correlation with $d_i^*(d_i)$ is 0.67(0.46);
CI width [1]= 95% confidence interval width.

Discussion

Our modeling of the reporting mechanism

- identify different sources of reporting error by specifying multiple types of reporting behaviors
- conform to the intuition and well explain the observed data

The proposed individual degree estimation

- blend the analysis of reporting behaviors and information of number of acquaintance in a subpopulation and self-recruitment rate
- yield modestly biased point estimation
- improve statistical inference when serving as sampling probability

Our framework

- is flexible to accommodate any distribution assumptions researchers believe underline the data-generating process
- is vulnerable to model misspecification as a model-based approach

- Avery, Lisa et al. (2021). "A review of reported network degree and recruitment characteristics in respondent driven sampling implications for applied researchers and methodologists". In: *Plos one* 16.4, e0249074.
- Bar, Haim Y and Dean R Lillard (2012). "Accounting for heaping in retrospectively reported event data–a mixture-model approach". In: *Statistics in medicine* 31.27, pp. 3347–3365.
- Brewer, Devon D (2000). "Forgetting in the recall-based elicitation of personal and social networks". In: *Social networks* 22.1, pp. 29–43.
- Gile, Krista J (2011). "Improved inference for respondent-driven sampling data with application to HIV prevalence estimation". In: *Journal of the American Statistical Association* 106.493, pp. 135–146.
- Killworth, Peter D et al. (1998). "Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach". In: *Evaluation review* 22.2, pp. 289–308.
- Lee, Juliette Roddy (2017). "Project positive attitudes towards health, Michigan, 2017". In.
- McCormick, Tyler H, Matthew J Salganik, and Tian Zheng (2010). "How many people do you know?: Efficiently estimating personal network size". In: *Journal of the American Statistical Association* 105.489, pp. 59–70.
- Salganik, Matthew J and Douglas D Heckathorn (2004). "Sampling and estimation in hidden populations using respondent-driven sampling". In: *Sociological methodology* 34.1, pp. 193–240.
- Volz, Erik and Douglas D Heckathorn (2008). "Probability based estimation theory for respondent driven sampling". In: *Journal of official statistics* 24.1, p. 79.