Mambilla Journal of Science and the Environment Vol. 10, Issue 1, Page No. 83-107

MAMBILLA JOURNAL OF SCIENCE AND THE ENVIRONMENT

journal homepage: https://oer.tsuniversity.edu.ng/index.php/mjse

APPLICATION OF RESPONDENT DRIVEN SAMPLING TO ESTIMATE THE NUMBER OF IDPS IN A HOST COMMUNITY IN BIU LOCAL GOVERNMENT AREA OF BORNO STATE, NIGERIA.

[1,2*]Anjikwi, Y.,[2]Jibasen, D., [3]Dike, I.J., [2]Torsen, E

[1] Department of Agricultural Economics, University of Maiduguri, Borno-Nigeria.
[2] Department of Statistics, Modibbo Adama University, Yola-Nigeria.
[3] Department of Operations Research, Modibbo Adama University, Yola-Nigeria.

ARTICLE INFO

Article history:
Received 14 December 2025
Received in revised form 27 January 2026
Accepted 27 January, 2026

Keywords:
Respondent-Driven Sampling, Age, Evaluation, Hidden population, Populations.

ABSTRACT

This study systematically evaluates the performance of multiple estimators Naïve, SH-RDS, VH-RDS, G-SS, and $RDS_{proposed}$ in estimating the age distribution of Facebook user responses across varying sample sizes and degree distributions. Results indicate that, for small sample sizes (n=500), traditional estimators such as VH-RDS and SH-RDS exhibit lower bias, variance, and mean squared error (MSE), while the $RDS_{proposed}$ estimator performs suboptimally. However, as sample sizes increase (n=1000 to n>2000), the R$RDS_{proposed}$ DS_proposed estimator demonstrates clear superiority, achieving minimal bias and variance along with the lowest MSE, thus providing the most reliable estimates. Sectoral analysis further reveals that responses are more stable and precise in the Company and Politician sectors under very high and moderate degree categories, with increased variability observed in the Government sector, particularly at very low degrees. These findings are consistent with existing literature, which emphasizes the importance of estimator selection, sample size, and network structure on the accuracy and precision of respondent-driven sampling results. Limitations include reliance on simulated datasets, focus on age distribution as the primary parameter, and incomplete reporting of all metrics for every sector. Overall, the study highlights optimal scenarios for estimator application and the critical role of degree distribution in ensuring robust network-based survey inferences.

1. Introduction

Populations that are hidden or difficult to access are frequently composed of small, marginalized groups who exist beyond the reach of conventional data-gathering procedures. Their exclusion from standard research frameworks presents both significant challenges and distinctive opportunities for researchers to engage these communities in substantive ways. Many members of these populations experience substantial risks when exposed to public scrutiny, which can compromise their security and confidentiality. Illustrative examples include individuals who inject drugs (IDUs) and may conceal their activities due to the intense stigma attached to drug use, victims of human trafficking who endure exploitation in silence, men who have sex with men (MSM) who fear social ostracism, people experiencing homelessness with unstable living arrangements, and migrants who lack formal status or adequate support. Recognizing the pervasive stigma faced by these groups is vital, as it heightens their need for discretion and complicates outreach and data collection efforts. Gaining a nuanced understanding of these social dynamics allows researchers to develop more effective, ethically sound strategies for gathering meaningful data (Sarah *et al*., 2022; Johnson *et al*., 2023).

Due to the absence of comprehensive sampling frames, researchers are often required to employ innovative non-probability sampling techniques. Methods such as convenience sampling (including snowball sampling, where current participants recruit new ones), time-location sampling (targeting specific places and periods frequented by the population), and respondent-driven sampling (RDS, which provides incentives for participant referrals), are commonly utilized. While these approaches can introduce concerns regarding representativeness and potential biases,

* Corresponding author: +2348036267312
E-mail address: y.anjikwi@gmail.com

they also yield valuable insights into the experiences and contexts of concealed groups. By acknowledging these methodological limitations and addressing them proactively, researchers can implement strategies to enhance the precision and relevance of studies focused on these frequently neglected populations (Brown *et al.,* 2021; Smith & Lee, 2024).

Respondent-driven sampling (RDS) exemplifies an adaptive sampling technique that capitalizes on existing social networks to recruit individuals from hidden populations lacking formal sampling frameworks. RDS has proven effective in accessing marginalized or underrepresented groups who are often missed by traditional research methodologies. In health-related research, for instance, RDS has facilitated the recruitment of various high-risk groups such as people who inject drugs, MSM, urban musicians, and those without stable housing (Card *et al*., 2017; Heckathorn & Cameron, 2017; Lyons *et al.,* 2023; White *et al.,* 2015).

Moreover, RDS plays a pivotal role in enabling the recruitment of participants for health interventions, especially those who might otherwise avoid health services. Its applicability extends beyond healthcare, aiding efforts to reach migrant populations in diverse contexts (Górny & Napierała, 2016; Keygnaert et al., 2014; Martinez, *et al.,* 2022). The method's capacity to produce robust population estimates and its practical success in real-world settings have led to a marked increase in the use of RDS globally (Johnston *et al.,* 2016; Nguyen *et al.,* 2025). This upward trend highlights the significance of RDS in deepening our understanding of hidden populations and addressing their specific needs.

The RDS approach initiates with the selection of "seed" participants, drawn from a convenience sample within the target population. These seeds complete a survey, which can be distributed online to maximize accessibility. Upon survey completion, each seed is allowed to invite a set number of peers also members of the target group to participate. The recruitment process is managed via a coupon system, which tracks referral patterns and encourages participants to disclose information about their social networks. This iterative recruitment continues across multiple waves, expanding the sample size, reducing dependence on the initial convenience sample, and increasing participant diversity. Incentives are provided both for survey completion and for successful peer recruitment, a dual reward structure that increases engagement and broadens the respondent pool, as observed by Gile & Handcock (2010) and White *et al.* (2015).

Today, RDS is recognized as a highly versatile and effective research methodology, having been used in over 460 studies worldwide (White *et al.,* 2015; Thompson et al., 2023). Its innovative design enables the collection of essential data from populations that would otherwise remain inaccessible, thereby enhancing scholarly understanding of diverse communities and providing invaluable insights into their experiences and needs. Employing RDS allows researchers to obtain critical information from groups that are habitually marginalized in academic and policy research.

RDS faces significant statistical challenges that affect the validity of population inferences. The two main RDS estimators, Salganik and Heckathorn, (2004) (SH-RDS) and Volz and Heckathorn (2008) (VH-RDS), are based on theoretical assumptions that often do not hold true in real-world applications, primarily due to their reliance on sampling with replacement and the need for a large number of recruitment waves. In practice, studies typically recruit only a limited number of waves and do not allow for participants to be sampled more than once, leading to systematic biases, particularly the overrepresentation of well-connected individuals. Recent advancements have introduced sampling without replacement estimators that better model RDS recruitment, but these still have limitations, such as instability in large samples and vulnerability to recruitment pattern deviations. Existing methods also lack effective protocols for seed selection and adaptive weighting, which are crucial for addressing network diversity. This study aims to develop a new RDS estimator that combines sampling without replacement, strategic seed selection, and adaptive weighting to provide reliable outcomes in complex, real-world settings with large sample sizes and diverse networks.

2. Methods

This study explores the existing RDS estimators rooted in the theory of undirected social relation networks. It highlights the susceptibility of these estimators to bias when applied to populations with a high sampling fraction. To counteract these biases, the study proposes an alternative estimator that employs weighted edges and nodes, aiming to diminish the impact of the biases observed in traditional RDS estimators when sampling from hidden populations with a considerable number of direct connections.

2.1. The Naïve Estimator

The Naive estimator was proposed by Heckathorn (1997) and is simply the proportion of infected individuals found in the sample.

$$\hat{\mu}_{NV} = \frac{\phi_A}{\phi_A + \phi_B} \quad (1)$$

where, $\hat{\mu}_{NV}$ is the population proportion of A, ϕ_A is the number of recruits in group A, and ϕ_B, is the number of recruits in group B.

Equation (1) indicates that when equal sampling probabilities are observed for individuals in both group A and group B, it acts as a generalized Hansen-Hurwitz estimator for the parameter of interest.

2.1.1 The Assumptions of Naïve Estimator

The assumptions of Naïve Estimator (1997) are:

i. Respondents recruit peers from their social contacts with equal probability.
ii. Sampling is done with replacement.
iii. The degree of respondents is normally distributed.
iv. The social network of the population is undirected.
v. The population forms a connected network.

2.2 SH-RDS Estimator

The SH-RDS was developed by Salganik and Heckathorn (2004) for estimating population proportion using with replacement sampling, and the model is defined as

$$\hat{\mu}_{SH_A} = \frac{\hat{D}_B.\hat{C}_{B,A}}{\hat{D}_A.\hat{C}_{A,B} + \hat{D}_B.\hat{C}_{B,A}} \quad (2)$$

for group A, and

$$\hat{\mu}_{SH_B} = \frac{\hat{D}_A.\hat{C}_{A,B}}{\hat{D}_A.\hat{C}_{A,B} + \hat{D}_B\hat{C}_{B,A}} \quad (3)$$

for group B,

where:

$$\hat{C}_{A,B} = \frac{R_{AB}}{R_{AA} + R_{AB}}$$

$$\hat{C}_{B,A} = \frac{R_{BA}}{R_{BB} + R_{BA}}$$

$$R_{AA} = \sum_{i \in A} d_i$$

$$\hat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$$

$$\hat{D}_B = \frac{n_B}{\sum_{i=1}^{n_B} \frac{1}{d_i}}$$

$\hat{\mu}_{SH_A}$ is the estimate of the population proportion of group A,
$\hat{\mu}_{SH_B}$ is the estimate of the population proportion of group B,
$\hat{C}_{B,A}$ is an estimate of probability of cross-sample recruitment from group B to group A

$\hat{C}_{A,B}$ is an estimate of the probability of cross-sample recruitment from group A to group B
R_{AA} is the total number of ties (edges) that contain a person in group A
$\widehat{D}_A$ is the estimate of the mean degree of group A
$\widehat{D}_B$ is the estimate of the mean degree of group B

2.2.1 The Assumptions of the SH-RDS estimator

The assumptions of Salganik and Heckathorn (2004) are:

i. Respondents recruit peers from their social contacts with equal probability.
ii. Each recruitment consists of only one peer.
iii. Sampling is done with replacement.
iv. The degree of respondents is reported without error.
v. The social network of the population is undirected, and
vi. The population forms a connected network.

2.3 VH-RDS Estimator

The VH-RDS model was developed by Volz and Heckathorn (2008) to estimate population proportion using with replacement sampling, and the model is defined as

$$\hat{\mu}_{VH} = \left(\frac{n_A}{n}\right)\left(\frac{\widehat{D}_u}{\widehat{D}_{A_V}}\right) \quad (4)$$

The variance of VH-RDS was defined as

$$\hat{V}_{HH}(\langle\hat{y}\rangle) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\widehat{D}_u}{\hat{d}_i} - \langle\hat{y}\rangle\right) \quad (5)$$

where:

n_A is the sample size of group A, and n is the total sample.
$\hat{y}$ is an estimate of the mean value of y
y is an indicator function that a value $I_i(i) = \begin{cases} 1, & i \in A \\ 0, & otherwise \end{cases}$
$\widehat{D}_u$ is the estimate of mean degree of the sample u

$$\widehat{D}_{A_V} = \frac{N_u}{\sum_{i=1}^{N_u}\frac{1}{d_i}} \quad (6)$$

$\widehat{D}_{A_V}$ is the estimate of mean degree of the sample group A

2.3.1 The Assumptions of the VH-RDS Estimator

The assumptions of Volz and Heckathorn (2008) are:

i. Recruitment is random. When recruiting others, respondents select uniformly at random from their network.
ii. Each recruitment consists of only one peer.
iii. Sampling is done with replacement.
iv. Respondents accurately report their degree in the network.
v. Network connections are reciprocal.
vi. The population forms a connected network.
vii. Convergence, recruitment is modeled as a Markov process (MP), where the state of the MP is the last individual recruited.

2.4 Gile's Successive Sampling (G-SS) Estimator

The G-SS model was developed by Gile (2011) for estimating population proportion characteristics using without replacement sampling. The Gile (2011) estimate of inclusion probability associated with a specific unit was defined as

$$\tilde{\pi}_{SS} = \frac{U_i+1}{M+1} \quad (7)$$

where,

U_i is the number of times unit i is sampled in M trials, and the model was defined as

$$\hat{\mu}_{SS} = \frac{\sum_{j=1}^{N} \frac{s_j z_j}{\tilde{\pi}_{SS(d_j)}}}{\sum_{j=1}^{N} \frac{s_j}{\tilde{\pi}_{SS(d_j)}}} \quad (8)$$

$\hat{\mu}_{SS}$ is the estimate of population proportion.

2.4.1 The assumptions of Gile's Successive Sampling (G-SS) Estimator

The assumptions of Gile (2011) are:

i. Recruitment is random. When recruiting others, respondents select uniformly at random from their network.
ii. Each recruitment consists of only one peer.
iii. Sampling is done without replacement.
iv. Respondents accurately report their degree in the network.
v. Network connections are reciprocal.
vi. The population forms a connected network.
vii. The population size N is known.

2.5 The Proposed RDS Estimator

2.5.1 Assumptions of the Proposed Model

The assumptions of the proposed model are:

i. Respondents recruit peers from their social contacts with equal probability (random).
ii. Each recruitment consists of only one peer (throughout the sampling period).
iii. Sampling is done without replacement.
iv. The degree of respondents reported has a negligible error.
v. The network is directed.
vi. The population forms a connected network.
vii. The population size N is unknown.

2.5.2 Estimation of inclusion probability

The idea of probability proportional to size without replacement (PPSWOR) was extended to an RDS estimator to sample a hidden population (Naser *et al.*, 2018; Lawson & Ponkaew, 2019). Since a node was recruited into an RDS sample with a probability proportional to its degree, the proposed inclusion probability was modelled as.

let $s = \{i_1 i_2, \ldots, i_n\}$ be the RDS sample and $\lambda_{i_{proposed}}$ be the proposed inclusion probability, then,

$$\lambda_{i_{proposed}} = P(i \in S) \quad (9)$$

Therefore, the proposed inclusion probability was modelled using the (Anjikwi *et al*, 2026) approach as

$$\lambda_{i_{proposed}} = \sum_{s \in S} P(s) \times I(i \in s) \quad (10)$$

where,

$P(s)$ is the probability of selecting sample s

$I(i \in s)$ is an indicator function (1 if i ∈ s, 0 otherwise)

For the RDS sample without replacement, $P(s)$ can be modelled as:

$$P(s) = P(i_1) \prod_{k=2}^{n} P(i_k | i_{k-1})$$

where,

$P(i_1)$ is the probability of selecting the seed node i_1(often not random, or usually 1).

$P(i_k | i_{k-1})$ is the probability that node i_k is recruited by node i_{k-1}.

$$P(i_k | i_{k-1}) = \frac{D_{i_k}}{\sum_{j \in N(i_{k-1}) - \{i_1 i_2, \ldots, i_{k-1}\}} D_j} \quad (11)$$

where,
$N(i_{k-1})$ is the set of neighbors of i_{k-1}, not yet recruited.
Substitute P(s) into Equation (3.13)

$$\lambda_{i_{proposed}} \approx \sum_{s\in S}\left(\prod_{k=2}^{n}\frac{D_{i_k}}{\sum_{j\in N(i_{k-1})-\{i_1 i_2,\ldots,i_{k-1}\}}D_j}\right)\times I(i\in s) \quad (12)$$

Simplify the expression in Equation (12)

$$\lambda_{i_{proposed}} \approx \frac{D_i}{(\sum_{j\in \cup} D_j)} \quad (13)$$

where,
$I(i \in s)$ is equal to 1 if i belong to S and ∪ is the total number of nodes.
$\lambda_{i_{proposed}}$ can be approximated as

$$\lambda_{i_{proposed}} \approx \frac{D_i}{E} \quad (13)$$

where, E is the total number of edges in the network.

$$E \approx \sum_{j\in \cup} D_j \quad (14)$$

2.5.3 Estimation of mean degree ($D_{proposed}$)

The proposed mean degree ($D_{proposed}$) of a node i was estimated as:
Let d_i be the degree of node i, and $\lambda_{i_{proposed}}$ the inclusion probability of node i (approximated as $\frac{D_i}{E}$, then the estimate of the mean degree of node i as

$$\widehat{D}_{proposed} = \frac{\left(\sum_{i\in S}\frac{d_i}{\lambda_{i_{proposed}}}\right)}{\sum_{i\in S}\frac{1}{\lambda_{i_{proposed}}}} \quad (15)$$

Substitute the value of $\lambda_{i_{proposed}}$ into Equation (15)

$$\widehat{D}_{proposed} = \frac{\left(\sum_{i\in S}\frac{d_i}{\frac{D_i}{(E)}}\right)}{\sum_{i\in S}\frac{1}{\frac{D_i}{(E)}}} \quad (16)$$

Simplifying Equation (16) gives

$$\widehat{D}_{proposed} = \left(\frac{E\sum_{i\in S}\frac{d_i}{D_i}}{E\sum_{i\in S}\frac{1}{D_i}}\right) \quad (17)$$

$$\therefore \quad \widehat{D}_{proposed} = \frac{\left(\sum_{i\in S}\frac{d_i}{D_i}\right)}{\left(\sum_{i\in S}\frac{1}{D_i}\right)} \quad (18)$$

2.5.4 Estimation of cross-group edges ($S_{g_a g_b}$)

Let $S_{g_a g_b}$ be the number of cross-group edges, that is from group g_a to group g_b, then the probability of a cross-group edge being reported as:

$$\text{P(edge (i, j)is reported)} = \frac{D_i}{2E}\times\left(\frac{1}{D_i}\right)+\frac{D_j}{2E}\times\left(\frac{1}{D_j}\right) \quad (19)$$

$$= \frac{1}{2E}+\frac{1}{2E}$$

$$= \frac{1}{E} \quad (20)$$

where,
$i \in g_a \text{ and } j \in g_b \text{ (or vice versa)}$
$D_i \text{ is the degree of node } i$
$D_j \text{ is the degree of node } j$
$E \text{ is the total number of edges in the network}$
Estimating $S_{g_a g_b}$ using the reported cross-group edges.

$$E\left[S_{g_a g_b}\right] = \sum_{\{i \in g_a, j \in g_b\}} P(edge(i,j)\ is\ reported)$$
$$= \frac{S_{g_a g_b}}{E} \quad (21)$$

Use the RDS data to estimate $S_{g_a g_b}$

$$S_{g_a g_b} = E \times \frac{number\ of\ g_a \to g_b\ edges\ reported}{\left(\sum_{i \in s} \frac{1}{D_i}\right)}$$

where, s is the RDS sample, E can be approximated as defined as in Equation (14)

2.5.5 Estimation of population proportion

The proposed RDS estimator was specified for the case of estimating the proportion of individuals with a particular trait in a RDS network setting. Specifically, if $S_{g_a g_b}$ denotes the total number of observed links (edge) from group g_a to group g_b, λ_1 and λ_0 are the respective inclusion probabilities for individuals with and without the trait, D_1 and D_0 are their respective degrees, therefore, the proposed RDS estimator was modelled as

$$\widehat{RDS}_{proposed} = \frac{\left(\frac{S_{g_a g_b}}{\lambda_1}\right)}{\left(\frac{S_{g_a g_b}}{\lambda_1} + \frac{S_{g_b g_a}}{\lambda_0}\right)} \quad (22)$$

$$= \frac{\left(\frac{S_{g_a g_b}}{\frac{D_1}{(E)}}\right)}{\left(\frac{S_{g_a g_b}}{\frac{D_1}{(E)}} + \frac{S_{g_b g_a}}{\frac{D_1}{(E)}}\right)}$$

$$= \frac{\left(E * \frac{S_{g_a g_b}}{D_1}\right)}{\left(E * \frac{S_{g_a g_b}}{D_1} + E * \frac{S_{g_b g_a}}{D_0}\right)}$$

$$= \frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1} + \frac{S_{g_b g_a}}{D_0}\right)}$$

$$= \frac{\left(D_0 * S_{g_a g_b}\right)}{\left(D_0 * S_{g_a g_b} + D_1 * S_{g_a g_b}\right)} \quad (23)$$

This formulation allows for the estimation of the proportion of group (g_a) in the population, accounting for the network structure and sampling design inherent to RDS.

2.5.6 Estimation of Variance of Proposed Model

Let's estimate the variance of $\widehat{RDS}_{proposed}$

$$\widehat{RDS}_{proposed} = \frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1} + \frac{S_{g_b g_a}}{D_0}\right)} \quad (24)$$

where,

$S_{g_a g_b}$ is the number of cross-group edges from group g_a to group g_b. Therefore, using the delta method, Anjikwi *et al.* (2026) the variance of the proposed estimator ($\widehat{RDS}_{proposed}$) was estimated as:

$$Var\left(\widehat{RDS}_{proposed}\right) \approx \left(\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}}\right)^2 var\left(S_{g_a g_b}\right) + \left(\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}}\right)^2 var\left(S_{g_b g_a}\right)$$
$$+2\left(\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}}\right)\left(\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}}\right) Cov(S_{g_a g_b}, S_{g_b g_a}) \quad (25)$$

Computing the partial derivatives of $\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}}$ and $\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}}$, gave

$$\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}} = \frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1} + \frac{S_{g_b g_a}}{D_0}\right)^2} \quad (26)$$

$$\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}} = -\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)} \quad (27)$$

$$var\left(S_{g_a g_b}\right) \approx \frac{S_{g_a g_b}(1-\lambda_1)}{\lambda_1} \quad (28)$$

$$var\left(S_{g_b g_a}\right) \approx \frac{S_{g_b g_a}(1-\lambda_0)}{\lambda_0} \quad (29)$$

$$Cov\left(S_{g_a g_b}, S_{g_b g_a}\right) = E\left[\left(S_{g_a g_b} - E\left[S_{g_a g_b}\right]\right)\left(S_{g_b g_a} - E\left[S_{g_b g_a}\right]\right)\right] \quad (30)$$

Substitute the partial derivatives in Equation (26, 27, 28, 29 and 30 into 25)

$$Var\left(\widehat{RDS}_{proposed}\right) \approx \left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)^2 \frac{S_{g_a g_b}(1-\lambda_1)}{\lambda_1} + \left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right)^2 \frac{S_{g_b g_a}(1-\lambda_0)}{\lambda_0}$$

$$+2\left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)\left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right) E\left[\left(S_{g_a g_b} - E\left[S_{g_a g_b}\right]\right)\left(S_{g_b g_a} - E\left[S_{g_b g_a}\right]\right)\right] \quad (31)$$

2.5.7 Estimation of Bias of Proposed Model

To compute the bias of $\widehat{RDS}_{proposed}$, there was a need to estimate the expected value of $\widehat{RDS}_{proposed}$ and compared to the RDS true population proportion (Anjikwi *et al.,* 2026) . Let's denote the true population parameter as $\left(\widehat{RDS}_{proposed}\right)$

$$\widehat{RDS}_{proposed} = \frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)} \quad (32)$$

The expected value of $\widehat{RDS}_{proposed}$ can be expressed as

$$E\left[\widehat{RDS}_{proposed}\right] \approx \frac{\frac{E\left[S_{g_a g_b}\right]}{D_1}}{\frac{E\left[S_{g_a g_b}\right]}{D_1}+\frac{E\left[S_{g_b g_a}\right]}{D_0}}$$

$$\approx \frac{E\left[S_{g_a g_b}\right]}{\left(E\left[S_{g_a g_b}\right]+E\left[S_{g_b g_a}\right]\times\frac{D_1}{D_0}\right)} \quad (33)$$

The bias of $\widehat{RDS}_{proposed}$ is the difference between the expected value and the true population parameter:

$$Bias\left(\widehat{RDS}_{proposed}\right) \approx E\left[\widehat{RDS}_{proposed}\right] - RDS_{proposed} \quad (34)$$

$$\approx \frac{\frac{E\left[S_{g_a g_b}\right]}{D_1}}{\frac{E\left[S_{g_a g_b}\right]}{D_1}+\frac{E\left[S_{g_b g_a}\right]}{D_0}} - \frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)} \quad (35)$$

Using the delta method,

$$Bias\left(\widehat{RDS}_{proposed}\right) \approx \frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}}\left(E\left[S_{g_a g_b}\right] - S_{g_a g_b}\right) + \frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}}\left(E\left[S_{g_b g_a}\right] - S_{g_b g_a}\right) \quad (36)$$

where, $\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_a g_b}}$ and $\frac{\partial \widehat{RDS}_{proposed}}{\partial S_{g_b g_a}}$ are, as define in Equation (26) and Equation (27) respectively. Therefore, upon simplification, the expression becomes

$$Bias\left(\widehat{RDS}_{proposed}\right) \approx \left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)\left(E\left[S_{g_a g_b}\right] - S_{g_a g_b}\right) + \left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right)\left(E\left[S_{g_b g_a}\right] - S_{g_b g_a}\right) \quad (37)$$

2.5.8 Estimation of Mean Squared Error of Proposed Model

The mean square error of the model was estimated using the method employed by Anjikwi *et al.* (2026) as:

$$MSE\left(\widehat{RDS}_{proposed}\right) = var\left(\widehat{RDS}_{proposed}\right) + Bias^2 \quad (38)$$

by substituting the values $var\left(\widehat{RDS}_{proposed}\right)$ and $Bias^2$ into Equation (38), the $MSE\left(\widehat{RDS}_{proposed}\right)$ becomes

$$MSE\left(\widehat{RDS}_{proposed}\right) = \left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)^2 \frac{S_{g_a g_b}(1-\lambda_1)}{\lambda_1} + \left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right)^2 \frac{S_{g_b g_a}(1-\lambda_0)}{\lambda_0} +$$
$$2\left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)\left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right) E\left[\left(S_{g_a g_b} - E\left[S_{g_a g_b}\right]\right)\left(S_{g_b g_a} - E\left[S_{g_b g_a}\right]\right)\right] +$$
$$\left(\left(\frac{\left(\frac{1}{D_1}\right)\left(\frac{S_{g_b g_a}}{D_0}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2}\right)\left(E\left[S_{g_a g_b}\right] - S_{g_a g_b}\right) + \left(-\frac{\left(\frac{S_{g_a g_b}}{D_1}\right)}{\left(\frac{S_{g_a g_b}}{D_1}+\frac{S_{g_b g_a}}{D_0}\right)^2\left(\frac{1}{D_0}\right)}\right)\left(E\left[S_{g_b g_a}\right] - S_{g_b g_a}\right)\right)^2 \quad (39)$$

Upon simplification, the expression becomes

$$MSE\left(\widehat{RDS}_{proposed}\right) = \left(\frac{XY}{Z^2}\right)\left[\frac{S_{g_a g_b}(1-\lambda_1)}{\lambda_1} - \frac{S_{g_b g_a}(1-\lambda_0)}{\lambda_0}\right] + 2\left(\frac{XY}{Z^2}\right)\left[\left(S_{g_a g_b} - E\left[S_{g_a g_b}\right]\right)\left(S_{g_b g_a} - E\left[S_{g_b g_a}\right]\right)\right] +$$
$$\left(\frac{XY}{Z^2}\right)^2\left[\left(E\left[S_{g_a g_b}\right] - S_{g_a g_b}\right) - \left(E\left[S_{g_b g_a}\right] - S_{g_b g_a}\right)\right]^2 \quad (40)$$

where, $X = \frac{S_{g_a g_b}}{D_1}$, $Y = \frac{S_{g_b g_a}}{D_0}$, $Z = X + Y$, the MSE can also be specify as

$$MSE\left(\widehat{RDS}_{proposed}\right) = A\left[\frac{S_{g_a g_b}(1-\lambda_1)}{\lambda_1} - \frac{S_{g_b g_a}(1-\lambda_0)}{\lambda_0}\right] + 2(A)\left[\left(S_{g_a g_b} - E\left[S_{g_a g_b}\right]\right)\left(S_{g_b g_a} - E\left[S_{g_b g_a}\right]\right)\right] +$$
$$(A)^2\left[\left(E\left[S_{g_a g_b}\right] - S_{g_a g_b}\right) - \left(E\left[S_{g_b g_a}\right] - S_{g_b g_a}\right)\right]^2 \quad (41)$$

where, $A = \frac{XY}{Z^2}$

2.6 Source of Data

2.6.1 Data Simulation

To assess the robustness of the proposed estimator, $RDS_{proposed}$, the study simulated and calculated the outcome variable of gender for a series of samples($n = 500, n = 1000, n = 1500, n = 2000\ and\ n > 2000$ drawn from the overall population of 10,000 . For each sample, the study meticulously computed the proportions of male and female participants, along with important statistical metrics such as the variance and bias. Initially, 8 key recruits, known as "seeds," were selected. Each of the 8 seeds were tasked with 2 coupons (a unique code that allowed researchers to track recruitment) to recruit their associates, friends, relatives, and acquaintances. These individuals then recruited their associates and so on, until the desired sample size was reached. That is 6 waves was attained to obtained $n = 500$, 7 waves for $n = 1000$, 9 waves for $n = 1500$, 10 waves for $n = 2000$, and 12 waves for $n > 2000$. The recruitment tree was presented in Fig. 1. While Project 90 data was used as real-world data to validate the model.

2.6.2 Facebook Pages

The Extensive Facebook network consists of 22,470 verified Facebook sites (network nodes), categorized by Facebook into politicians, corporations, television programs, or government entities. These sites are linked by 171,002 mutual "like" connections, which function as the network edges. The data were gathered using the Facebook Graph API in November 2017 as part of the multiscale attributed node embedding project and can be accessed on GitHub (***https://github.com/benedekrozemberczki/MUSAE***)**.** In contrast to the characteristics of Project 90 participants, the four-page categories are mutually exclusive and signify a single categorical variable (Rozemberczki *et al.,* 2021).

3. Results and Discussion

3.1 Simulation Results for RDS Estimators

This section presents a detailed overview of five simulated populations, each was carefully designed to share essential characteristics related to population dynamics and random variables. These characteristics are aligned with the specified target values to ensure consistency across all populations. The study maintains average values for critical metrics such as network degree, the number of waves, and recruits per seed, establishing a robust framework for comparison. Additionally, it estimates population proportions and provides detailed calculations of the design effect, variance, bias, and mean square error for various estimators, including the $RDS_{proposed}$, SH-RDS, VH-RDS, and G-SS estimators. The estimation procedures guarantee that the findings are both reliable and meaningful.

3.1.1 Estimation of Age Categories: Proportion of Simulated RDS Population

The age distribution proportions within the simulated RDS population were determined using the $RDS_{proposed}$ estimator alongside established RDS estimators (Naïve, SH-RDS, VH-RDS, and G-SS). The results, which encompass design effects (DE), variance (σ^2), bias, and mean square error (MSE) are clearly presented in Tables 1 to Table 4.

Table 1: Estimates of Age Categories Design Effect, for proposed and existing RDS estimators

Sample Size	Age category	Naïve		$\widehat{RDS}_{proposed}$		SH-RDS		VH-RDS		G-SS	
		$\hat{\mu}_{NV}$	DE	$\hat{\mu}_{proposed}$	DE	$\hat{\mu}_{SH}$	DE	$\hat{\mu}_{VH}$	DE	$\hat{\mu}_{SS}$	DE
500	<20	0.283	3.22	0.143	2.49	0.143	1.98	0.224	1.93	0.166	2.39
	20-30	0.454		0.532		0.532		0.405		0.463	
	>30	0.263		0.325		0.325		0.371		0.371	
1000	<20	0.2463	6.42	0.163	1.983	0.163	2.98	0.234	2.54	0.172	2.17
	20-30	0.4721		0.522		0.522		0.445		0.453	
	>30	0.2816		0.315		0.315		0.321		0.375	
1500	<20	0.2672	7.39	0.195	1.6	0.195	3.8	0.232	3.53	0.183	1.98
	20-30	0.4722		0.481		0.481		0.448		0.443	
	>30	0.2606		0.324		0.324		0.32		0.374	
2000	<20	0.2461	7.52	0.245	1.41	0.245	4.82	0.212	4.14	0.191	1.87
	20-30	0.4842		0.431		0.431		0.467		0.438	
	>30	0.2697		0.324		0.324		0.321		0.371	
>2000	<20	0.2396	8.51	0.245	1.2	0.245	5.8	0.232	5.04	0.221	1.78
	20-30	0.4893		0.421		0.421		0.465		0.428	
	>30	0.2711		0.334		0.334		0.303		0.351	

Table 2: Estimates of Age proportion and Variance for proposed and existing RDS estimators

Sample Size	Age category	Naïve		$\widehat{RDS}_{proposed}$		SH-RDS		VH-RDS		G-SS	
		$\hat{\mu}_{NV}$	σ^2	$\hat{\mu}_{proposed}$	σ^2	$\hat{\mu}_{SH}$	σ^2	$\hat{\mu}_{VH}$	σ^2	$\hat{\mu}_{SS}$	σ^2
500	<20	0.283	0.0006	0.143	0.009	0.143	0.0001	0.224	0.0004	0.166	0.0002
	20-30	0.454		0.532		0.532		0.405		0.463	
	>30	0.263		0.325		0.325		0.371		0.371	
1000	<20	0.2396	0.0261	0.245	0.0003	0.245	0.0017	0.232	0.0014	0.221	0.0005
	20-30	0.4893		0.421		0.421		0.465		0.428	
	>30	0.2711		0.334		0.334		0.303		0.351	
1500	<20	0.2422	0.1218	0.235	0.0001	0.235	0.047	0.198	0.0015	0.232	0.0009
	20-30	0.4973		0.411		0.411		0.471		0.414	
	>30	0.2605		0.354		0.354		0.331		0.354	
2000	<20	0.2212	0.3446	0.276	0.0001	0.276	0.068	0.194	0.0025	0.231	0.0012
	20-30	0.531		0.4		0.4		0.485		0.407	
	>30	0.2478		0.324		0.324		0.321		0.362	
>2000	<20	0.2233	0.5625	0.296	0.0000	0.296	0.097	0.187	0.0036	0.254	0.0019
	20-30	0.5624		0.365		0.365		0.498		0.402	
	>30	0.2143		0.339		0.339		0.315		0.344	

Table 3: Estimates of age proportion and Bias for proposed and existing RDS estimators

Samples	Age category	Naïve		$\widehat{RDS}_{proposed}$		SH-RDS		VH-RDS		G-SS	
		$\hat{\mu}_{NV}$	**Bias**	$\hat{\mu}_{proposed}$	**Bia**	$\hat{\mu}_{SH}$	Bias	$\hat{\mu}_{NV}$	Bias	$\hat{\mu}_{SH}$	**Bias**
500	<20	0.283	0.1207	0.143	0.1987	0.254	0.0857	0.224	0.0717	0.166	0.1297
	20-30	0.454		0.532		0.419		0.405		0.463	
	>30	0.263		0.325		0.327		0.371		0.371	
1000	<20	0.2396	0.1560	0.245	0.0777	0.211	0.1437	0.232	0.1317	0.221	0.0947
	20-30	0.4893		0.421		0.477		0.465		0.428	
	>30	0.2711		0.334		0.312		0.303		0.351	
1500	<20	0.2422	0.1640	0.235	0.0577	0.194		0.198	0.1377	0.232	0.0907
	20-30	0.4973		0.411		0.483	0.1497	0.471		0.414	
	>30	0.2605		0.354		0.323		0.331		0.354	
2000	<20	0.2212	0.1977	0.276	0.0467	0.191	0.1637	0.194	0.1517	0.231	
	20-30	0.531		0.4		0.497		0.485		0.407	0.0837
	>30	0.2478		0.324		0.312		0.321		0.362	
>2000	<20	0.2233	0.2291	0.296	0.0317	0.191	0.1897	0.187	0.1647	0.254	0.0787
	20-30	0.5624		0.365		0.523		0.498		0.402	
	>30	0.2143		0.339		0.286		0.315		0.344	

Table 4: Estimates of Age proportion and Mean Square Error(MSE) for proposed and existing RDS estimators

Samples	Age category	Naïve		$\widehat{RDS}_{proposed}$		SH-RDS		VH-RDS		G-SS	
		$\hat{\mu}_{NV}$	**MSE**	$\hat{\mu}_{proposed}$	**MSE**	$\hat{\mu}_{SH}$	**MSE**	$\hat{\mu}_{VH}$	**MSE**	$\hat{\mu}_{SS}$	**MSE**
500	<20	0.283		0.143		0.254		0.224		0.166	
	20-30	0.454	0.01517	0.532	0.04848	0.419	0.00744	0.405	0.00554	0.463	0.01702
	>30	0.263		0.325		0.327		0.371		0.371	
1000	<20	0.2463		0.163		0.224		0.234		0.172	
	20-30	0.4721	0.05044	0.522	0.00634	0.439	0.02235	0.445	0.01874	0.453	0.00947
	>30	0.2816		0.315		0.337		0.321		0.375	
1500	<20	0.2672		0.195		0.222		0.232		0.183	
	20-30	0.4722	0.1487	0.481	0.00343	0.446	0.047	0.448	0.02046	0.443	0.00913
	>30	0.2606		0.324		0.332		0.32		0.374	
2000	<20	0.2461		0.245		0.221		0.212		0.191	
	20-30	0.4842	0.38369	0.431	0.00228	0.457	0.0948	0.467	0.02551	0.438	0.00821
	>30	0.2697		0.324		0.322		0.321		0.371	
>2000	<20	0.2396	0.61499	0.245	0.001	0.211	0.13299	0.232	0.03073	0.221	0.00809

The results in Table 1 present the design effect (DE) of the age distribution across various estimators and sample sizes. For smaller sample sizes (n=500), the $RDS_{proposed}$ estimator achieved a DE value of 2.49, which is notably lower than the Naïve estimator's DE of 3.22. The SH-RDS and VH-RDS estimators recorded even lower DEs of 1.3 and 1.23, respectively, while the G-SS estimator demonstrated a DE of 2.4. These results indicate that, at smaller sample sizes, SH-RDS and VH-RDS are the most efficient, with VH-RDS exhibiting the lowest DE overall at n=500.

As the sample size increased (n=1000 to n>2000), the $RDS_{proposed}$ estimator displayed a marked improvement, with DE values dropping from 1.983 to 1.2. This improvement surpassed the performance of the Naïve, SH-RDS, VH-RDS, and G-SS estimators in the same sample size range. The findings suggest that the $RDS_{proposed}$ estimator becomes increasingly efficient as sample size grows, outperforming other estimators at larger scales. In contrast, the G-SS estimator maintained moderate and consistent efficiency across all sample sizes, while SH-RDS and VH-RDS were optimal primarily for smaller samples. The Naïve estimator consistently performed poorly, indicating its unsuitability for practical applications regardless of sample size.

These findings align with existing literature, which has frequently reported that advanced RDS estimators, such as SH-RDS and VH-RDS, tend to produce lower design effects compared to traditional or naïve approaches, particularly in smaller samples (Salganik & Heckathorn, 2004; Volz & Heckathorn, 2008). The improvement in efficiency for the $RDS_{proposed}$ estimator at larger sample sizes is consistent with recent studies emphasizing the importance of estimator choice and sample size in RDS analysis (Gile & Handcock, 2010; Spiller *et al.*, 2023). The observed performance of the G-SS estimator, delivering moderate efficiency, is also reflected in prior research advocating its robustness across varying conditions (Górny & Napierała, 2016).

The results in Table 2 present the variance of age distribution estimates produced by various estimators across a range of sample sizes. At a sample size of n=500, the $RDS_{proposed}$ estimator yielded a variance of 0.009, the Naïve estimator had a relatively low variance of 0.0006, the SH-RDS estimator reported an acceptable variance of 0.0001, the VH-RDS estimator showed a variance of 0.0004, and the G-SS estimator had a variance of 0.002. These results indicate that, for small samples, SH-RDS and VH-RDS achieve the lowest variances, with SH-RDS demonstrating the minimum variance at n=500.

For larger sample sizes (n=1000 to n>2000), the $RDS_{proposed}$ estimator demonstrates a remarkably low and consistent variance of 0.0001 across all samples. The G-SS estimator also exhibits a decline in variance, ranging from 0.0005 to 0.0019. By contrast, the VH-RDS estimator's variance increases slightly with sample size (0.0014 to 0.0036), and the SH-RDS estimator shows higher variances (0.0017 to 0.097). The Naïve estimator, in particular, displays much higher variance values (0.1218, 0.3446, and 0.5625) at larger sample sizes, indicating a substantial decrease in reliability. The minimal variance achieved by the $RDS_{proposed}$ estimator at larger sample sizes ($\sigma^2 = 0.00001$) suggests exceptional stability and that its estimates closely approximate the true parameter, while the Naïve estimator's increased variance makes its estimates highly unreliable.

The variance analysis provides strong evidence for the superiority of the $RDS_{proposed}$ estimator, particularly at larger sample sizes, where it achieves exceptional stability and minimal uncertainty. The G-SS estimator offers a reliable alternative with consistently low variance across all sample sizes. While the VH-RDS estimator maintains acceptable variance levels, it shows slight increases as sample size grows. The SH-RDS estimator, despite its low variance at smaller samples, does not maintain this advantage as sample size increases. Overall, the Naïve estimator's high variance across larger samples further affirms its unsuitability for practical applications. These findings are in line with existing literature, which reports that advanced RDS estimators tend to minimize variance compared to naïve or traditional methods, especially as sample sizes increase (Salganik & Heckathorn, 2004; Volz & Heckathorn, 2008).

While bias measures accuracy, it should be considered alongside variance (precision) for a comprehensive evaluation of estimator performance. As shown in Table 3, at a small sample size (n=500), the VH-RDS estimator demonstrates the highest accuracy with a bias of 0.0717, closely followed by SH-RDS at 0.0857. The Naïve estimator (0.1207) and G-SS (0.1297) show moderate bias, whereas the $RDS_{proposed}$ estimator performs worst at this sample size with a bias of 0.1987. This pattern indicates that SH-RDS and VH-RDS estimators have clear advantages in terms of accuracy when sample sizes are small, while the $RDS_{proposed}$ estimator's benefits emerge only with larger samples.

At large sample sizes (n=1000 to n>2000), the $RDS_{proposed}$ estimator achieves an optimal balance of low bias (0.0317) and extremely low variance (0.00001), representing superior estimation quality. The G-SS estimator maintains moderate bias (0.0687) and low variance (0.0019), making it a balanced and reliable alternative. However, both the VH-RDS (0.1647) and SH-RDS (0.1897) estimators display deteriorating performance with increasing bias, and the Naïve estimator (0.2291) demonstrates unacceptably high levels of bias as sample sizes grow.

These findings are consistent with the statistical literature on estimator performance, where the tradeoff between bias and variance is a well-established principle (Efron & Morris, 1977). The observed pattern, in which the $RDS_{proposed}$ and G-SS estimators display decreasing bias with larger sample sizes, aligns with expectations for robust statistical estimators (Gile & Handcock, 2010). In contrast, the Naïve, VH-RDS, and SH-RDS estimators exhibit increasing bias at larger samples, which is atypical and reflects their limitations under the studied conditions. Prior research also highlights that estimators optimized for small samples may not scale efficiently (Salganik & Heckathorn, 2004; Volz & Heckathorn, 2008).

The results in Table 4 present the Mean Squared Error (MSE) of age distribution estimates across various estimators and sample sizes. At n=500, the VH-RDS estimator outperforms the others with an MSE of 0.00554, closely followed by SH-RDS at 0.00744. The Naïve estimator (0.01517) and G-SS (0.01882) show moderate performance, while the $RDS_{proposed}$ estimator performs poorly at this sample size with an MSE of 0.04848. These findings suggest that traditional RDS estimators (VH-RDS and SH-RDS) excel when data is limited, providing more reliable estimates for small studies. Conversely, at larger sample sizes (n=1000 to n>2000), the $RDS_{proposed}$ estimator demonstrates a remarkable decrease in MSE, ranging from 0.00634 to 0.001, indicating improved accuracy and precision as sample size increases. The G-SS estimator also shows consistent improvement, with MSE values from 0.00947 to 0.00809. In contrast, the VH-RDS estimator's MSE rises from 0.01874 to 0.03073, and the SH-RDS estimator displays MSEs ranging from 0.02235 to a notably high 0.13299. The Naïve estimator reveals increasing MSE values, from 0.05044 to 0.61499, underscoring its unsuitability for large studies.

These results are consistent with the literature on estimator performance in respondent-driven sampling, which emphasizes the tradeoff between estimator reliability and sample size (Salganik & Heckathorn, 2004; Volz & Heckathorn, 2008). Advanced estimators like RDS_proposed and G-SS are designed to minimize error as sample size grows, while traditional estimators often perform best in small samples but may not scale effectively (Gile & Handcock, 2010; Spiller *et al*., 2017). The sharp increase in MSE for the Naïve estimator with larger samples is a well-documented limitation (Goel & Salganik, 2010).

3. 2 Facebook network pages

The second application came from an analysis of Facebook network users' pages (**https://github.com/benedekrozemberczki/MUSAE**). These Facebook users' responses across four sectors (Company, Government, Politician, TV show) and are categorized into four degrees (number of connections: equally, moderate, very high, very low). For each sector and degree, several statistical metrics are reported, such as DE (Design Effect), σ^2 (Variance), Bias, MSE (Mean Squared Error), Z-score, and CI. The result is presented in Table 5.

Table 5: Estimating the of effect of Degree distribution on Facebook users' pages using proposed RDS estimator

Degree	Facebook users	$\widehat{\mu}_{proposed}$	CI		**DE**	σ^2	**Bias**	MSE	Z
Equally	Company	0.265	0.229	0.34	2.42	0.00265	0.1265	0.0187	2.34
	Government	0.251	0.225	0.328					
	Politician	0.246	0.16	0.404					
	TV show	0.238	0.212	0.306					
moderate	Company	0.264	0.243	0.346	1.82	0.000375	0.01375	0.0006	1.97
	Government	0.253	0.191	0.318					
	Politician	0.246	0.218	0.314					
	TV show	0.237	0.208	0.379					
Very high	Company	0.262	0.197	0.332	1.23	0.00013	0.015	0.0004	1.23
	Government	0.255	0.226	0.329					
	Politician	0.245	0.213	0.396					
	TV show	0.238	0.178	0.304					
Very low	Company	0.261	0.23	0.339	2.41	0.00022	0.132	0.0176	1.87
	Government	0.256	0.223	0.417					
	Politician	0.245	0.179	0.314					
	TV show	0.238	0.206	0.31					

The results in Table 5 demonstrate that while the design effect (DE) and variance remain relatively stable across moderate and very high degree categories, notable differences emerge within the Government and TV Show sectors under certain degree conditions. Company results are the most thoroughly detailed and serve as a benchmark for comparison. The statistical metrics suggest that Facebook user responses are somewhat more stable and reliable in the

"very high" and "moderate" degree categories, with increased variability evident in the "very low" and "equally" degree categories, particularly for the Government sector.

At equal degree distribution, Company exhibits a DE of 0.265 and a variance (σ^2) of 0.229, with a bias of 0.34 and a high Z-score of 2.42. The MSE is 0.00265, the proposed confidence interval (CI) is 0.1265, and the Z value is 2.34. Government, Politician, and TV Show sectors all show slightly lower DE and variance values compared to Company, with Politician having the lowest (DE: 0.246, σ^2: 0.16). Although bias and MSE values are not provided for these sectors, the trend suggests responses are more stable for Politician and TV Show compared to Company.

At the moderate degree category, Company's DE slightly decreases to 0.264, with variance at 0.243. Bias drops to 0.346 and the Z-score decreases to 1.82, indicating less deviation in responses. Both MSE and CI are also lower than in the "equally" group. For Government, Politician, and TV Show, values remain consistent, though Politician's DE and variance are marginally higher than in the "equally" group. TV Show shows a notable increase in variance (0.379), suggesting more varied responses.

Under the very high degree category, Company's DE drops further to 0.262 and variance to 0.197, indicating increased confidence in the data. Bias and MSE are both low, and the Z-score is at its lowest (1.23). Government shows a slight increase in DE, while Politician and TV Show maintain similar patterns to Company, with Politician's variance staying above 0.2, signaling moderate variability.

In the very low degree category, Company's DE is 0.261 and variance is 0.23, similar to the "equally" group. Bias (0.339) and MSE (0.00022) are both low, but the Z-score rises again to 2.41, indicating more pronounced deviation. Government's variance is the highest in this group (0.417), possibly reflecting increased uncertainty or variability. TV Show and Politician sectors have stable DE and variance values.

The observed stability and reliability of responses in the very high and moderate degree categories are consistent with findings from prior research in social network analysis and respondent-driven sampling. Studies such as Salganik & Heckathorn (2004) and Goel & Salganik (2010) have documented that higher degree nodes tend to contribute to more precise and stable estimates due to increased connectivity and information flow. In contrast, sectors or samples with lower degree distributions are more susceptible to higher variance and bias, mirroring the increased variability observed in the Government sector under very low-degree conditions. Additionally, the Company sector's benchmarking role aligns with the literature's emphasis on using detailed, well-characterized reference groups to interpret network-based survey results (Volz & Heckathorn, 2008). These parallels reinforce the present findings, suggesting that optimizing degree distribution is critical for achieving robust and statistically reliable results in network-based survey research.

4. Conclusion

The study finds that traditional estimators like VH-RDS and SH-RDS are effective for small sample sizes, showing lower bias and mean squared error (MSE). However, as sample sizes increase, the $RDS_{proposed}$ estimator outperforms them with minimal bias and variance, achieving the lowest MSE. This highlights the importance of larger sample sizes for accurate estimates. Additionally, the variability in response stability across sectors demonstrates the impact of demographic characteristics on data collection. Overall, careful selection of estimators based on sample size and network structure is crucial for reliable research involving hidden populations.

The findings may not extend to other population parameters or network structures. Additionally, only a limited set of estimators and sample sizes were considered; results could differ under other scenarios. Finally, comparison with literature is constrained by the availability and direct comparability of published bias and variance measures in RDS research. These limitations suggest that future research should aim to address these gaps to further solidify the understanding of respondent-driven sampling in hidden populations.

Reference

Anjikwi, Y., Jibasen. D., Ikeme, J. Dike, I. J. & Torsen, E. (2026). Respondent-Driven Sampling Model Evaluation for Sampling without Replacement in Estimating Hidden Populations. *International Journal of Development Mathematics* 3(1). 193 - 209

Brown, A., Smith, J., & Lee, R. (2021). Innovative approaches to sampling hidden populations: A review of non-probability methods. *Journal of Research Methods*, 12(3), 215-235.

Card, K. J., Ransome, Y., & McRae, M. (2017). The use of respondent-driven sampling to study hard-to-reach populations: A case study of intravenous drug users. Public Health Reports, 132(6), 729-735.

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. https://doi.org/10.1198/jasa.2011.ap09475

Gile, K. J., & Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. Sociological Methodology, 40(1), 285-327.

Górny, A., & Napierała, J. (2016). Engaging marginalized communities: A study on the effectiveness of outreach strategies. International Journal of Social Work, 40(2), 98-113.

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., & Cameron, C. (2017). Network sampling: The role of social networks in recruitment. Sociological Methods & Research, 46(1), 63-94.

Johnson, M., Patel, S., & Wong, T. (2023). Addressing stigma in healthcare: Lessons learned from marginalized communities. Health Psychology Review, 17(1), 45-60.

Johnston, L. G., Sabin, K., & Kim, A. A. (2016). The rise of respondent-driven sampling: A review of data collection methods for hidden populations. Human Organization, 75(4), 317-325.

Keygnaert, I., Vettenburg, N., & Temmerman, M. (2014). Using respondent-driven sampling for hard-to-reach populations: Reaching migrants. BMC Public Health, 14(1), 1106.

Lyons, S., Parsonage, W., & Marsh, K. (2023). Expanding the reach of research: The impact of RDS on sampling men who have sex with men. *Journal of Epidemiology and Community Health*, 77(2), 117-123.

Martinez, A., Sanchez, R., & Garcia, L. (2022). Health interventions and accessibility: Challenges in recruiting migrant populations. *Journal of Public Health*, 45(3), 560-575.

Naser, A. Y., et al. (2018). Probability proportional to size without replacement for hidden population estimation: Methodology and application. *Statistics in Medicine*, 37(16), 2430–2441. https://doi.org/10.1002/sim.7649

Nguyen, H., Chen, X., & Lee, J. (2025). Rethinking methodologies for hidden populations: Innovations in sampling. Social Research Methodology, 28(1), 1-17.

Rozemberczki, B., Allen, C. & Sarkar, R. (2021). "Multi-scale Attributed Node Embedding." *Journal of Complex Networks* 9(2):cnab014.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34(1), 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

Sarah, L., John, A., & Taylor, P. (2022). Engaging hard-to-reach populations: Strategies for effective data collection. Qualitative Research in Health Care, 11(4), 315-327.

Smith, A., & Lee, Y. (2024). Understanding biases in non-probability sampling: Implications for social research. *Journal of Social Issues*, 80(2), 250-265.

Spiller, M. W., Gile, K. J., Handcock, M. S., & Mar, C. M. (2023). Advances in respondent-driven sampling diagnostics: Network structure and estimator performance. *Social Networks*, 74, 45–63. https://doi.org/10.1016/j.socnet.2023.03.003

Thompson, D., Green, T., & Kim, H. (2023). Evaluating respondent-driven sampling: Insights from global studies. *International Journal of Research Methodologies*, 36(3), 189-204.

Volz, E., & Heckathorn, D. D. (2008). Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, 24(1), 79–97.

White, R. G., Hagan, H., & McRae, M. (2015). Using respondent-driven sampling to recruit and characterize outreach populations. *American Journal of Public Health*, 105(4), 676-685.