# Sampling Hard-to-Reach Populations

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

UCLA

*Some joint work with*

Krista J. Gile Katherine R. McLaughlin Brian Kim
Ian E. Fellows Henry F. Raymond

Supported by NIH NICHD Grant HD041877 and NSF award MMS-0851555.

*7th Italian Conference on Survey Methodology*
*Session on Sampling Network Populations, 9 June, 2022*

# Challenges to Traditional Survey Sampling

- Eroding survey response rates and non-ignorable non-response
  - computation of inclusion probabilities is difficult
  - estimation of inclusion probabilities is required

- Allure of model-assisted and model-based modes of inference
  - models require assumptions
  - assumptions may be difficult to validate

# Some types of non-probability samples

General types:

- Convenience sampling: no formal design, the goal is acquisition
- Online panels: using internet, social media, etc, to select and recruit
- Sample matching: stratify on important population characteristics
- Network sampling: applicable when the population is networked

Modes of inference for non-probability samples:

- Quasi-randomization: Design-based via a model for inclusion probabilities: $\pi_i = P(S_i = 1)$
- Super-population: Model-based via a model for outcomes: $P(Y_i)$

# Hard-to-Reach Population Sampling: Motivating Questions

- What proportion of sub-Saharan migrants to Morocco have children?
- What proportion of semi-rural people are at high-risk for opioid addiction?
- What proportion of unregulated workers in New York City experience workplace violations of code?
- What proportion of Injecting Drug Users in Kampala are HIV Positive?
- What proportion of sex workers in rural China belong to ethnic minorities?

Limitation: No practical conventional sampling frame.

# Adaptive Network Sampling

Suppose:

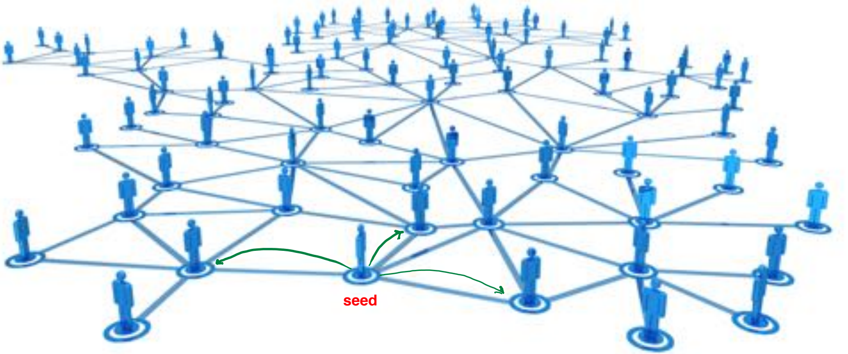
- The population is joined by informal social network of relationships.
- Researchers can access some members of the population.

*Sampling design*:

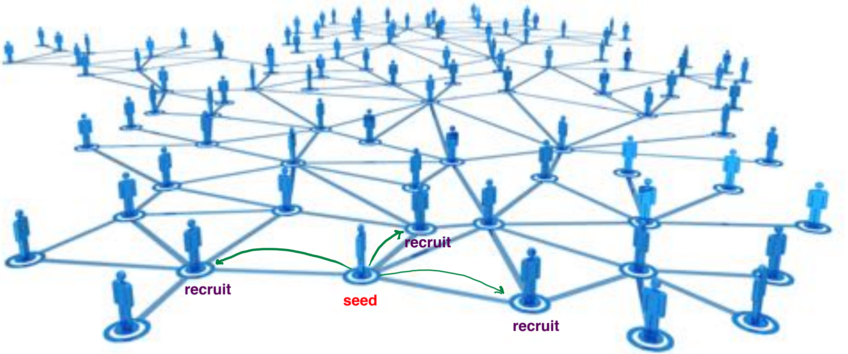
- Begin with a reachable (convenience) sample (the ***seeds***)
- Expand the sample by the researchers sampling those tied to those already in the sample.
  a process called *link tracing*.

# Start with a seed person

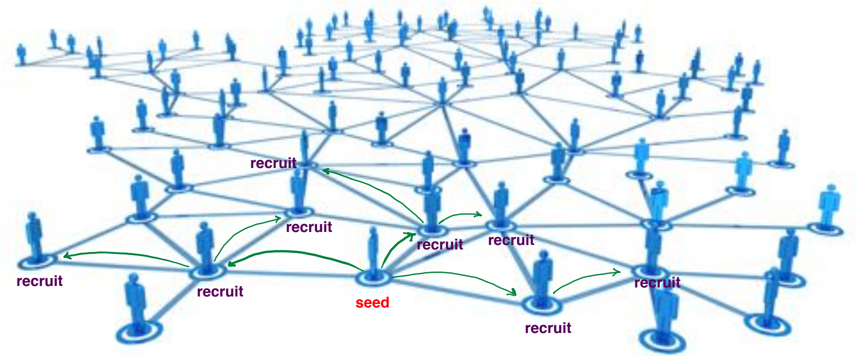
# Contact other people via the seed's social network



# Contact other people via the seed's social network



# Contact other people via the seed’s social network



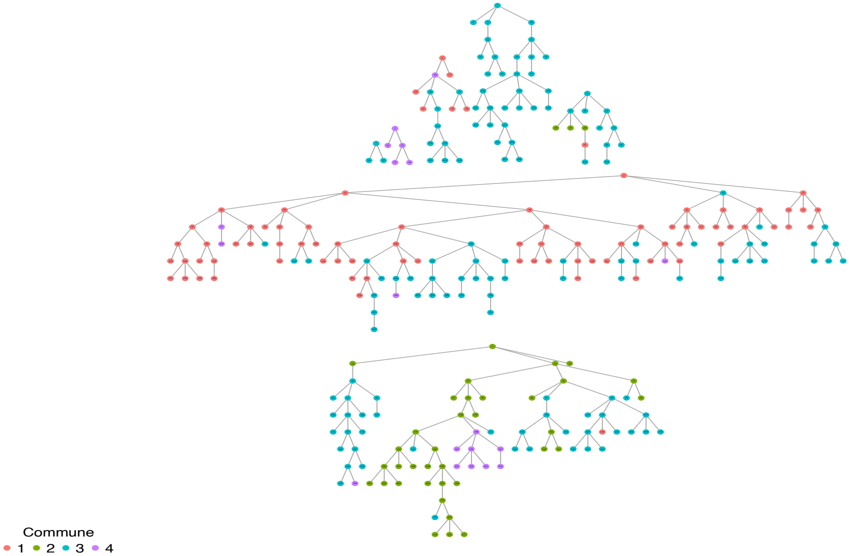


Figure: Graphical representation of the recruitment tree for the sampling of PWID. The nodes are the respondents and the wave number increases as you go down the page. The node color indicates the geographic neighborhood.

# Adaptive Network Sampling

*Strengths*:

- Exploits information in the network of relationships
- Network structure used to improve the design
- Increases the range of possible designs
- Adjusts for discovered features in the population
- Leads to increased efficiency of sampling

# Adaptive Network Sampling

*Strengths*:

- Exploits information in the network of relationships
- Network structure used to improve the design
- Increases the range of possible designs
- Adjusts for discovered features in the population
- Leads to increased efficiency of sampling

*Issues*:

- Seed Dependence: final sample depends on sampling mechanism of seeds
- Privacy: some populations prefer to stay "hidden"
- Link-tracing can be challenging: confidentiality, logistics
- Estimation: The sample and sampling probabilities depend on the unknown network

# Sampling depends on network: design-based

Observable sampling probabilities:

| Sampling | Nodal Probabilities $\pi_i$ | | Dyadic Probabilities $\pi_{ij}$ | |
|---|---|---|---|---|
| Scheme | Undirected | Directed | Undirected | Directed |
| Simple Random | Yes | Yes | Yes | Yes |
| One-Wave | Yes | No | No | No |
| $k-$Wave, $1 < k < \infty$ | No | No | No | No |
| Saturated | Yes | No | No | No |

(Unconditional) sampling probabilities unknown for many simple sampling strategies

*Snijders, T.A.B., 1992, "Estimation on the basis of snowball samples: how to weight." Bulletin Methodologie Sociologique, 36, 59-70.*
*Handcock, M.S. and K.J. Gile, 2010, "Modeling social networks from sampled data." , Annals of Applied Statistics, 4, Number 1, 5-25.*

# A peculiar case: Respondent-Driven Sampling

- *Sampling design*: Require respondents to choose from among their social circle rather than the researcher chooses.
- *Seed Dependence*: follow only a few links from each sampled
- *Privacy*: ***respondent-driven:*** respondents distribute uniquely identified coupons. no names.
- *Link-tracing*: none by researchers, done by respondents.
- *Estimation*: Challenging to get valid estimates
- Effective at obtaining large varied samples in many populations.
- Widely used: over 100 studies, in over 30 countries. Often HIV-risk populations.

*Heckathorn, D.D., "Respondent-driven sampling: A new approach to the study of hidden populations." Social Problems, 1997.*

*Handcock, M.S. and K.J. Gile, "On the Concept of Snowball Sampling." Sociological Methodology, 2011.*

# Classic Design-Based Inference: Generalized Horvitz-Thompson Estimators

- Goal: Estimate the population mean of $y$:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$$

where

$$y_i = \begin{cases} 1 & i \text{ ``positive"} \\ 0 & i \text{ ``negative"}. \end{cases}$$

- Hajek Estimator:

$$\hat{\mu} = \frac{\sum_i \frac{S_i}{\pi_i} y_i}{\sum_i \frac{S_i}{\pi_i}}$$

where

$$S_i = \begin{cases} 1 & i \text{ sampled} \\ 0 & i \text{ not sampled} \end{cases} \qquad \pi_i = P(S_i = 1).$$

- The key point: Estimator requires $\pi_i = P(S_i = 1) \quad \forall i : S_i = 1$

# One Approach: Random walk approximation

Respondent-driven Sampling:

- Approximate link-tracing process by a Markov chain representation
- Assume sample can be treated as from stationary distribution
- Then sampling probabilities proportional to degree.

Volz-Heckathorn Estimator (VH): inverse probability weighted by degrees

$$\hat{\mu}_{\mathrm{VH}} = \frac{\sum_i S_i \frac{y_i}{d_i}}{\sum_i S_i \frac{1}{d_i}}$$

where $d_i$ = degree of node $i$, $S_i$ sample indicator, $y_i$ quantity of interest.

Volz, E., and D.D. Heckathorn, "Probability Estimation Theory for Respondent Driven Sampling," *Journal of Official Statistics,* 2008.

# Improvements to RDS design-based estimators

The key is the modeling of the sampling process

- Salganik and Heckathorn (2004): simple Markov Chain model over classes. Struggles with Seed bias and finite population, good on homophily
- Volz and Heckathorn (2008): Markov Chain model over people. Seed bias, finite population, differential activity, homophily
- Gile (2008, 2011): Develops a model based on the successive sampling of people in time.
Adjusts for without-replacement and finite population effects
- Fellows (2018) introduced the homophily configuration graph (HCG) estimator that has the good features of the SH and SS estimators.

# Fitting Models to Partially Observed Social Network Data

- Focus on the joint distribution of $Z = (Y, X)$.
- Types of data: Observed relations, nodal and dyadic variables ($z_{obs} = (y_{obs}, w_{obs})$), and indicators of relations and covariates
- $Z = (Z_{obs}, Z_{unobs})$

$$
\begin{aligned}
L(\boldsymbol{\eta}, \psi) &\equiv P(Z_{obs} = z_{obs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{z_{unobs}} P(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{z_{unobs}} P(D|Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}) \\
&= \sum_{z_{unobs}} P(D|Z, \psi) \times P_{\boldsymbol{\eta}}(Z = z)
\end{aligned}
$$

sampling design × network model

- $\boldsymbol{\eta}$ is the network model parameter (“super population")
- $\psi$ is the sampling parameter

UCLA

# Adaptive Sampling Designs

- A sampling design adaptive if:

$$P(D = d|Z_{obs}, Z_{mis}, \psi) = P(D = d|Z_{obs}, \psi) \qquad \forall z \in \mathcal{Z}.$$

that is, it uses information collected during the survey to direct subsequent sampling, but the sampling design depends only on the observed data.

- adaptive sampling designs satisfy a "*missing at random*" condition from Rubin (1976) in the context of missing data.
- **Result:** standard network sampling designs such as conventional, adaptive web, and multi-wave link-tracing sampling designs are adaptive
  ⇒ Thompson and Frank (2000), Handcock and Gile (2007).

UCLA

# When is sampling non-adaptive?

- Individual sample based on unobserved properties of non-respondents - like infection status or illicit activity.
- Link-tracing sample starting where links are followed dependent on unobserved properties of alters.

# Adaptive Sampling Designs and their Amenable Models

Definition: Consider a sampling design governed by parameter $\psi \in \Psi$ and a stochastic network model $P_{\boldsymbol{\eta}}(Z = z)$ governed by parameter $\boldsymbol{\eta} \in \Xi$. We call the sampling design amenable to the model if the sampling design is adaptive and the parameters $\psi$ and $\boldsymbol{\eta}$ are distinct.

Result: If the sampling design is amenable to the model the likelihood for $\boldsymbol{\eta}$ and $\psi$ is

$$L[\boldsymbol{\eta}, \psi | Z_{obs} = z_{obs}, D = d] \propto L[\psi | D = d, Z_{obs} = z_{obs}] L[\boldsymbol{\eta} | Z_{obs} = z_{obs}]$$

sampling design likelihood × face-value likelihood

$$L[\psi | D = d, Z_{obs} = z_{obs}] = P(D | Z_{obs} = z_{obs}, \psi)$$

$$L[\boldsymbol{\eta} | Z_{obs} = z_{obs}] = \sum_{z_{unobs}} P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

UCLA

# Adaptive Sampling Designs and their Amenable Models

Result: If the sampling design is ***not*** amenable to the model the likelihood for $\boldsymbol{\eta}$ and $\psi$ is

$$L(\boldsymbol{\eta}, \psi) = \sum_{z_{unobs}} P(D | Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

and the design will need to be represented.

Clearly $P(D|Z, \psi)$ can be modeled when it is unknown.

# Doing better: A Network Model-Based Estimator

- Fit a network model to observed data (ERGM, using `statnet` R package)
- Estimate sampling probabilities based on network model, and weight sample appropriately
- Can estimate conditional on seed selection, to reduce bias induced by seed selection.

# Network Model-Assisted Estimator

- Interested in sampling probabilities $\pi_i = P_y(S_i = 1)$.
- Should reflect:
  - Nodal degree $\boldsymbol{d_i}$
  - Sample fraction
  - Seed selection
  - Homophily and branching structure of sampling
- This is very difficult to do without known the underlying social network $\boldsymbol{y}$
- So we develop a “super-population” representation for $\boldsymbol{y}$ with the purpose of “assisting” the design-based inference

# Network Model-Assisted Estimator

- Approach: Retain design-based framework, but estimate the unknown finite-population sampling probabilities $\pi_i(y) = \mathbf{E}(S_i | Y = y)$.

Idea:

1. For given network $\boldsymbol{y}$, can compute

$$\pi_i(y) = \mathbf{E}(S_i | Y = y)$$

2. Estimate $\pi_i$ via

$$\hat{\pi}_i = \sum_{y_{unobs}} \pi_i(y) P_\eta(Y = y | Y_{obs} = y_{obs})$$

3. We do not know $\eta$, so we estimate it from the data.

# The ERGM Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of $Y$ e.g. $\{0,1\}^N$
and $\mathcal{X}$ be the sample space of $X$.
Model the multivariate distribution of $Y$ given $X$ via:

$$P_\eta(Y = y|X = x) = \frac{\exp\{\eta{\cdot}g(y|x)\}}{c(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}, \quad x \in \mathcal{X}$$

Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^d$ $d$-vector of parameters
- $g(y|x)$ $d$-vector of *graph statistics*.
  $\Rightarrow$ $g(Y|x)$ are jointly sufficient for the model
- $c(\eta, x, \mathcal{Y})$ distribution normalizing constant

$$c(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta{\cdot}g(y|x)\}$$

# Extensive development of conditional models

- Classes of $g(y|x)$ (Generative Theory, Structural signatures)
- Inference on the loglikelihood function,

$$\ell(\eta|y_{\mathrm{obs}}; x_{\mathrm{obs}}) = \eta{\cdot}g(y_{\mathrm{obs}}|x_{\mathrm{obs}}) - \log c(\eta|x_{\mathrm{obs}})$$

$$c(\eta|x_{\mathrm{obs}}) = \sum_{z \in \mathcal{Y}} \exp\{\eta{\cdot}g(z|x_{\mathrm{obs}})\}$$

- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)

# Returning to RDS: Fitting the Model

Use networks statistics $g(y)$:

- $\{y_{1+}, y_{2+}, \ldots, y_{n+}\}$, the ***degree sequence*** of the network.
- the number of ties between those positive and those negative.

Problem: Requires (unknown) networked population statistics $g(y)$.
Solution: Use design-based estimators

$$\hat{g}(\eta) = \sum_{i=1}^{N} \frac{\mathbf{S}_i \tilde{g}(y_{obs})}{\hat{\pi}_i}$$

where $\tilde{g}(y_{obs})$ are corresponding sample statistics.

Problem: This, in turn, requires sampling probabilities.
Solution: Novel iterative algorithm to find self-consistent solution.

UCLA

# Model-Assisted Estimator: Algorithm

- Goal: Estimate sampling probabilities ($\pi_i$).
- A function of homophily ($\eta$), and population of degrees and infection N.
- Initiate via $\hat{\pi}_i$ estimated by simple rule.
- Iterate the following steps:
  - Estimate $\hat{g}(\eta)$ using $\hat{\pi}_i$.
  - Find corresponding model parameter $\eta$ (`ergm` R package)
  - Simulate $M$ networks, and samples from networks. Estimate $\hat{\pi}_i$ by simulation.
- Use the resulting estimated probabilities, $\hat{\pi}_i$, to form weighted estimator.

$$\hat{\mu}_{\mathrm{MA}} = \frac{\sum_i S_i \frac{\mathbf{y}_i}{\hat{\pi}_i}}{\sum_i S_i \frac{1}{\hat{\pi}_i}}.$$

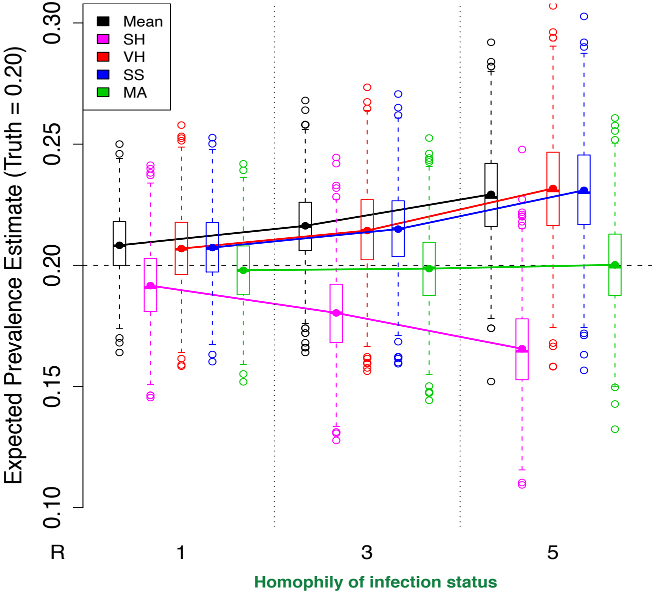
# Standard Error Estimation

Population Bootstrap:

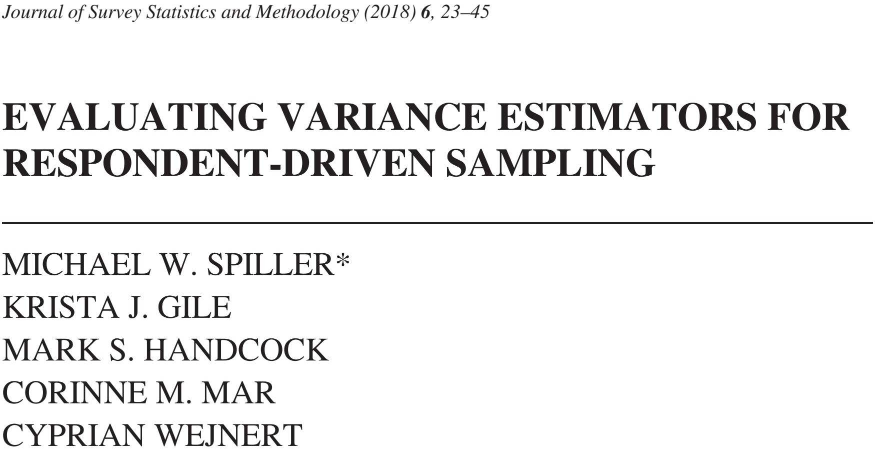
- Simulate *M* populations
  - Conditional on $z_{obs}$
  - Use model parameter $\eta$
- Simulate adaptive network samples over each population
- Compute MA estimates. Average estimates over M populations
- Results:
  - Performs well when statistics are homophily, degree distribution and differential activity (*w*)
  - Computationally expensive

*Krista J. Gile and Mark S. Handcock, "Network Model-Assisted Inference from Respondent-Driven Sampling Data." JRSS, A, 178, 3, 619-639, 2015.*

# Simulation Study
## comparing design-based to model-assisted estimators



# More and deeper studies



*Journal of Survey Statistics and Methodology* (2018) **6**, 23–45

## EVALUATING VARIANCE ESTIMATORS FOR RESPONDENT-DRIVEN SAMPLING

MICHAEL W. SPILLER*
KRISTA J. GILE
MARK S. HANDCOCK
CORINNE M. MAR
CYPRIAN WEJNERT

# Spiller *et. al* 2018

- The first systematic evaluation of the different RDS variance estimators
- Evaluation based on statistical performance on realistic but simulated populations
- Compare over simulated populations close to those of interest to the CDC
- Based on the CDC's National HIV Behavioral Surveillance system (NHBS)
  - NHBS sampled persons-who-inject-drugs (PWID) in 20 U.S. cities in both 2009 and 2012
  - Different surveys for heterosexuals, MSM and PWID, ongoing, 5 rounds since 2003.
  - a standardized protocol is used
  - 40 populations are simulated using information from the $2 \times 20 = 40$
- Primary focus is on estimates of confidence intervals (i.e., *coverage*)
- Confidence intervals are the primary RDS estimates

# Conclusions:  Methodological

- Coverage of nominally 95% RDS CI are usually above 90%
- Suggests that accurate RDS CI estimation is feasible
- The SS/SS-BS combination performed the best
  - the SH and VH CI estimators are poor when differential activity is low and homophily is high
- Fellows (2018) introduced the homophily configuration graph (HCG) estimator. It is model assisted and based on the homophily configuration graph model. It has the good features of the SH and SS estimators. It appears to be the best estimator.
- Improvements for cases with extreme low prevalence possible using alternative CI types
  - the combined Agresti-Coull and the bootstrap-***t*** interval of Mantalos and Zografos (2008)
- Both CI coverage rates and design-effects are often acceptable but not perfect.

# Conclusions: Broader

- Focus on the *effective sample size* rather than *design effect* or *sample size*.
- CI are a lower bound on the actual uncertainty
- The studies suggest that they are close to the actual uncertainty if the sampling is executed well.
- The availability of estimates and sensitivity methods in *user-friendly software* is an essential research contribution

# Strengths and weaknesses of Respondent-Driven Sampling

Strengths:

- Effective at obtaining large varied samples in many populations.
- Can be used in situations where a sampling frame does not exist.
- Unlike other link-tracing methods, does not require initial probability sample.
- Widely used: over 150 studies, in over 30 countries. Often populations at high risk for HIV.

Weaknesses:

- Still subject to many assumptions, especially data quality
- The degree to which it can be considered a probability sample depends on the quality of the implementation and network characteristics.
- Requires case-specific sensitivity analysis to justify its validity.

*Gile, K.J., and M.S. Handcock,* "*Respondent-Driven Sampling: An Assessment of Current Methodology,*" Sociological Methodology, ***40, 2010, available on arXiv***.