

# Hard-to-survey populations



---

## Overview of methods and applied research

Angelo Cozzubo

Research Programs Committee

October 2023

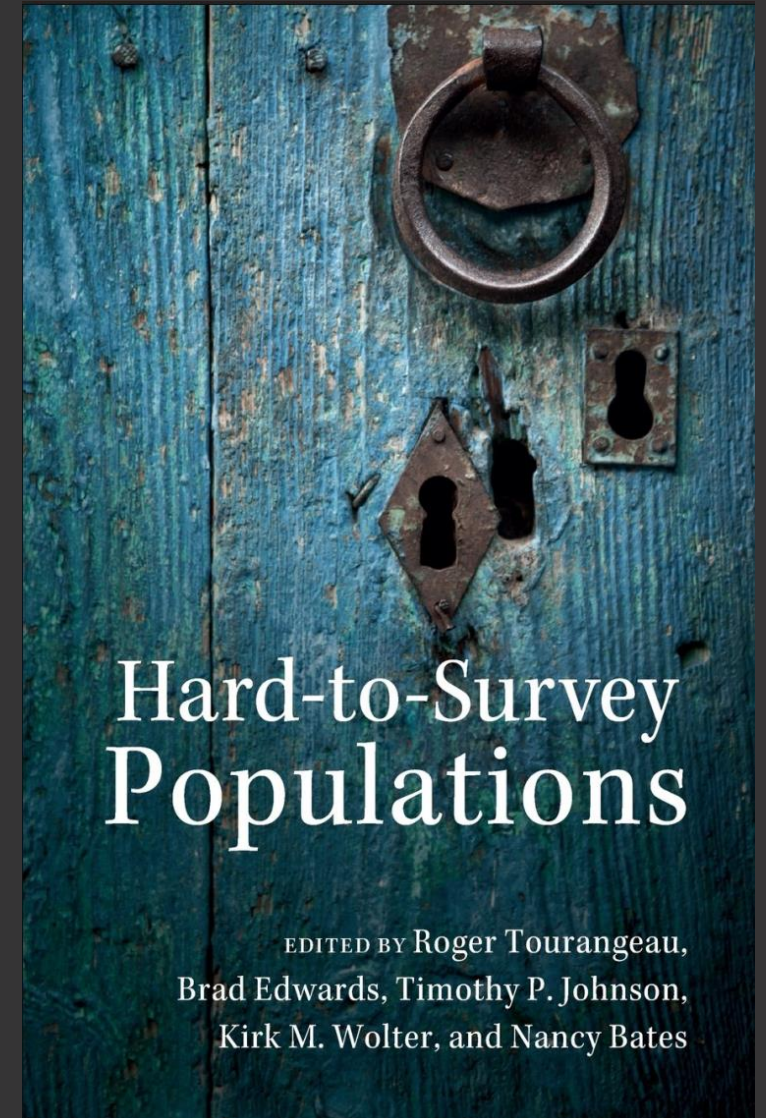


## What are Hard-to-survey Populations (HSP)?

- **Problem:** populations that present various challenges that make them harder to survey (Tourangeau et al., 2014)

### Challenges:

- Hard-to-sample: rare populations (homeless)
- Hard-to-identify: hidden by stigma, sensitivity and motivated misreporting (men who have sex with men)
- Hard-to-reach: difficult contact, group do not want to be identified, barriers to locate (undocumented immigrants)
- Hard-to-persuade: unwilling to answer (high income)
- Hard-to-interview: health impairments, language (prisoners)





## Hard-to-reach populations (HRP)

**Problem:** populations that are difficult to contact, that do not want to be identified, or have barriers to access

### Challenges:

- No list frame to draw the sample from
  - High mobility of the individuals
  - Hard to locate in the territory
- Barriers: restricted, dangerous or remote areas

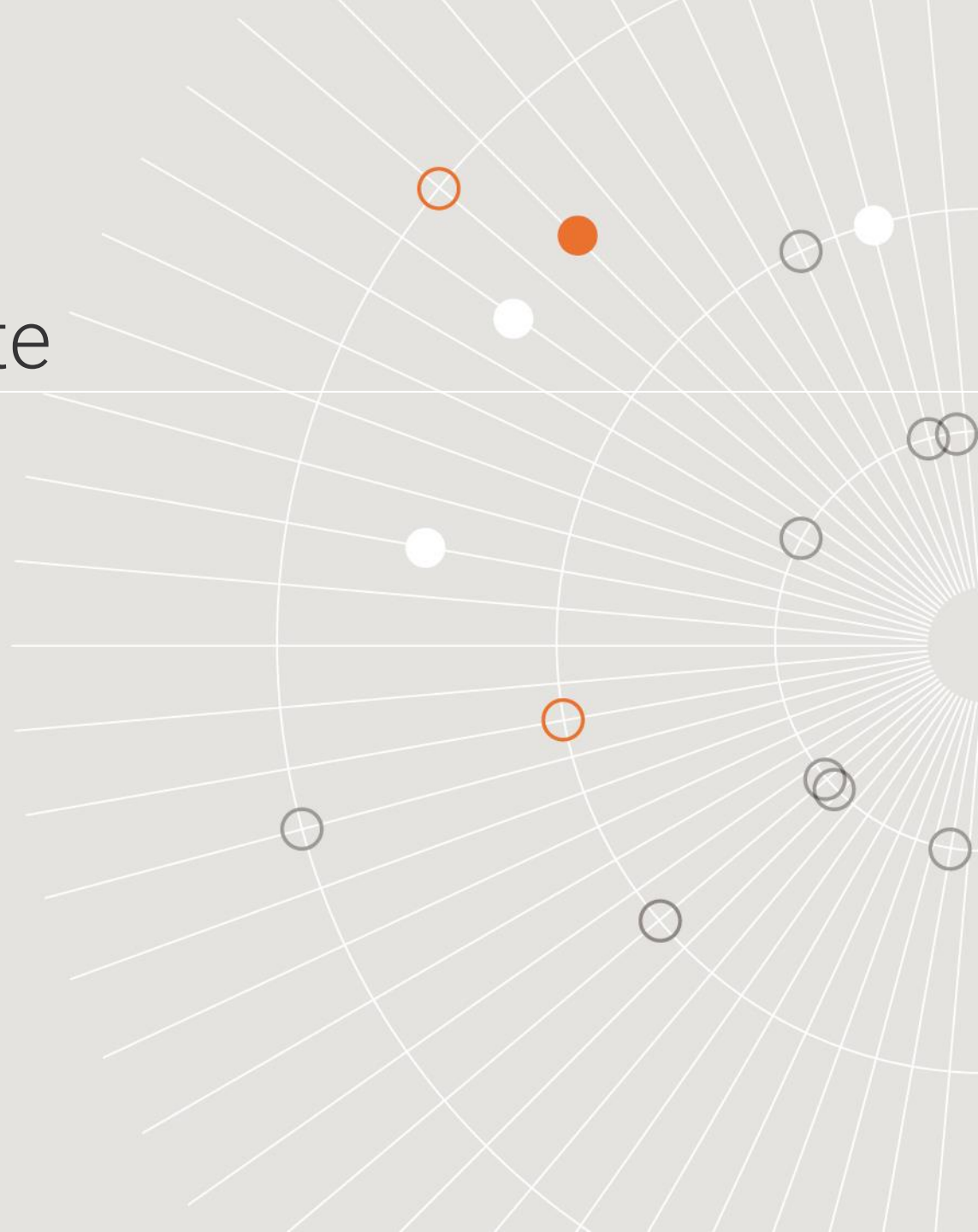
**Examples:** nomadic or displaced workers, undocumented immigrants, forced labor victims, etc.



## Why investigate Hard-to-reach populations (HRP)?




- Historically underrepresented groups (Berry & Gunn 2014)
  - Academic interest, lack of evidence
  - Concern about safety and well-being of these groups
  - Concern about risks they may pose for others
- Marginalized status, associated with stigma or fear of legal repercussions
- Improve representativeness and generalizability of findings and interventions (Raifman et al, 2022)
- Need for tailored public policy

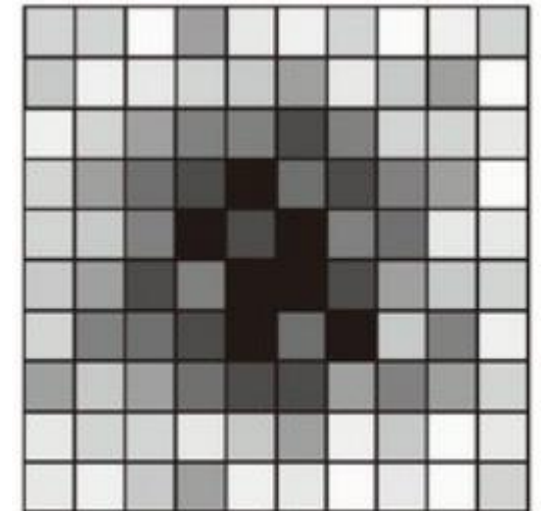
# Methodologies to investigate Hard-to-Reach Populations



## Different methods to investigate HRP

**Objective: Produce statistically sound estimates for HRP → trait prevalence, size, etc.**

- Based on especial modules in traditional surveys, use of administrative data, or (pseudo/non) probabilistic samples.
- Could also involve the combination of different sources
- Methods that aim to recover population size estimates
  - E.g., how many homeless individuals are in the city of Lima?
- Alternatively, (pseudo/non) probability methods that aim to recover “representative” estimates of size, prevalence, etc. of the HRP
  - Network scale-up 
  - Mark-recapture 
  - Respondent Driven Sampling and extensions (RDS+) 



# Network Scale-up method (NSUM)

**Intuition: Use information from known population groups to estimate the size of the hidden population**

- Ask respondents about **their personal network** in a probability sample for the general population. Module easily appended to questionnaire
- Accurate estimate for each respondent's personal network size and number of people in this network that are part of the hidden population
- Extrapolate sample information to estimate the size/prevalence of hidden population
- Reference NSUM: "How many Jorges do you know?" "How many policemen?"
  - Admin. data: we know the size of those population → estimate personal network
  - Use the personal network size to "weight" the number of individuals in hidden pop.
- Summation NSUM: direct estimates based on relation types (friends, colleagues, etc.)
- NSUM module can easily be appended to a survey questionnaire



<b>Hidden population(s)</b>	<b>Location</b>	<b>Citation</b>
Mortality in earthquake	Mexico City, Mexico	( <a href="#">Bernard et al., 1989</a> )
Rape victims	Mexico City, Mexico	( <a href="#">Bernard et al., 1991</a> )
HIV prevalence, rape, and homelessness	U.S.	( <a href="#">Killworth et al., 1998b</a> )
Heroin use	14 U.S. cities	( <a href="#">Kadushin et al., 2006</a> )
Choking incidents in children	Italy	( <a href="#">Snidero et al., 2007, 2009, 2012</a> )
Groups most at-risk for HIV/AIDS	Ukraine	( <a href="#">Paniotto et al., 2009</a> )
Heavy drug users	Curitiba, Brazil	( <a href="#">Salganik et al., 2011a</a> )
Groups most at-risk for HIV/AIDS	Kerman, Iran	( <a href="#">Shokoohi et al., 2012</a> )
Men who have sex with men	Japan	( <a href="#">Ezoe et al., 2012</a> )
Groups most at-risk for HIV/AIDS	Almaty, Kazakhstan	( <a href="#">Scutelnicuic, 2012a</a> )
Groups most at-risk for HIV/AIDS	Moldova	( <a href="#">Scutelnicuic, 2012b</a> )
Groups most at-risk for HIV/AIDS	Thailand	(Aramrattan and Kanato, 8 30)
Groups most at-risk for HIV/AIDS	Rwanda	( <a href="#">Rwanda Biomedical Center, 2012</a> )
Groups most at-risk for HIV/AIDS	Chongqing, China	( <a href="#">Guo et al., 2013</a> )
Groups most at-risk for HIV/AIDS	Tabriz, Iran	( <a href="#">Khounigh et al., 2014</a> )
Men who have sex with men	Taiyuan, China	( <a href="#">Jing et al., 2014</a> )
Drug and alcohol users	Kerman, Iran	( <a href="#">Sheikhzadeh et al., 2014</a> )
Men who have sex with men	Shanghai, China	( <a href="#">Wang et al., 2015</a> )



# Mark-recapture or Multiple System Estimation (MSE)

**Intuition: Combine different lists of individuals to identify overlapping members and estimate size of hidden pop.**

- Select multiple samples (capture occasions). Read across arising lists to measure overlap and **estimate abundance** (population size)
- For human subjects → mark means collecting PI data
- Modifications: open/closed populations, lists heterogeneity, etc.

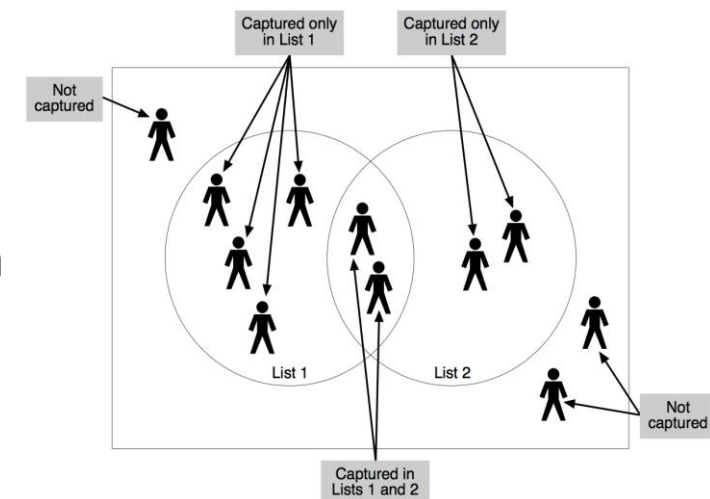
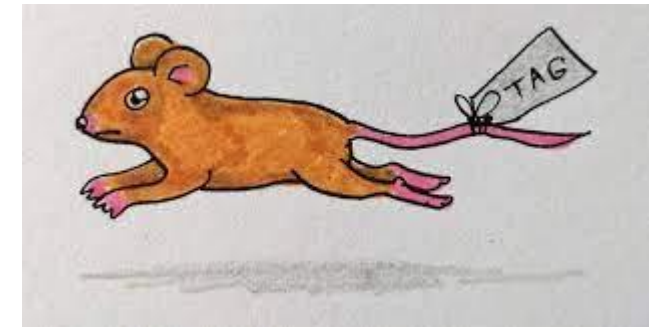
## Datasets:

- Primary: lists of individuals from different sampling occasions strategies (site sampling, respondent-driven sampling, or a household sampling design)

MSE is a generalization or mark-recapture: overlap individuals between lists

- Allows for secondary data: NGOs, law enforcement records, local information (village leaders), etc.

In Perú, Truth and Reconciliation Commission: **~25,000 deaths** documented.  
MSE analysis, HRDAG **~70,000 deaths** estimated



# Respondent Driven Sampling (RDS+)

---



# Respondent Driven Sampling (RDS+)

**Intuition: start with convenience sample and use incentives so respondents redirect us to other members of target pop.**

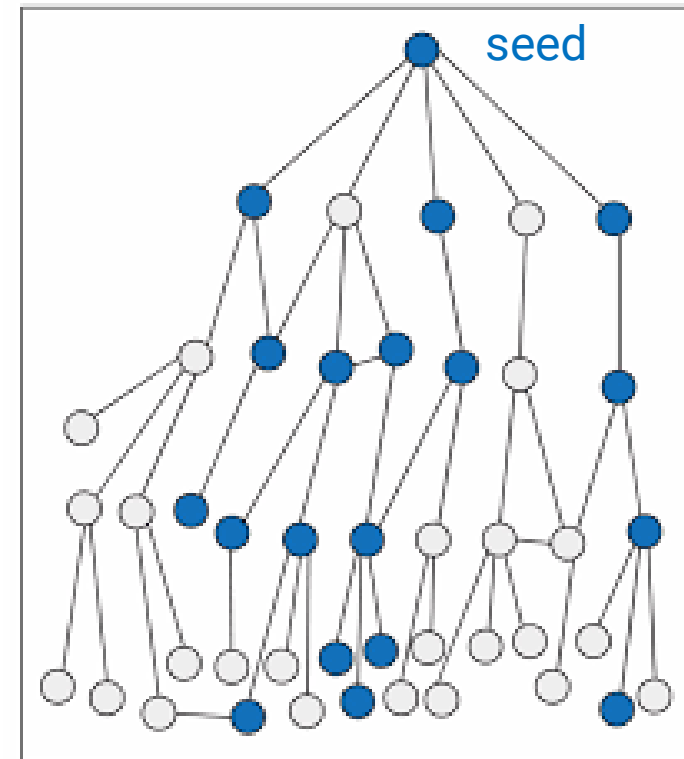
- Strategy for accessing, recruiting, and studying a hidden/hard-to-reach population through pseudo-probability sampling

## Process

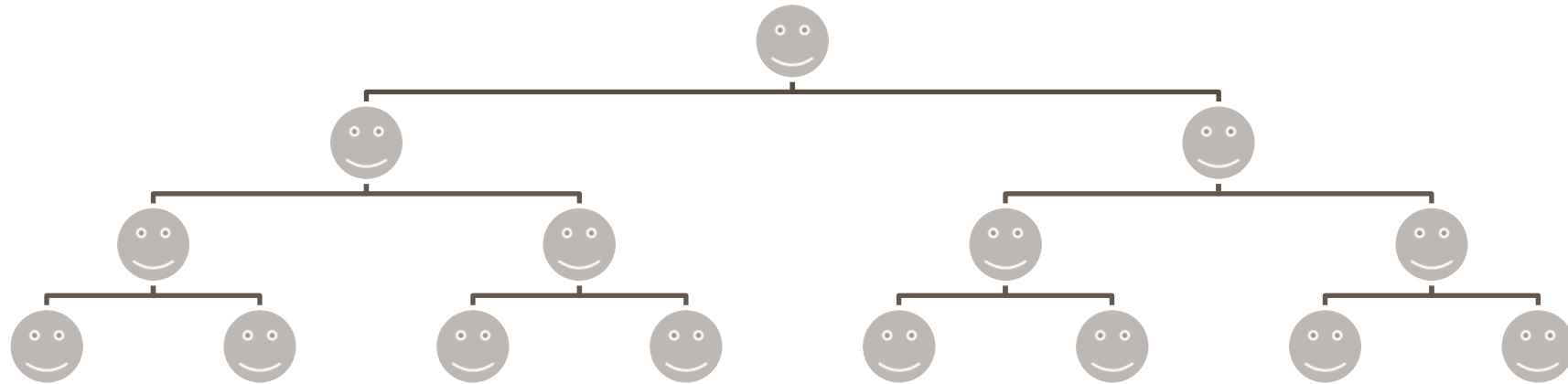
- Start with convenience sample of individuals in the target population (**seeds**)
- Individuals refer to others in the target population (**branching**). A subset of these **nominees** are selected **randomly**
- These process occurs several times in subsequent **waves**
- Usually involve monetary (double) incentives for participation

## Estimation

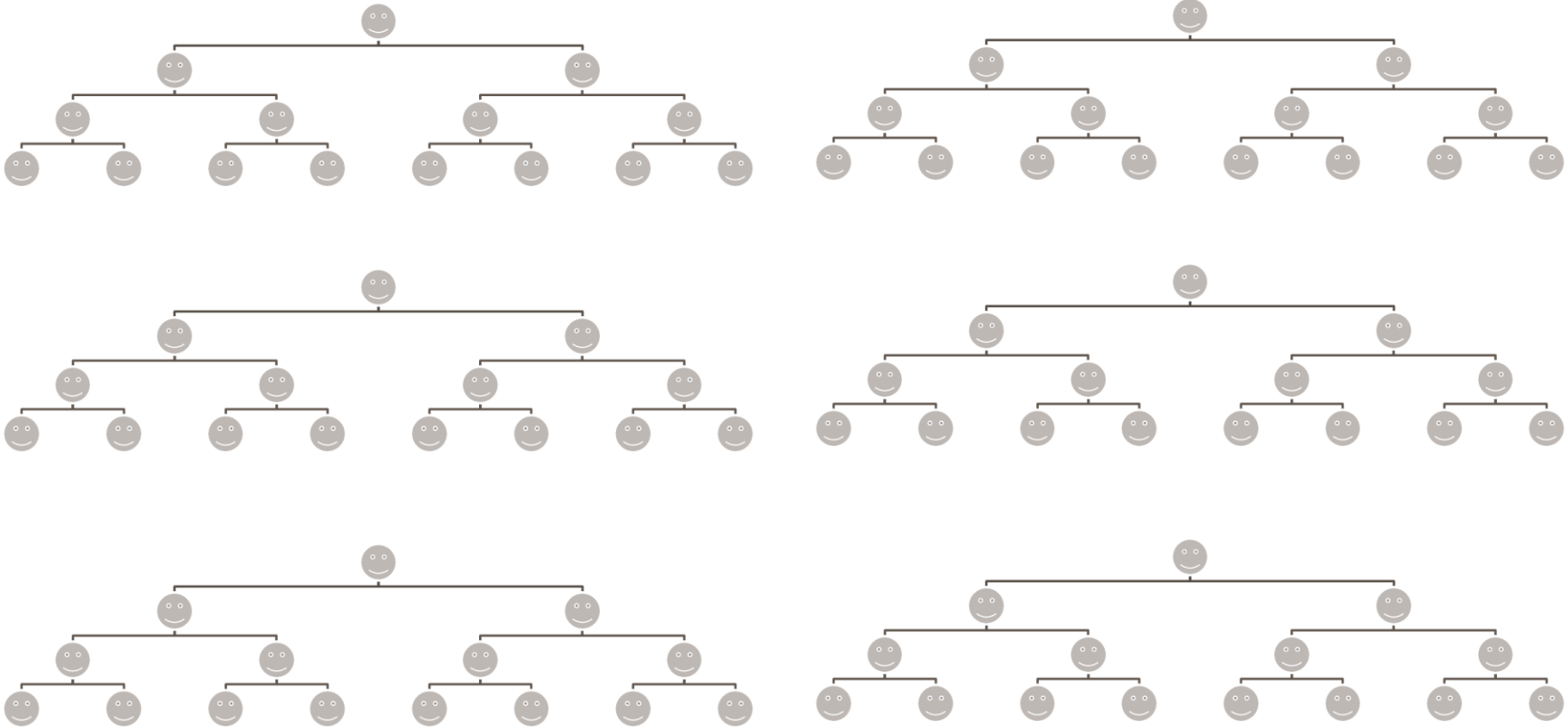
- Approximate sampling weights to use as a “traditional sample” (Markov chain)
- Several estimators: Salganick-Heckathorn, Voltz-Heckathorn, Gile-Hancock resampling estimator, NE4NS (Thompson & Vincent)



# Research Methodology: Respondent Driven Sampling



# Research Methodology: Respondent Driven Sampling



# Assumptions

---

Estimation relies heavily on Markov chain theory:

- A random walk is taken over the population network.
- Start with a pre-chosen node and move to a neighbor with uniform probability. Repeat.
- With a large enough sample size, inclusion probabilities are proportional to network size.

## Assumptions

- The entire study population is comprised of a single network, i.e., a link exists between any two individuals.
- Links in the population network are non-directional, i.e., symmetric.
- Sampling occurs with replacement.
- Respondents can accurately report their network size. Recruitment occurs at random from personal networks

# Network Data Sources

Data to map overlaps comes from two sources:

1. Unique codes on referral coupons issued to respondents
  - QR codes (unique for each respondent-referral)
  - Activated in SurveyCTO live database
  - Protection against deceive (re-use, photocopy, cutting edges, etc)
2. Covariates captured at the end of the survey on potential nominations
  - Only a subset of nominated individuals are randomly selected
  - However, we can still use information on the non-selected

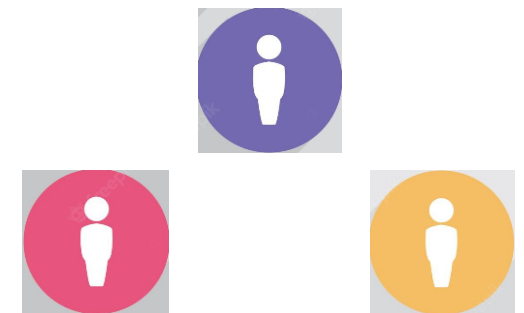


Respondent

Nominees



Selected referrals



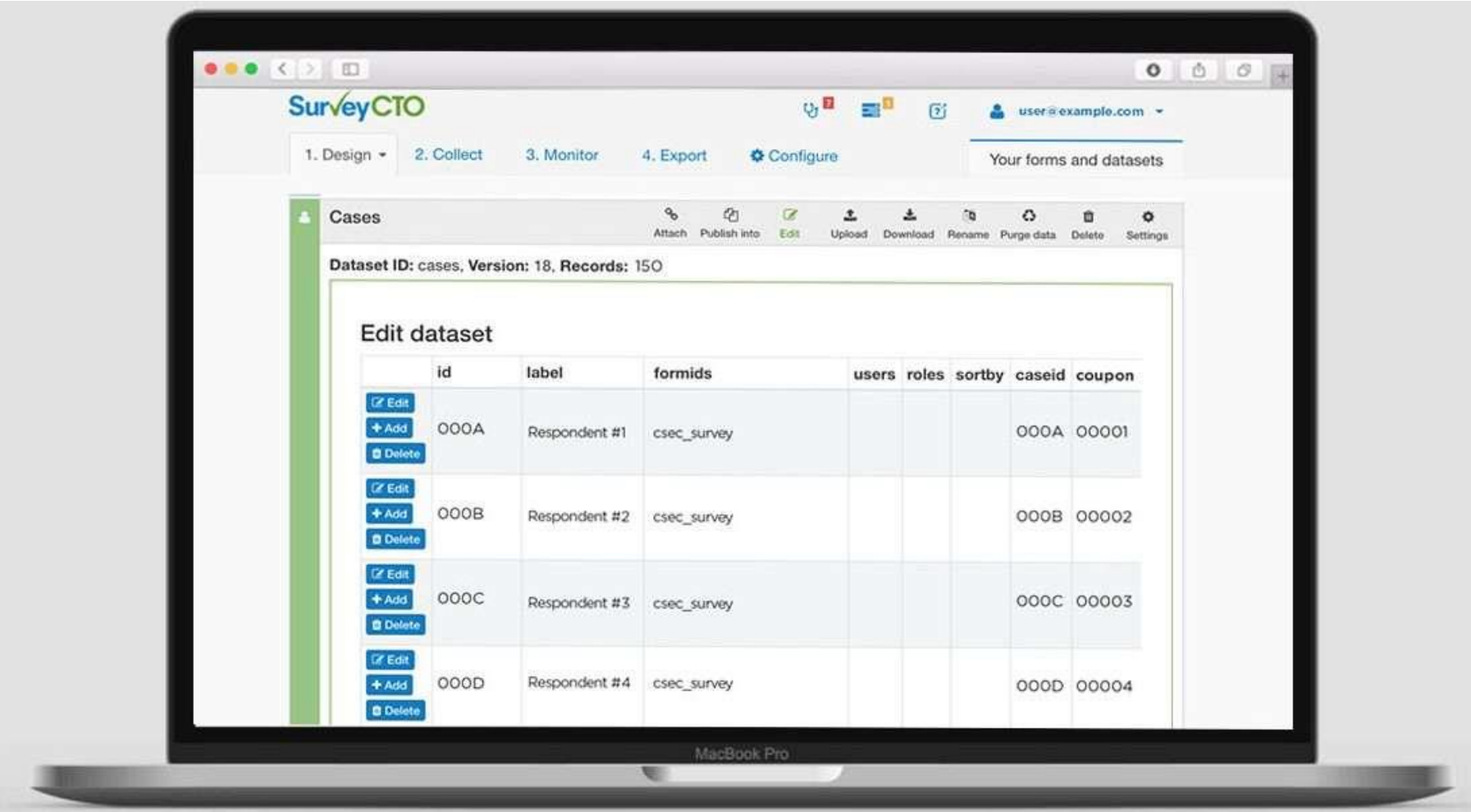
# Data collection software

Details on how NORC uses SurveyCTO for RDS

[www.surveycto.com/case-studies/norc-respondent-driven-sampling/](http://www.surveycto.com/case-studies/norc-respondent-driven-sampling/)



SCAN ME





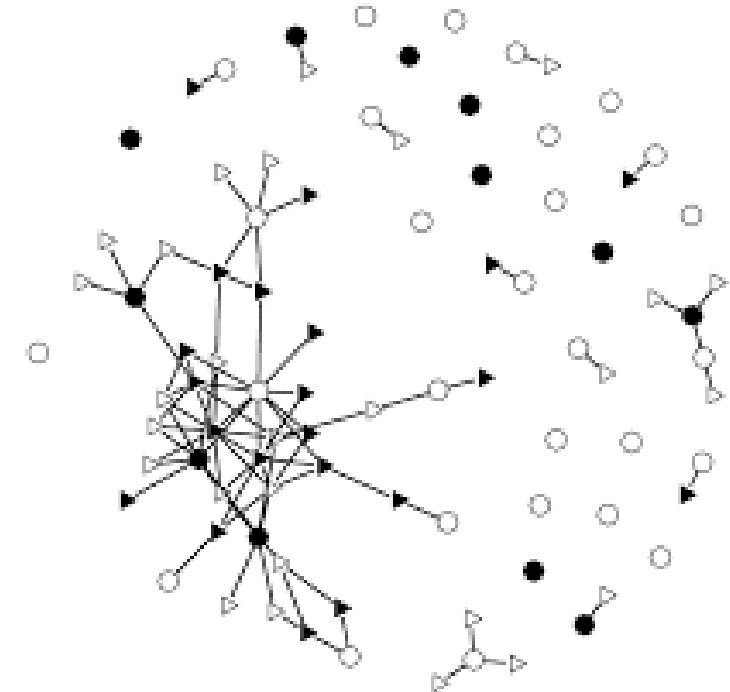
## What about the “+” in RDS+?

Sometimes called **Link-tracing sampling**. Exploit additional PI to better capture the network

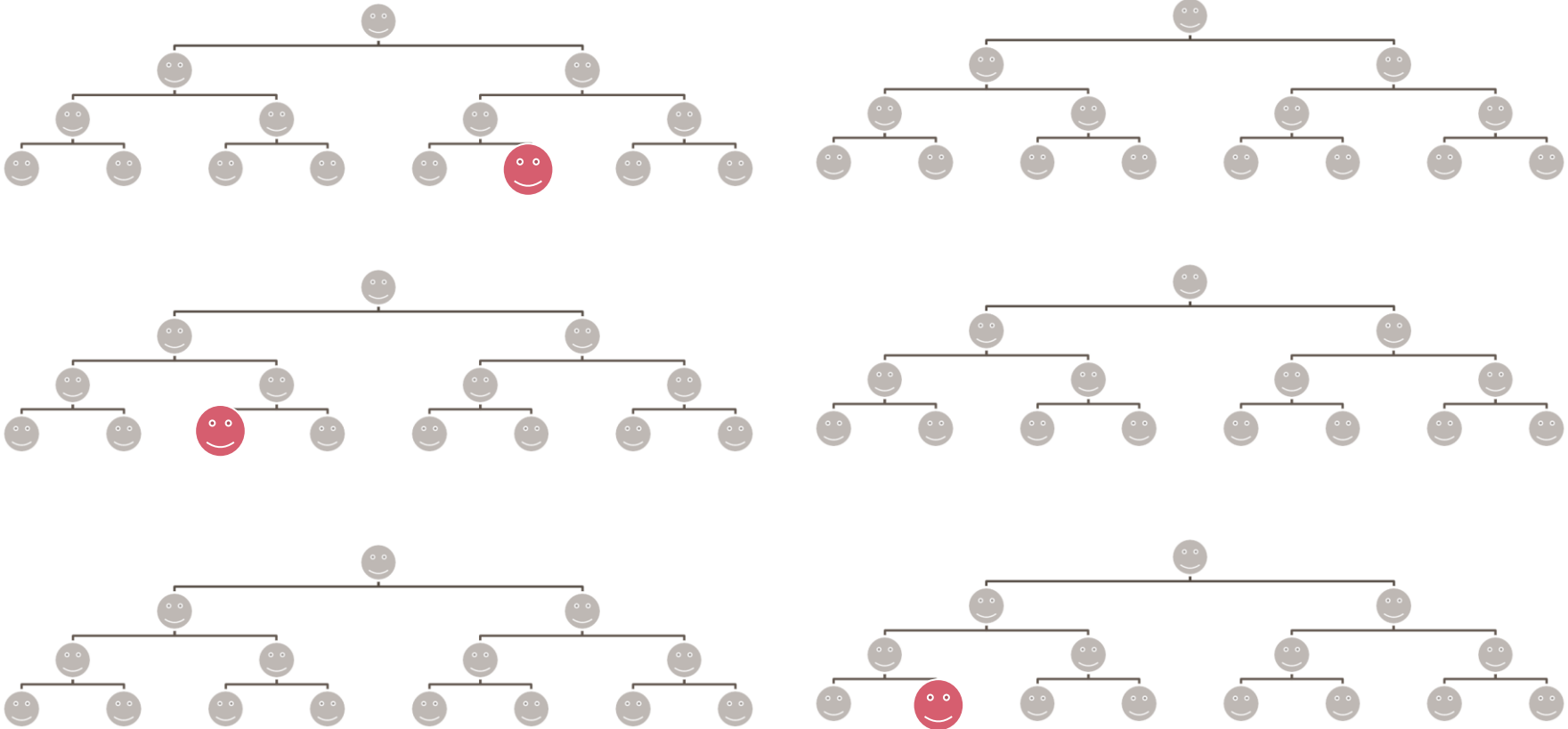
- Use PI information of **non-selected nominees**: post data linkage to observe all personal connections disclosed during the interview
- Allow re-interviews (**recaptures**) if individuals are nominated through different branches

### Advantages

- Identify overlaps in the sample networks
- Recaptures: combine RDS with MSE strategies to estimate population size
- A seminal estimator of pop. size is proposed by Frank and Snijders (1994)
- Estimation considers the quantity of links within the initial sample against the quantity of links that stretch out of the initial sample.
- SS-PSE method (Handcock et al., 2015) utilizes a Bayesian approach and a successive sampling strategy to estimate population size



# Research Methodology: Respondent Driven Sampling



## Volz-Heckathorn (2008) estimator

- Improved afterward by dual-model estimator
- The recruitment process controls for differential bias and effectiveness

### Assumptions

- i. Initial sample members are chosen independently and proportional to network degree. Weights are the inverse of the degree and missing for seeds
- ii. Relationships within the population are symmetric (i.e., if A is a contact of B, then B is also a contact of A)
- iii. Participants recruit uniformly at random from their contacts. No need to be a single recruit compared to SH
- iv. Recruited individuals always participate in the study
- v. Individuals can be recruited into the sample more than once
- vi. Number recruits does not depend on individual traits
- vii. Respondents accurately report their social network degree.

$$\hat{p} = \frac{\sum_{j \in I} 1/d_j}{\sum_{j \in S} 1/d_j}$$

where  $\hat{p}$  is the estimated proportion,  $S$  is the full sample,  $I$  is sample members with trait  $y_i$  and  $d_j$  is the self-reported 'degree' of respondent  $j$

Gile and Handcock (2010) show several weaknesses of the VH estimator due to unrealistic assumptions.

## Thompson (2020)

- Recent estimator → exploits topology of the sampled network and survey design.
- Results: ↓↓ MSE, ↓ Bias  $\approx 0$
- Inclusion probability not only depends on the degree (VH) but also on the position in topology
- Uses topology & sampling design, relax some assumptions
  - Without replacement
  - Limited branching by number of coupons
  - Resample from the sampled network. Count the number of times each node appear (frequencies) to compute the  $\text{Pr}(\textit{selection})$  in the real network
- What is a resample?
  - Remove nodes randomly at each step
  - Reseed at each step
  - Obtain a smaller network

a) **Assumption:** Matching resampling method to real sample, the resample probability of inclusion  $\phi_i$  are proportional to the probability of inclusion in the real unobserved population

$$\pi_i \times \phi_i = c\pi_i$$

a) **Assumption:** inclusion frequencies  $f_i$  tend in probability to  $\phi_i$  for many resamples  $T$

$$f_i \xrightarrow{p, T \rightarrow \infty} \phi_i$$

c) If these assumptions hold, then

$$\widehat{\mu}_f \equiv \frac{\sum(y_i/f_i)}{\sum 1/f_i} \xrightarrow{p} \widehat{\mu}_f \equiv \frac{\sum(y_i/\phi_i)}{\sum 1/\phi_i}$$

d) And given  $\phi_i = c\pi_i$

$$\widehat{\mu}_f \xrightarrow{p} \widehat{\mu}_\pi \equiv \frac{\sum(y_i/\pi_i)}{\sum 1/\pi_i}$$

Some examples of RDS applied to HRP



Title	Authors	Publication Year	Population Studied
Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations	Heckathorn, D.	1997	Drug injectors, USA
Estimating the Size of Hidden Populations Using Respondent-Driven Sampling Data: Case Examples from Morocco	Johnston, Lisa G., et al.	2006	Drugs injectors, sex workers, men who have sex with other men, and migrants in Morocco.
Sampling and estimation in hidden populations using respondent-driven sampling.	Salganik, Matthew J., et al.	2004	Jazz musicians in New York and San Francisco
Model-based Respondent-driven sampling analysis for HIV prevalence in brazilian MSM	Robineau, O., et al	2020	HIV prevalence in Brazil
Searching for sex trafficking victims: Using a novel link-tracing method among commercial sex workers in Muzaffarpur, India.	Vicent, Zhang & Dank	2019	Sex workers, India

# Sampling in Practice

# How we sample our respondents at NORC studies

- Complementarily to contact information from NGOs, public sector, etc. our sampling methodology starts with a time/location sampling
- We select locations where individuals from the target population tend to congregate, establishing times where we will visit.
- At these locations we will interview “seeds”
- Each successful seed interview will nominate up to 7 individuals. From these, the software will randomly select up to 3 (referrals).
- The enumerators gives coupons assigned to each referral and activates them in the database. These follow-up surveys are called “waves”.
- The local firms will screen “wave” interviews and organize interviews based on the teams’ location until the desired sample size is reached.



# Explaining Referrals to respondents: Example of mining project

At the end of the interview, you must clearly explain the referral procedures and eligibility criteria.

They will be given three coupons to distribute to up to three contacts that meet the eligibility criteria:

- The referral is at least 18 years of age
- The referral has worked in gold mining in the last 12 months (approx. January/February 2023)
- The referral lives in the state of Pará

# Referral's interviews

- Interviews for individuals that are not seeds will be established based on the coupons from previous respondents.
- 7 nominees, 3 randomly selected referrals by the SurveyCTO software



NORC Projects using RDS+

---

Forced Labor in mining sector (Brazil)



## Tapajos Project:

- Tapajos river basin in Para, Brazil. High concentration of (informal) gold mining activity in the Amazonian rainforest
- A data-driven and evidence-based **intervention** addressing prevention, protection, and prosecution in terms of labor exploitation
- Use **context-specific research** to inform intervention design, targeting, implementation, and potential for replication and scaling up

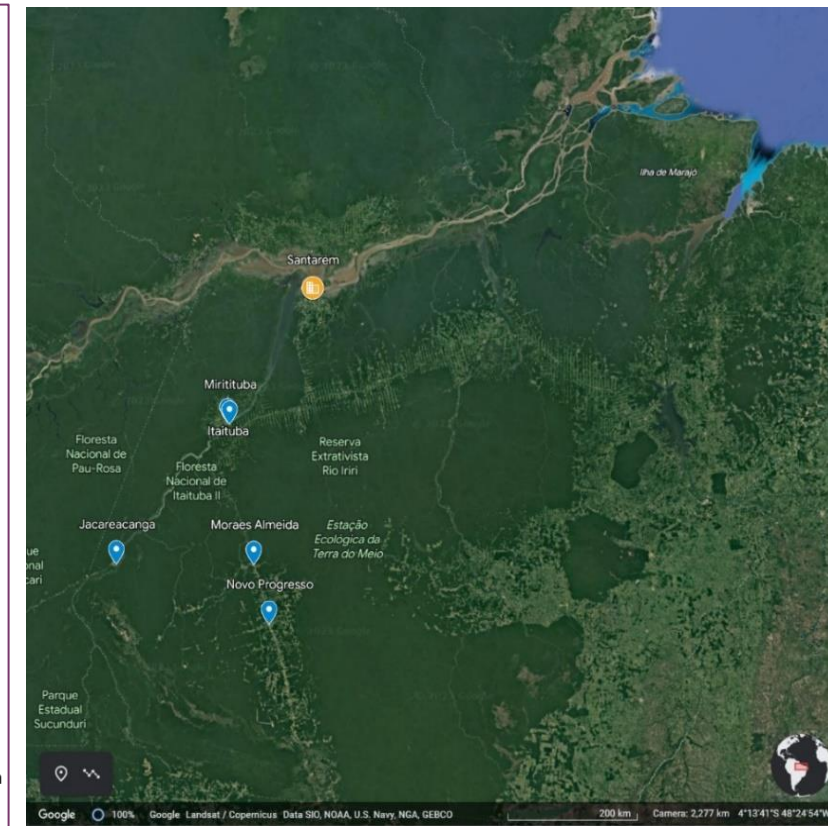
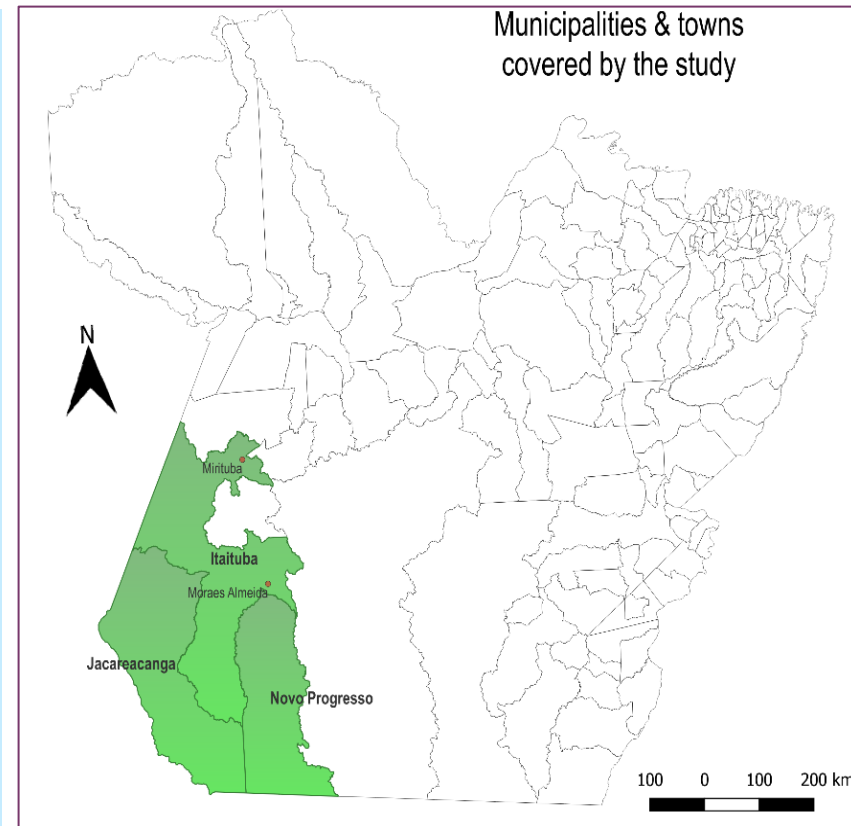
## Phase 1 comprised of 3 components:

- (1) Formative Assessment (FGDs, KIIs)
- (2) Intervention Development Research (IDR)
- (3) Forced Labor Prevalence Estimation (quantitative survey)

Tapajos is called “The Amazonian El Dorado”



- RDS+ sampling methodology
- 3 sampled municipalities within the Tapajós River Basin: Itaituba, Jacareacanga, and Miritituba



## Survey Instrument (Measurement Domains)

### *Individual Background:*

1. Demographic information
2. Work history and conditions

### *Substantive to forced labor dimensions:*

1. Rights violations
2. Vulnerability and occupational hazards
3. Service priorities and protection needs

## Defining Forced Labor

- Indicators were derived from The University of Georgia's Prevalence Reduction Innovation Forum (PRIF) measurement, which have been used in 6 countries, and currently endorsed by the funding agency (JTIP)
- Potential victims are those who have experienced two or more abuses from two separate categories

Threshold 1	Threshold 2
	Abuses during recruitment
Losing freedom of movement through surveillance, isolation, or being locked in the workplace, or losing the freedom to communicate with friends or family	Abusive employment practices and penalties
	Abuses in personal life and properties
Having to perform sex acts to pay off debts	Degrading work conditions
	Debt bondage or dependency
	Violence and threats of violence

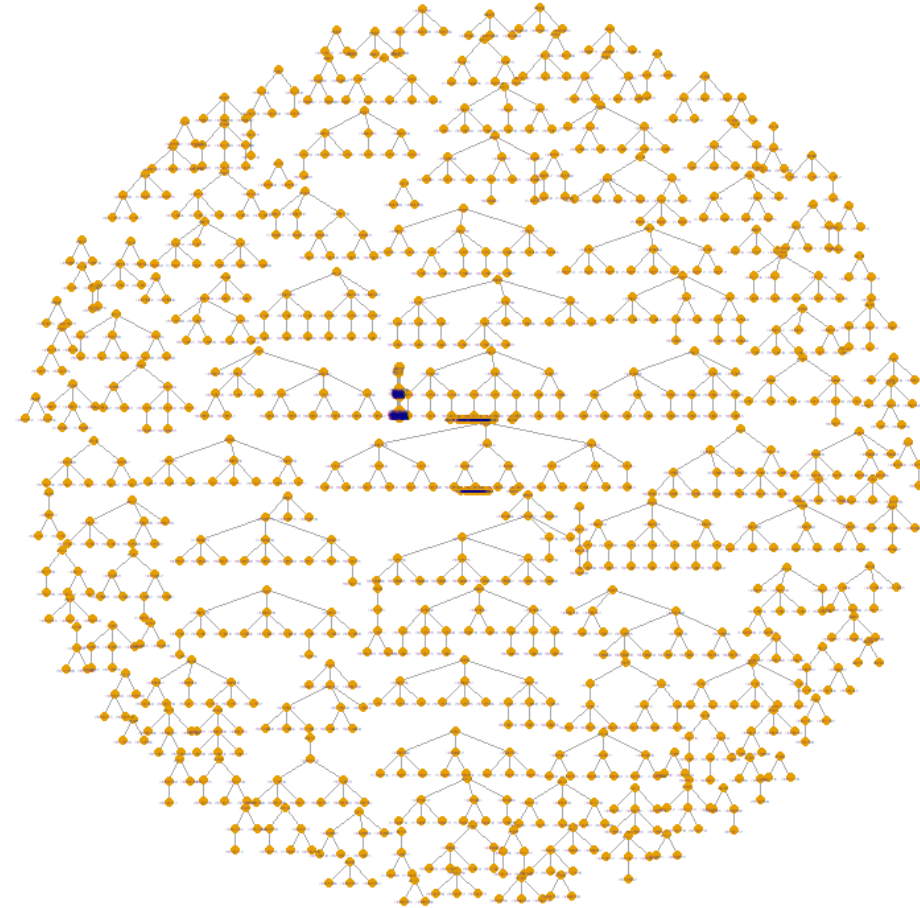
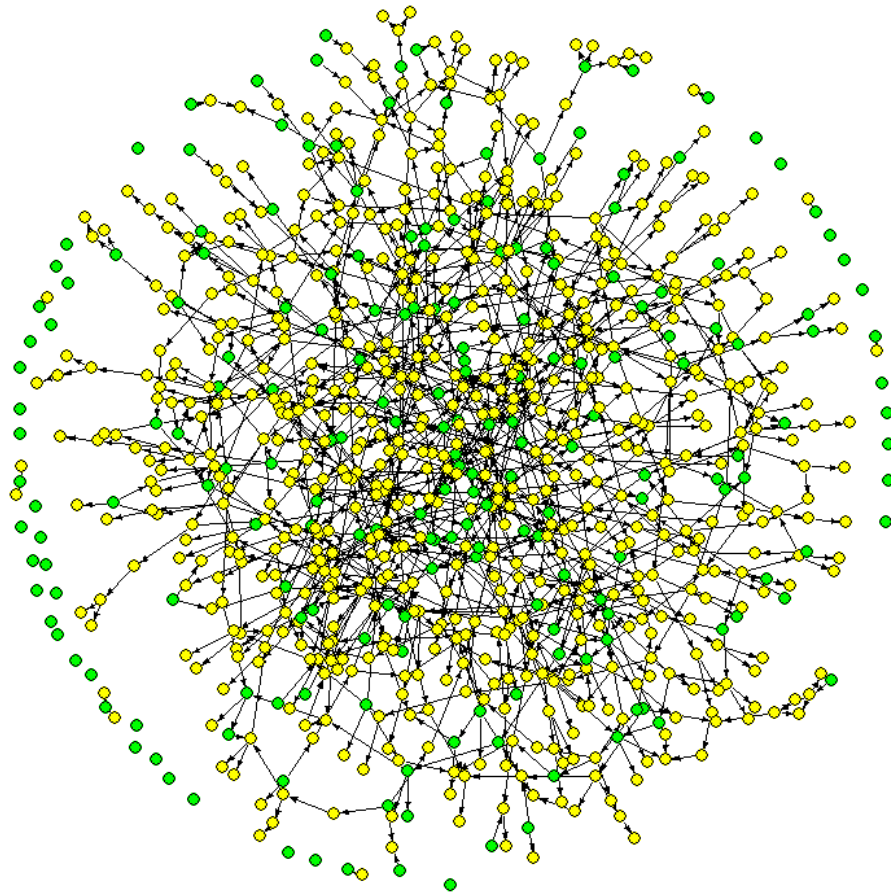


---

# Findings

## Prevalence of Forced Labor in Gold Mining

Seeds in Green



- Final Sample: 863 (183 seed interviews; 680 referrals)
- Miner population estimate: **4 550** with CI 95% of [2 853; 6 248]

## Sample Profile

---

Demographics 91% males

---

Avg. age: 43 years

---

Most miners started working from age 20

---

67% multiracial

---

79% literate

---

Work  
Schedule 12.5 hours/day

---

6.5 days/week

---

## Prevalence Estimation for Target Population

### Overall

- **40% target population** are potential victims of Trafficking in Persons for forced labor



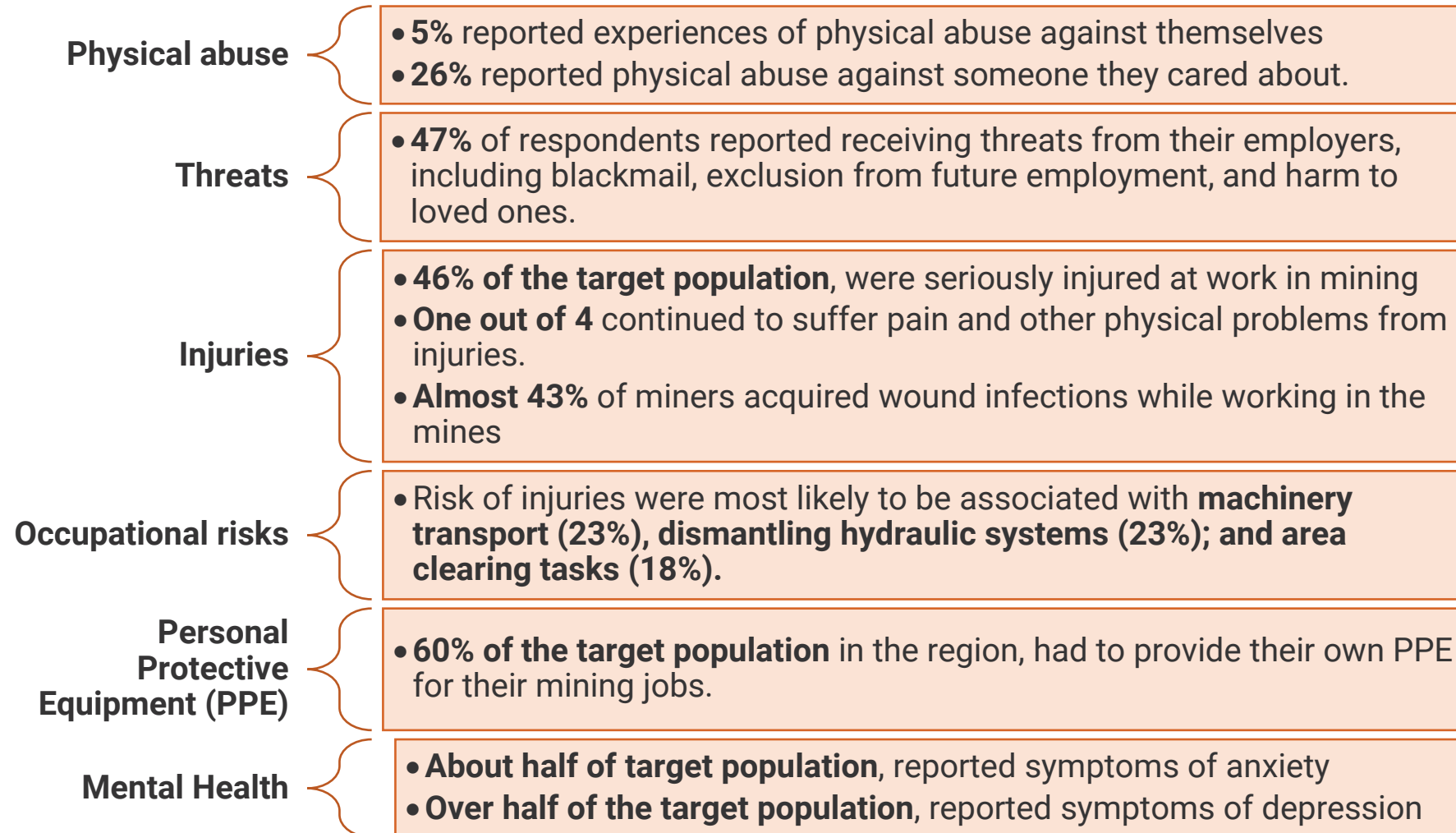
**Deceptive Recruitment: 57% of target population** reported their current job duties differed from what they had been told



**40% of target population** reported being charged inflated prices for goods purchased from employers



About **23% of target population** reported being surveilled



Different types of **Hard-to-survey** populations. Hard-to-reach is one of them: difficult to contact, identify, or have barriers to access

RDS approximates a **probabilistic sampling** using chain of referrals and introducing randomization. Several estimators available.

For the Tapajos project, we used an RDS+ design to study forced labor in the Brazilian mining sector. We found **40% of miners** are potential victims of Trafficking in Persons



**Complementary techniques** develop to study population size and prevalence of characteristic in HRP: NSUM, MSE, RDS

**Extensions** (RDS+) allow recaptures (MSE) and exploits PI information of non-respondent to better map the network

RDS designs can be applied to **other HRP populations**: migrants, sex workers, MSM, jazz musicians, etc.

---

Questions?



# Thank you.

**Angelo Cozzubo**

Data Scientist I

cozzubo-angelo@norc.org



@acozzubo

SCAN ME



---

 Research You Can Trust™

---

 **NORC LABS**



---

# Appendix