Estimating the size of hidden populations using the network scale-up method: Evidence from Brazil and Rwanda

Matthew J. Salganik
Department of Sociology, Princeton University

Joint with:
(Dennis Feehan)
(Ale Abdo, Maeve Mello, Dimitri Fazito, Neilane Bertoni, & Chico Bastos)
(Mary Mahy, Wolfgang Hladik, & Aline Umubyeyi)

April 27, 2015
Sociology 504, Princeton University

This research is supported by the NIH (R01-HD075666), NSF(CNS-0905086), UNAIDS, and USAID.

There are between 33 million and 37 million people worldwide living with HIV/AIDS. In most countries, the disease is concentrated in three high-risk groups:

- injection drug users
- commercial sex workers
- men who have sex with men

Better information about these group can be used to understand and control the spread of HIV: “know your epidemic”

Questions:

- What percent of drug injectors in New York have HIV?
- How many drug injectors are there in New York?

Methods:

- respondent-driven sampling
- network scale-up method

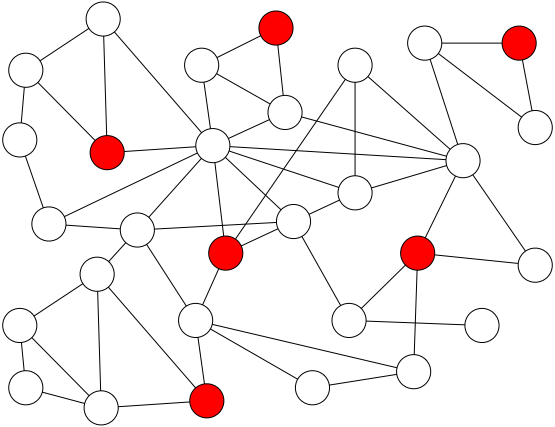
Modeling
counting with multiplicity

$\leftrightarrow$

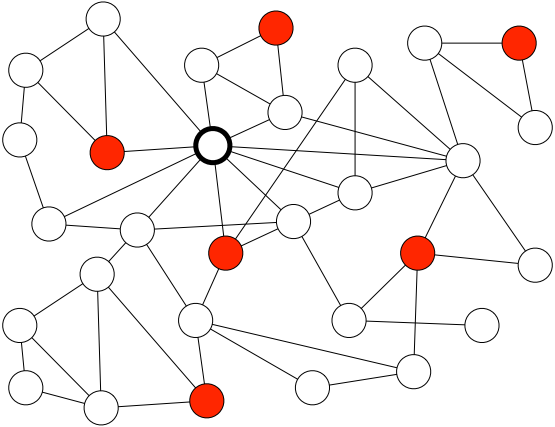
Empirical
Brazil
Rwanda

Basic insight from Bernard et al. (1989)

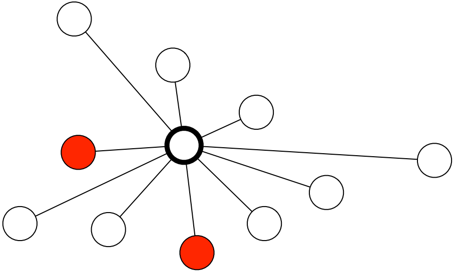
Network scale-up method



Network scale-up method



Network scale-up method



$$\hat{N}_H = \frac{2}{10} \times 30 = 6$$

If $\underbrace{y_{i,k} \sim Bin(d_i, N_k/N)}_{\text{basic scale-up model}}$, then maximum likelihood estimator is

$$\hat{N}_H = \frac{\sum_i y_{i,H}}{\sum_i \hat{d}_i} \times N$$

- $\hat{N}_H$: number of people in the hidden population
- $y_{i,H}$: number of people in hidden population known by person i
- $\hat{d}_i$: estimated number of people known by person i
- N: number of people in the population

See Killworth et al., (1998)

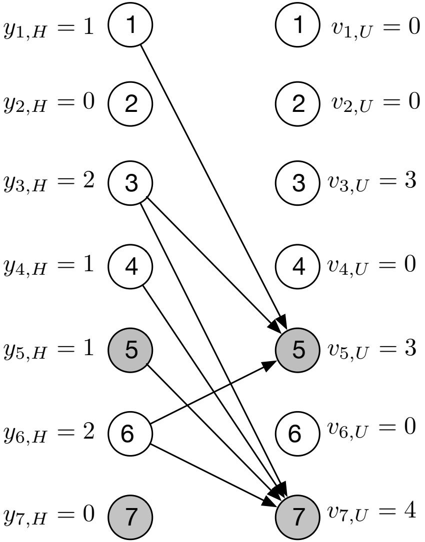
- Requires a random sample from the entire population
- Respondents are asked:
    - How many people do you know who are drug injectors?
    - How many women do you know that have given birth in the last 12 months?
    - How many people do you know who are middle school teachers?
    - ...
    - How many people do you know named Michael?
- "Know" typically defined: you know them and they know you and have you been in contact with them over the past two years

If $\underbrace{y_{i,k} \sim Bin(d_i, N_k/N)}_{\text{basic scale-up model}}$, then maximum likelihood estimator is

$$\hat{N}_H = \frac{\sum_i y_{i,H}}{\sum_i \hat{d}_i} \times N$$

- $\hat{N}_H$: number of people in the hidden population
- $y_{i,H}$: number of people in hidden population known by person i
- $\hat{d}_i$: estimated number of people known by person i
- N: number of people in the population

See Killworth et al., (1998)


$y_{1,H} = 1$
$y_{2,H} = 0$
$y_{3,H} = 2$
$y_{4,H} = 1$
$y_{5,H} = 1$
$y_{6,H} = 2$
$y_{7,H} = 0$
1
2
3
4
5
6
7
1
2
3
4
5
6
7
$v_{1,U} = 0$
$v_{2,U} = 0$
$v_{3,U} = 3$
$v_{4,U} = 0$
$v_{5,U} = 3$
$v_{6,U} = 0$
$v_{7,U} = 4$

total out-reports $=$ total in-reports

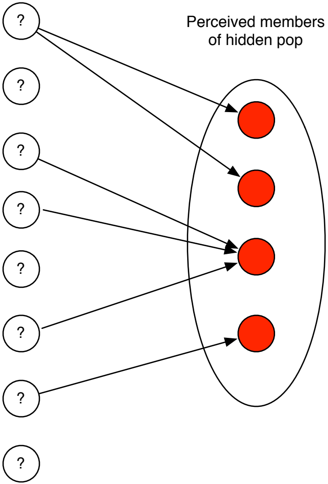
$$\text{total out-reports} = \text{total in-reports}$$

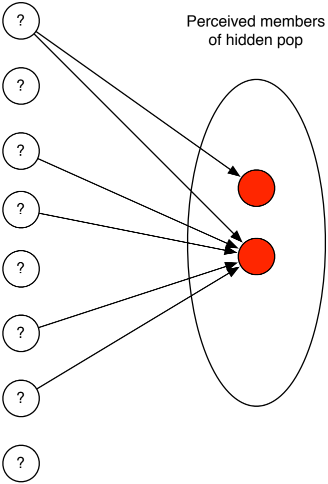
$$\text{total out-reports} = \text{size of hidden pop} \times \text{in-reports per member of hidden pop}$$

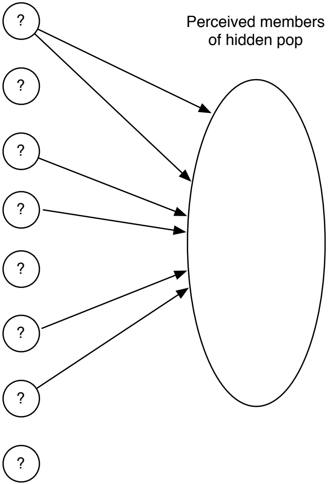
$$\text{total out-reports} = \text{total in-reports}$$

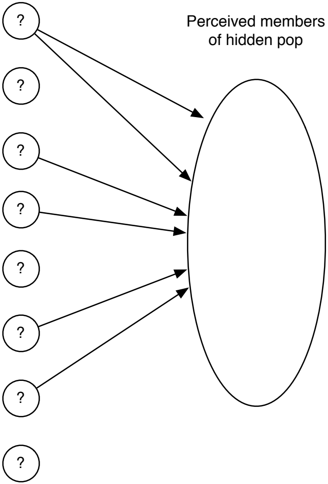
$$\text{total out-reports} = \text{size of hidden pop} \times \text{in-reports per member of hidden pop}$$

$$\text{size of hidden pop} = \frac{\text{total out-reports}}{\text{in-reports per member of hidden pop}}$$


?
?
?
?
?
?
?
?
Perceived members
of hidden pop


?
?
?
?
?
?
?
?
Perceived members
of hidden pop


?
?
?
?
?
?
?
?
Perceived members of hidden pop


?
?
?
?
?
?
?
?
Perceived members of hidden pop

Counting with multiplicity approach:

- no assumptions about the underlying social network
- extends naturally to incomplete social awareness
- extends naturally to incomplete frames
- extends naturally to complex sample designs

Counting with multiplicity approach:

- no assumptions about the underlying social network
- extends naturally to incomplete social awareness
- extends naturally to incomplete frames
- extends naturally to complex sample designs

Motivates empirical work to

- estimate the visible degree of the target population
- optimize definition of a network tie

- Target population: Heavy drug users, people who had used illegal drugs other than marijuana more than 25 times in the past 6 months
- Location: Curitiba, Brazil (1.8 million people)
- Funded by UNAIDS and Brazilian Ministry of Health

Map source: Wikipedia

Interviewer shuffles a deck of 24 playing cards

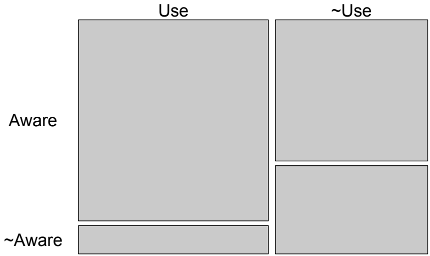
A card is pulled from the deck and the respondent is asked:

How many people do you know named [Amadeu]?

The respondent will pick up this many blocks and place them:

Record answers; clear board; repeated for 24 names.

294 participants told us about 4,173 alters



Evidence of:

- selective exposure
- selective disclosure

Many data quality checks in our paper (Salganik et al., 2011)

A useful decomposition:

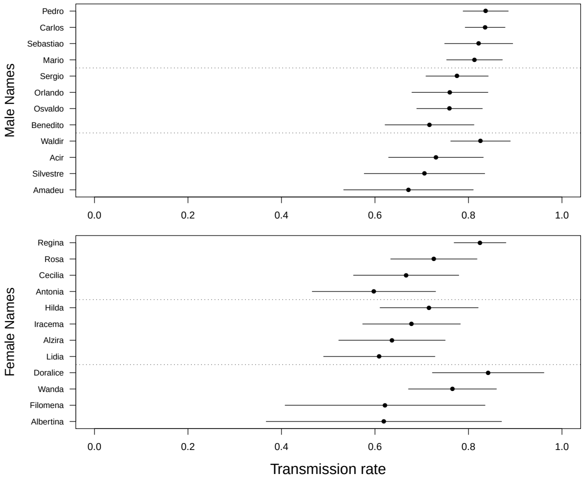
$$\underbrace{\bar{v}_{H,F}}_{\text{average visibile degree of hidden pop}} = \bar{d}_{F,F} \times \underbrace{\frac{\bar{d}_{H,F}}{\bar{d}_{F,F}}}_{\text{degree ratio } (\delta)} \times \underbrace{\frac{\bar{v}_{H,F}}{\bar{d}_{H,F}}}_{\text{true positive rate } (\tau)}$$

True positive rate

$$\hat{\tau} = \frac{\sum y_{ik[aware]}}{\sum y_{ik}} = 0.77 \quad [0.73, 0.83]$$

True positive rate

$$\hat{\tau} = \frac{\sum y_{ik[aware]}}{\sum y_{ik}} = 0.77 \quad [0.73, 0.83]$$

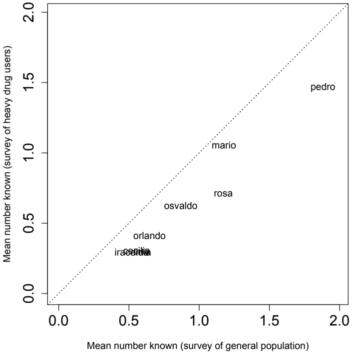


Degree ratio

$$\hat{\delta} = \frac{\sum_{i \in s_H} y_{i,k} / n_H}{\sum_{i \in s_F} y_{i,k} / n_F} = 0.69 \quad [0.60, 0.79]$$

Degree ratio

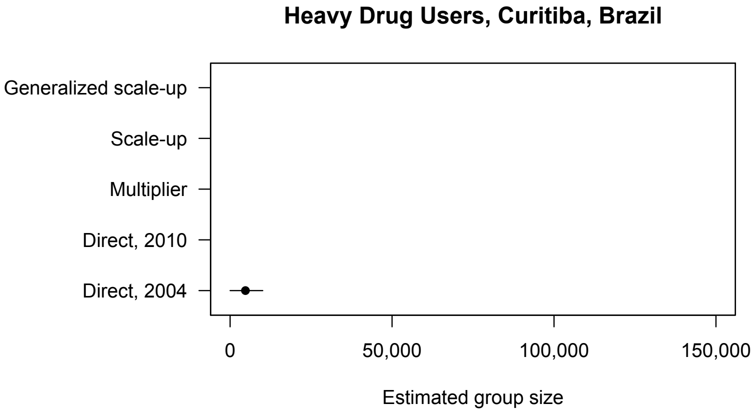
$$\hat{\delta} = \frac{\sum_{i \in s_H} y_{i,k} / n_H}{\sum_{i \in s_F} y_{i,k} / n_F} = 0.69 \quad [0.60, 0.79]$$

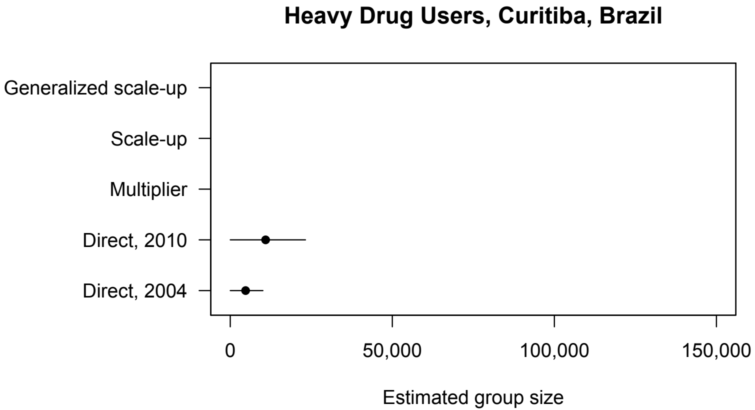


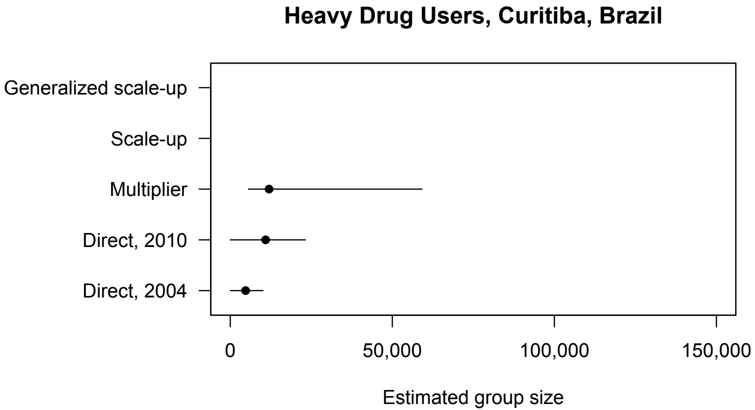
$$\hat{\bar{v}}_{H,F} = \hat{\bar{d}}_{F,F} \times \underbrace{\widehat{\frac{\bar{d}_{H,F}}{\bar{d}_{F,F}}}}_{\text{degree ratio } (\delta)} \times \underbrace{\widehat{\frac{\bar{v}_{H,F}}{\bar{d}_{H,F}}}}_{\text{true positive rate } (\tau)}$$

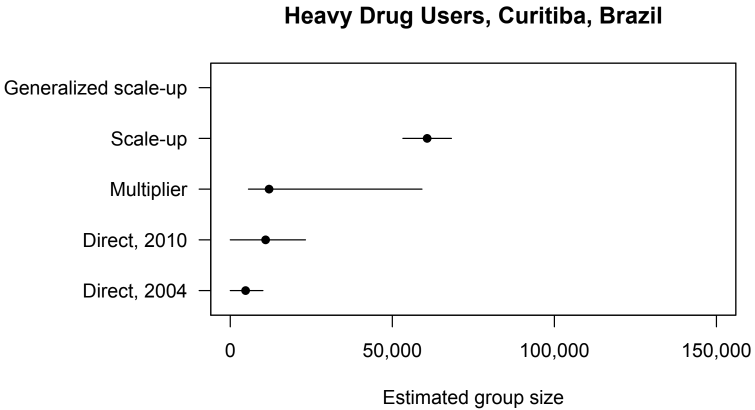
$$\hat{\bar{v}}_{H,F} = 184 \times 0.69 \times 0.77 \approx 100$$

Average visible degree of the hidden population is very different from the average degree of the population

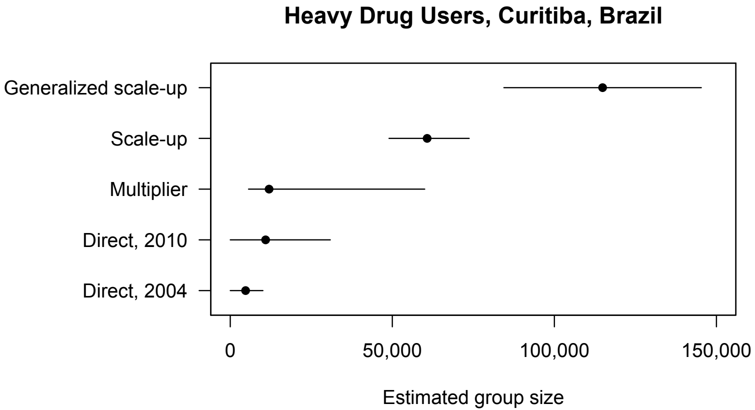

Heavy Drug Users, Curitiba, Brazil
Generalized scale-up
Scale-up
Multiplier
Direct, 2010
Direct, 2004
0
50,000
100,000
150,000
Estimated group size


Heavy Drug Users, Curitiba, Brazil
Generalized scale-up
Scale-up
Multiplier
Direct, 2010
Direct, 2004
0
50,000
100,000
150,000
Estimated group size


Heavy Drug Users, Curitiba, Brazil
Generalized scale-up
Scale-up
Multiplier
Direct, 2010
Direct, 2004
0
50,000
100,000
150,000
Estimated group size



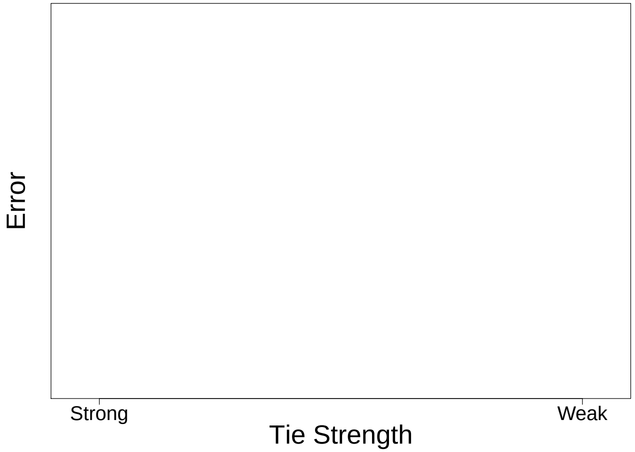
Heavy Drug Users, Curitiba, Brazil
Generalized scale-up
Scale-up
Multiplier
Direct, 2010
Direct, 2004
0
50,000
100,000
150,000
Estimated group size

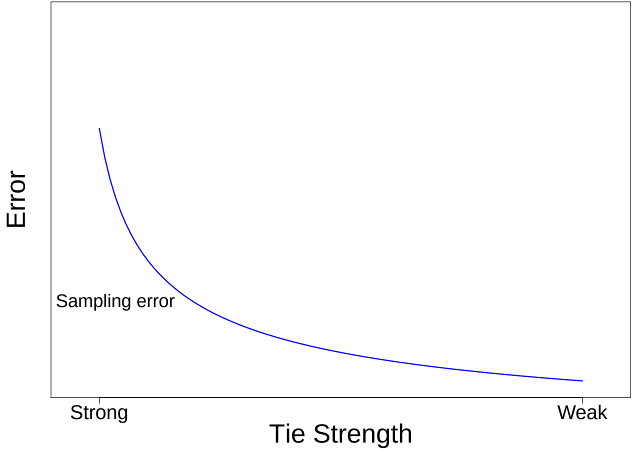


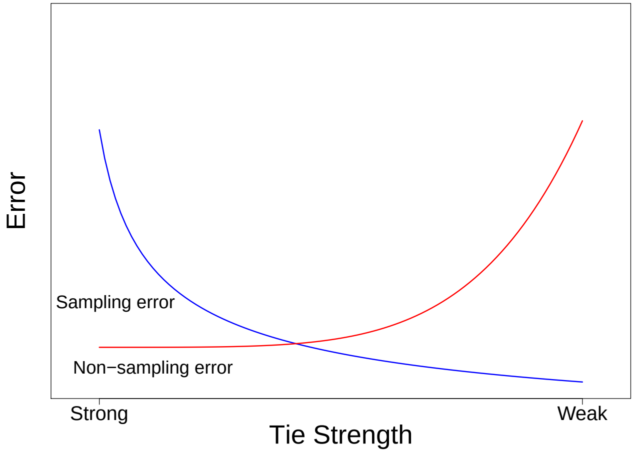
Heavy Drug Users, Curitiba, Brazil
Generalized scale-up
Scale-up
Multiplier
Direct, 2010
Direct, 2004
0
50,000
100,000
150,000
Estimated group size

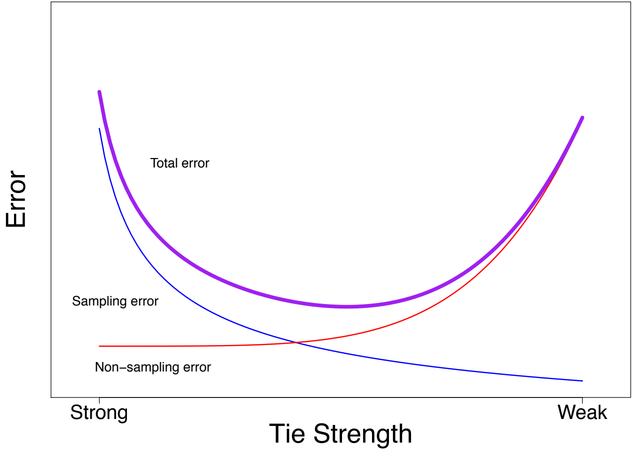
Motivates empirical work to

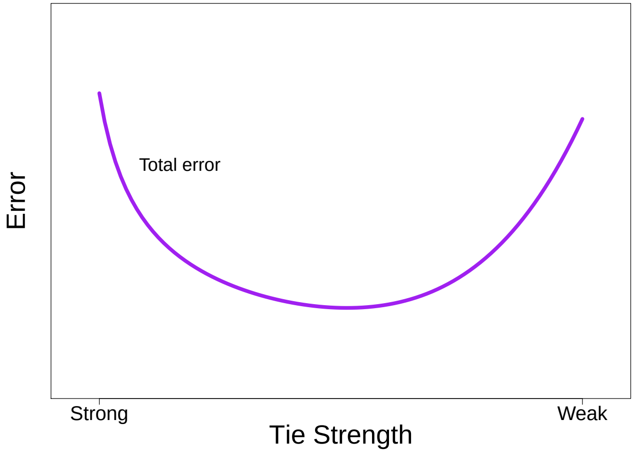
- estimate the visible degree of the hidden population
- optimize the definition of a network tie


Error
Strong
Weak
Tie Strength


Error
Sampling error
Strong
Weak
Tie Strength

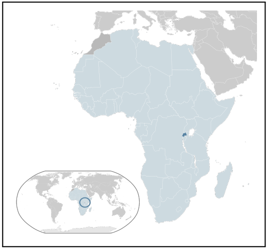

Error
Sampling error
Non−sampling error
Strong
Weak
Tie Strength


Error
Total error
Sampling error
Non–sampling error
Strong
Weak
Tie Strength


Error
Total error
Strong
Weak
Tie Strength

Survey experiment in Rwanda

- Nationally representative sample of 5,000 Rwandans
- Target populations: drug injectors, female sex workers, clients of female sex workers, men who have sex with men
- Funded by UNAIDS and USAID



Map source: Wikipedia

Let's unpack this:
"Our sample was drawn from the preparatory frame constructed for the 2012 Rwanda Census, which contained a complete list of 14,837 villages, which are the smallest administrative units in the country. We used a stratified, two-stage cluster design with these villages as the primary sampling units."

Let's unpack this:

$$\hat{N}_H = \frac{\sum_i y_{i,H}}{\sum_i \hat{d}_i} \times N$$

$$\hat{N}_H = \frac{\sum_i \frac{y_{i,H}}{\pi_i}}{\sum_i \frac{\hat{d}_i}{\pi_i}} \times N$$

- $\hat{N}_H$: number of people in the hidden population
- $y_{i,H}$: number of people in hidden population known by person i
- $\hat{d}_i$: estimated number of people known by person i
- N: number of people in the population

- Sampling is about trade-offs of cost and precision
- With standard probability designs, just see if what you are doing makes sense in a census
- The first question I always ask when someone comes to me with a sampling problem: what would you do if cost was not a constraint? If you can't answer that question, it is hard to develop a good sampling plan.

Basic definition $(n = 2,500)$

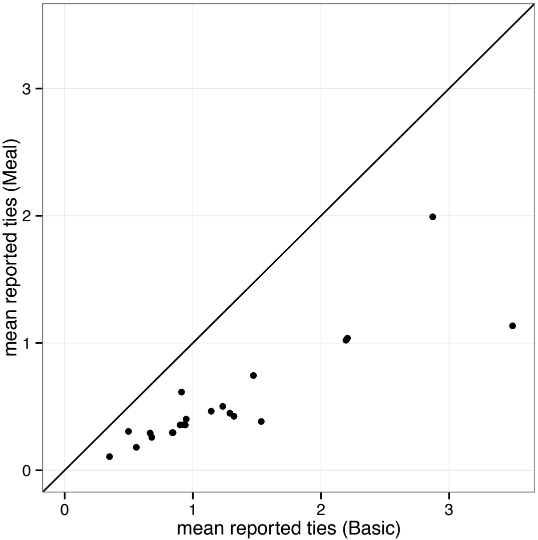
- people you know by sight and name and who also know you by sight and name
- people you have had some contact with in the past 12 months
- people of all ages who live in Rwanda

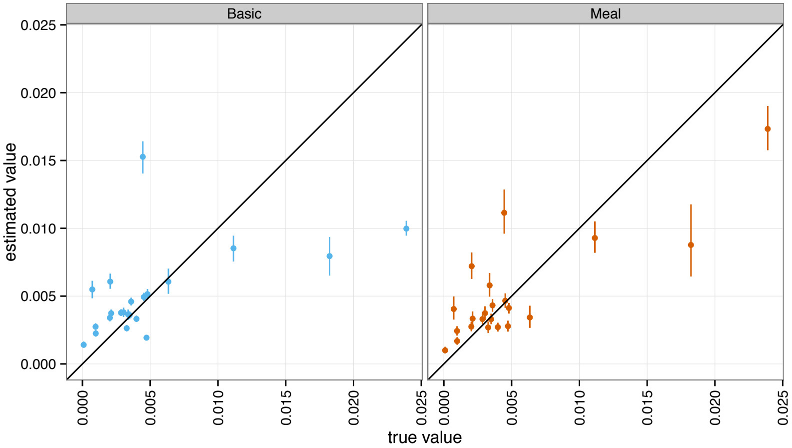
Meal definition $(n = 2,500)$

- people you know by sight and name and who also know you by sight and name
- people you have shared a meal or drink with in the past 12 months
- people of all ages who live in Rwanda

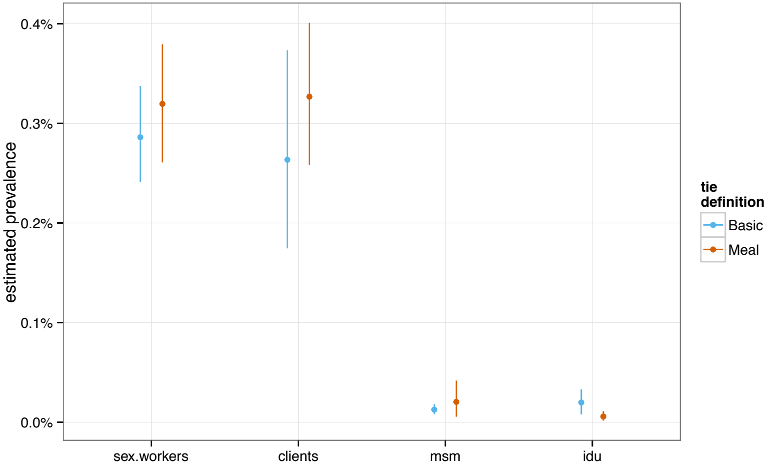
Priests
Nurses or Doctors
Male Community Health Worker
Widowers
Teachers
Divorced Men
Incarcerated people
Women who smoke
Muslim
Women who gave birth in the last 12 mo.

Twahirwa
Mukandekezi
Nyiraneza
Ndayambaje
Murekatete
Nsengimana
Mukandayisenga
Ndagijimana
Bizimana
Nyirahabimana
Nsabimana
Mukamana


mean reported ties (Meal)
mean reported ties (Basic)
0
1
2
3
0
1
2
3



Meal definition has lower error (RMSE, MAE, MRE)

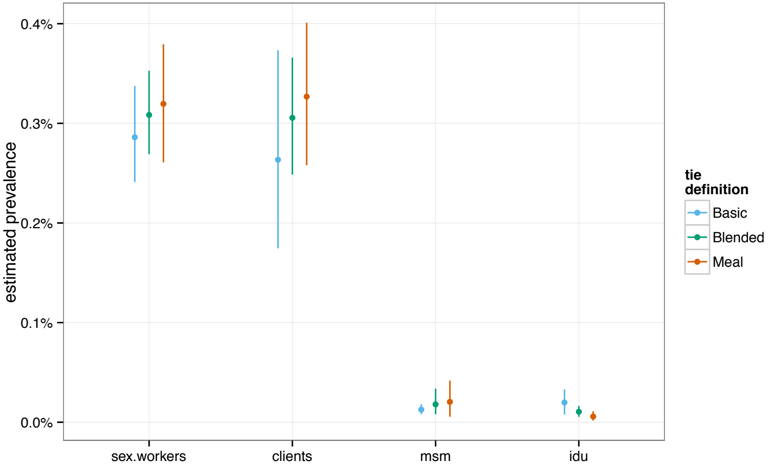


0.4%
0.3%
0.2%
0.1%
0.0%
estimated prevalence
sex.workers
clients
msm
idu
tie definition
Basic
Meal

$$\hat{N}_H = w \cdot \hat{N}_{H[meal]} + (1 - w) \cdot \hat{N}_{H[basic]}$$

$$\hat{N}_H = w \cdot \hat{N}_{H[meal]} + (1 - w) \cdot \hat{N}_{H[basic]}$$

$$w = \frac{\widehat{\sigma}^2_{basic}}{\widehat{\sigma}^2_{basic} + \widehat{\sigma}^2_{meal}}$$

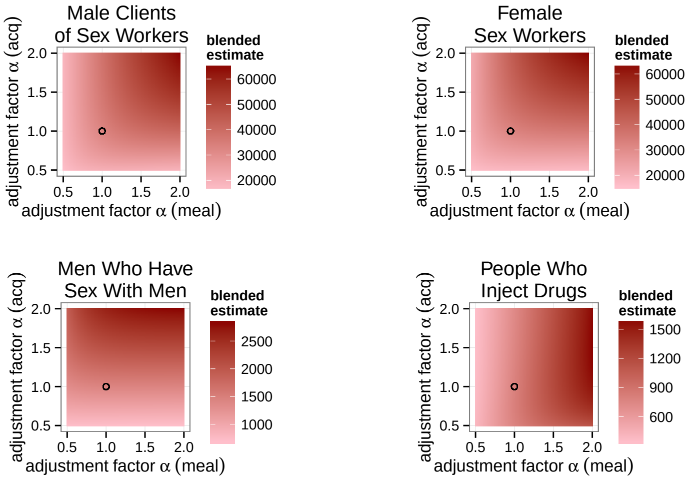


0.4%
0.3%
0.2%
0.1%
0.0%
estimated prevalence
sex.workers
clients
msm
idu
tie
definition
Basic
Blended
Meal

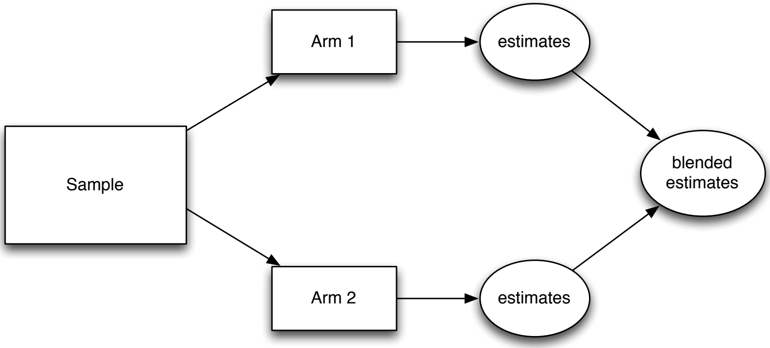
$$N_H = \alpha \hat{N}_H$$

where

$$\alpha = \underbrace{\left(\frac{\eta_F}{\tau_F}\right)}_{\text{reporting distortions}} \times \underbrace{\left(\frac{1}{\phi_F \delta_F}\right)}_{\text{structural distortions}}$$


Male Clients
of Sex Workers
blended
estimate
60000
50000
40000
30000
20000
adjustment factor α (acq)
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
adjustment factor α (meal)
Female
Sex Workers
blended
estimate
60000
50000
40000
30000
20000
adjustment factor α (acq)
adjustment factor α (meal)
Men Who Have
Sex With Men
blended
estimate
2500
2000
1500
1000
adjustment factor α (acq)
adjustment factor α (meal)
People Who
Inject Drugs
blended
estimate
1500
1200
900
600
adjustment factor α (acq)
adjustment factor α (meal)

Survey experiment with blending



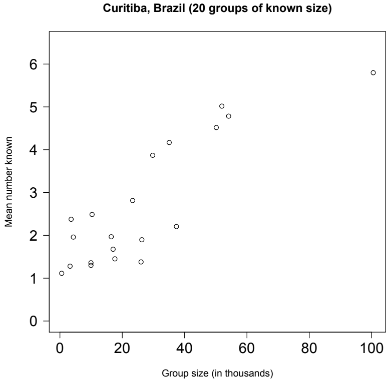
Modeling $\longleftrightarrow$ Empirical

Modeling $\longleftrightarrow$ Empirical

- false positive rate

Modeling $\longleftrightarrow$ Empirical

- false positive rate
- response error



Estimator	Imperfect assumptions	Effective estimand
$\widehat{\bar{d}}_{F,F}$ (Result B.3)	(i) $\widehat{N}_{\mathcal{A}} = c_1\ N_{\mathcal{A}}$ (ii) $\bar{d}_{\mathcal{A},F} = c_2\ \bar{d}_{F,F}$	$\frac{c_2}{c_1}\ \bar{d}_{F,F}$
$\widehat{\bar{d}}_{U,F}$ (Result B.4)	(i) $\widehat{N}_{\mathcal{A}} = c_1\ N_{\mathcal{A}}$ (ii) $\bar{d}_{\mathcal{A},F} = c_2\ \bar{d}_{U,F}$	$\frac{c_2}{c_1}\ \bar{d}_{U,F}$
$\widehat{\phi}_F$ (Result B.6)	(i) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_1\ \bar{d}_{F,F}$ (ii) $\widehat{\bar{d}}_{U,F} \rightsquigarrow c_2\ \bar{d}_{U,F}$	$\frac{c_1}{c_2}\ \phi_F$
$\widehat{\bar{v}}_{H,F}$ (Result C.2)	(i) $\widehat{N}_{\mathcal{A}\cap F} = c_1\ N_{\mathcal{A}\cap F}$ (ii) $\tilde{v}_{H,\mathcal{A}\cap F} = c_2\ v_{H,\mathcal{A}\cap F}$ (iii) $\frac{v_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = c_3\ \frac{v_{H,F}}{N_F}$	$\frac{c_3\ c_2}{c_1}\ \bar{v}_{H,F}$
$\widehat{\delta}_F$ (Result C.6)	(i) $\widehat{\bar{d}}_{H,F} \rightsquigarrow c_1\ \bar{d}_{H,F}$ (ii) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_2\ \bar{d}_{F,F}$	$\frac{c_1}{c_2}\ \delta_F$
$\widehat{\tau}_F$ (Result C.7)	(i) $\widehat{\bar{v}}_{H,F} \rightsquigarrow c_1\ \bar{v}_{H,F}$ (ii) $\widehat{\bar{d}}_{H,F} \rightsquigarrow c_2\ \bar{d}_{H,F}$	$\frac{c_1}{c_2}\ \tau_F$
$\widehat{N}_H$ (Result C.8)	(i) $\widehat{\bar{v}}_{H,F} \rightsquigarrow c_1\ \bar{v}_{H,F}$	$\frac{1}{c_1}\ N_H$
$\widehat{N}_H$ (Result C.10)	(i) $\widehat{\bar{d}}_{F,F} \rightsquigarrow c_1\ \bar{d}_{F,F}$ (ii) $\widehat{\delta}_F \rightsquigarrow c_2\ \delta_F$ (iii) $\widehat{\tau}_F \rightsquigarrow c_3\ \tau_F$	$\frac{1}{c_1\ c_2\ c_3}\ N_H$

Appendix D of Salganik and Feehan (2014)

Fails when:

- adjustment factors measured poorly
- adjustment factors not measured
- reporting not consistent with awareness
- variance estimation fails
- . . . .

Papers:

- Feehan and Salganik (2014) Estimating the size of hidden populations using the generalized network scale-up estimator. *arXiv*.
- Salganik, Mello, Abdo, Bertoni, Fazito, and Bastos (2011) The game of contacts: Estimating the social visibility of groups. *Social Networks*.
- Salganik, Fazito, Bertoni, Abdo, Mello, and Bastos (2011) Assessing network scale-up estimates for groups most at risk for HIV/AIDS: Evidence from a multiple method study of heavy drug users in Curitiba, Brazil. *American Journal of Epidemiology*.

Data and code:

- http://opr.princeton.edu/archive/nsum/
- http://opr.princeton.edu/archive/gc/
- R package `networkreporting`, available on CRAN (stable) & github (development)

$$\underbrace{N_T}_{\text{size of target pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in U} v_{i,F} / N_T\right)}_{\text{in-reports per member of target pop}}}$$

$$\underbrace{N_T}_{\text{size of target pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in U} v_{i,F} / N_T\right)}_{\text{in-reports per member of target pop}}}$$

If there are no false positives,

$$\underbrace{N_T}_{\text{size of target pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in T} v_{i,F} / N_T\right)}_{\text{avg visible degree of target pop}}}$$

Generalized scale-up identity

$$\underbrace{N_T}_{\text{size of target pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in T} v_{i,F} / N_T \right)}_{\text{avg visible degree of target pop}}}$$

Basic scale-up estimator

$$\hat{N}_T = \frac{\sum_{i \in s_F} y_{i,T}}{\sum_{i \in s_F} \hat{d}_i} \times N$$

Generalized scale-up identity

$$\underbrace{N_T}_{\text{size of target pop}} = \frac{\overbrace{\sum_{i \in F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\left(\sum_{i \in T} v_{i,F} / N_T \right)}_{\text{avg visible degree of target pop}}}$$

Basic scale-up estimator

$$\underbrace{\hat{N}_T}_{\text{est size of target pop}} = \frac{\overbrace{\sum_{i \in s_F} y_{i,T}}^{\text{total out-reports}}}{\underbrace{\sum_{i \in s_F} \hat{d}_{i,U} / N}_{\text{avg degree of pop}}}$$