

Class 22: Network scale-up method

Matthew J. Salganik

Sociology 204: Social Networks
Princeton University



Announcements:

- ▶ Welcome back

Announcements:

- ▶ Welcome back
- ▶ Feedback on the feedback about the Facebook Files

Announcements:

- ▶ Welcome back
- ▶ Feedback on the feedback about the Facebook Files
- ▶ Information about the final exam was posted this morning

Community Minute

There are an estimated 38 million people [31.6 million–44.5 million] living with HIV in 2019. In most countries, the disease is concentrated in three high risk groups:

- ▶ 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/AIDS: “know your epidemic”

Two main questions:

- ▶ prevalence of some trait within a hidden population (e.g., What proportion of sex workers in New York City have HIV/AIDS?):

Two main questions:

- ▶ prevalence of some trait within a hidden population (e.g., What proportion of sex workers in New York City have HIV/AIDS?): respondent-driven sampling

Two main questions:

- ▶ prevalence of some trait within a hidden population (e.g., What proportion of sex workers in New York City have HIV/AIDS?): respondent-driven sampling
- ▶ size of hidden population (e.g., How many sex workers are there in New York City?)

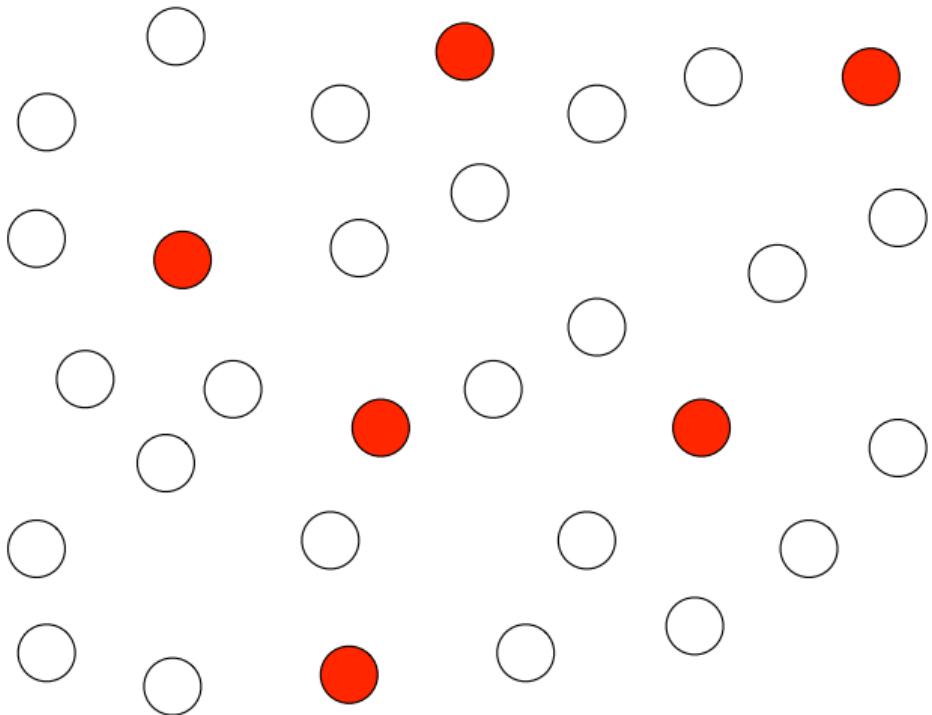
Two main questions:

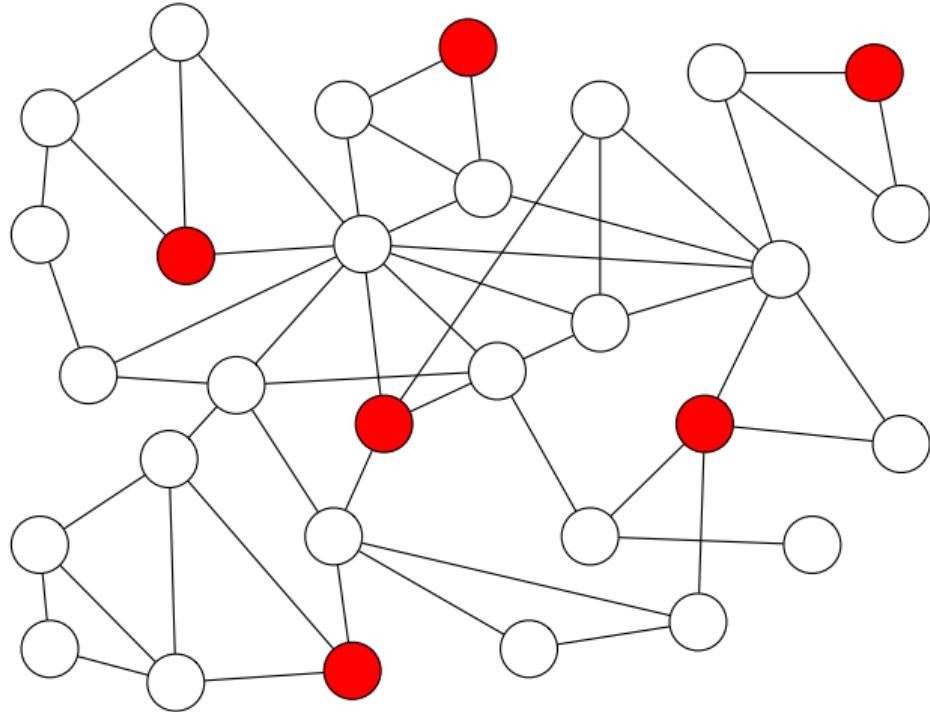
- ▶ prevalence of some trait within a hidden population (e.g., What proportion of sex workers in New York City have HIV/AIDS?): respondent-driven sampling
- ▶ size of hidden population (e.g., How many sex workers are there in New York City?) network scale-up method (this lecture)

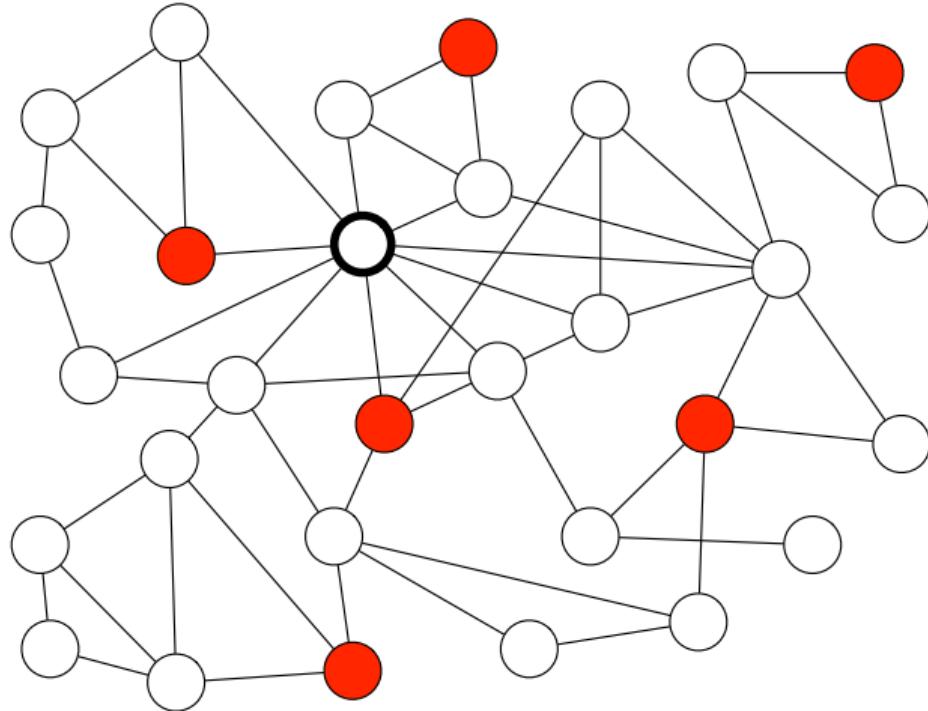
Network scale-up method

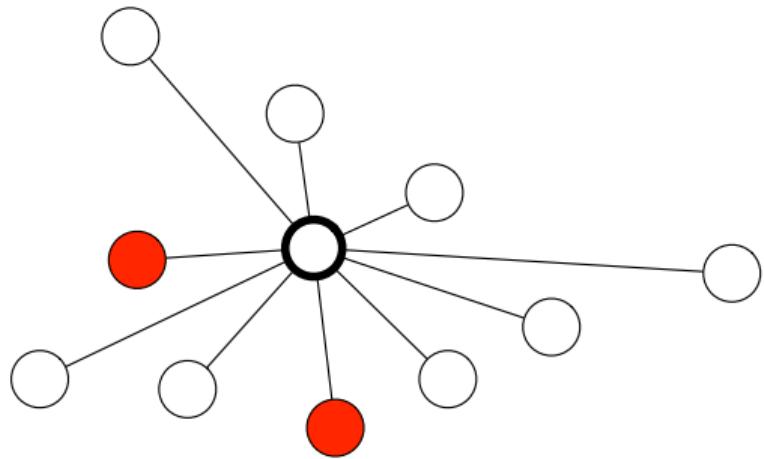


Basic insight from Bernard et al. (1989)



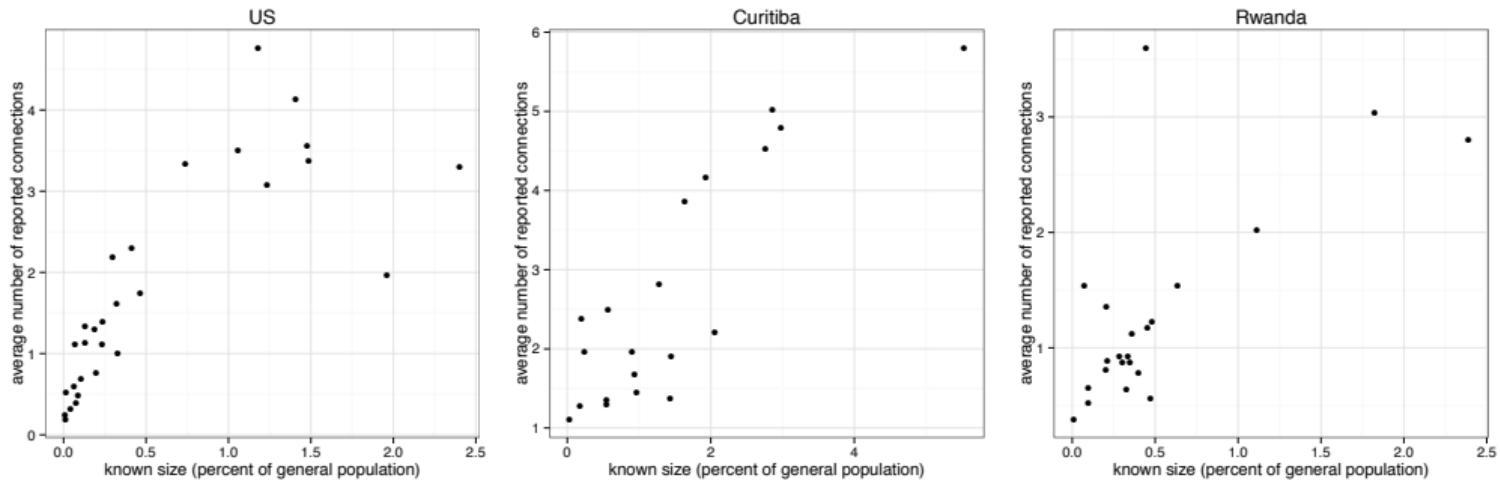






$$\hat{N}_H = \frac{2}{10} \times 30 = 6$$

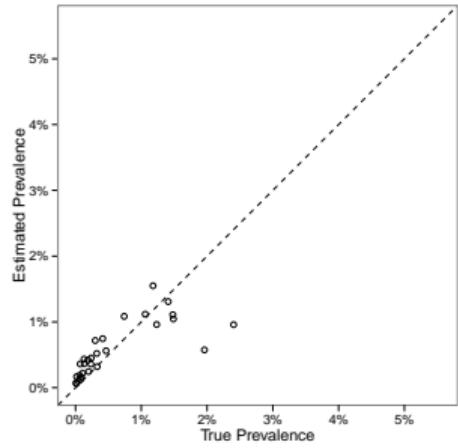
- ▶ Requires a random sample from the entire population
- ▶ Respondents are asked:
 - ▶ How many people do you know who are drug injectors?
 - ▶ How many women do you know that have given birth in the last 12 months?
 - ▶ How many people do you know who are middle school teachers?
 - ▶ ...
 - ▶ How many 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



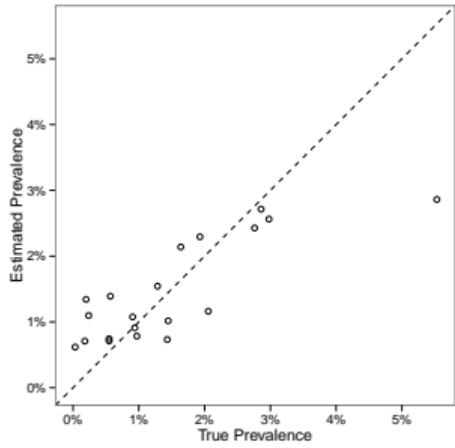
On average, these answers are not crazy.

Other size estimation methods are problematic, and scale-up method has many nice properties:

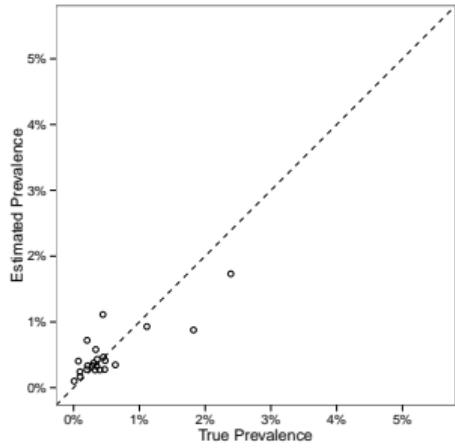
- ▶ Requires a random sample of the general population, not specific contact with the hard-to-reach population
- ▶ Can be added as a module (5-10 minutes) in any existing survey
- ▶ Can estimate many target populations in a single survey
- ▶ Can be applied at the city-level, sub-national-level, or national-level
- ▶ Statistical methods are improvable
- ▶ Partially self-validating because it uses groups of known size



(a) United States



(b) Curitiba



(c) Rwanda

But, basic scale-up also has problems. I will focus on insights about basic scale-up that we discovered from developing the generalized scale-up.

Personal background

How did I end up working on this research?

How did I end up working on this research?

- ▶ Working on sampling at the Census Bureau

How did I end up working on this research?

- ▶ Working on sampling at the Census Bureau
- ▶ When I began grad school I knew about sampling and was interested in networks and Doug Heckathorn was working on respondent-driven sampling

How did I end up working on this research?

- ▶ Working on sampling at the Census Bureau
- ▶ When I began grad school I knew about sampling and was interested in networks and Doug Heckathorn was working on respondent-driven sampling
- ▶ Respondent-driven sampling lead to the network scale-up method

Modeling
counting with multiplicity



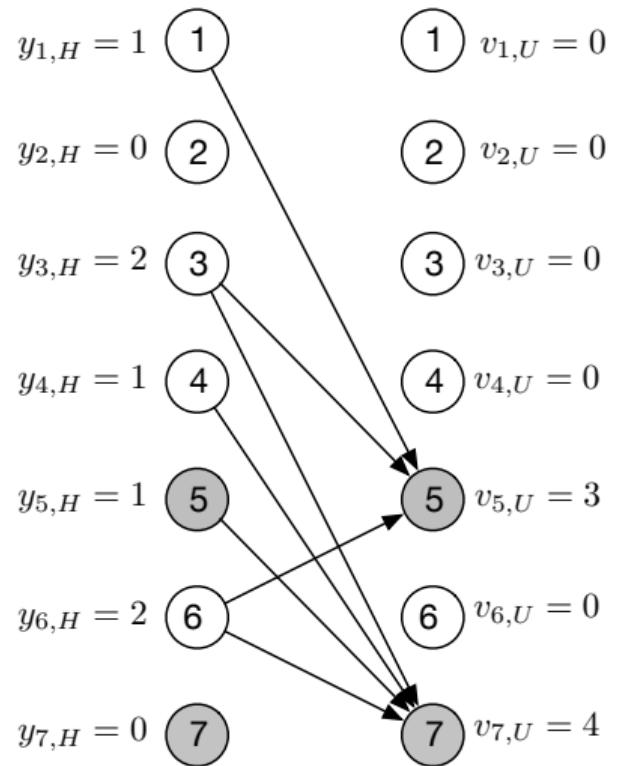
Empirical
Rwanda (this lecture)
Brazil (next lecture)

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)



total out-reports = total in-reports

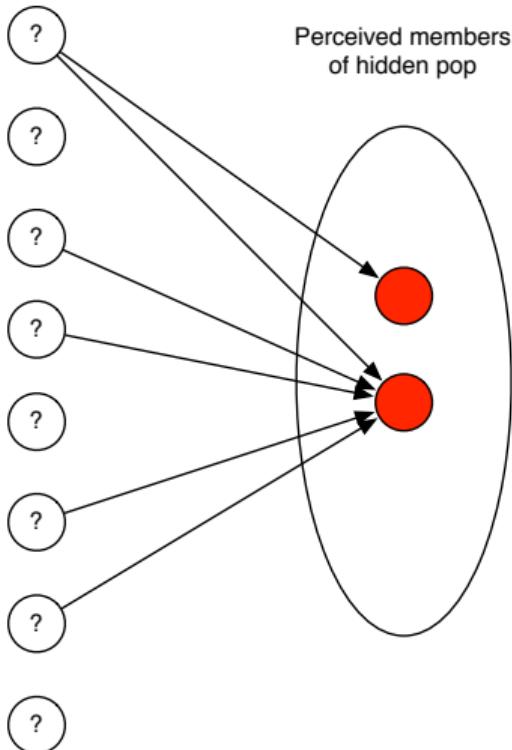
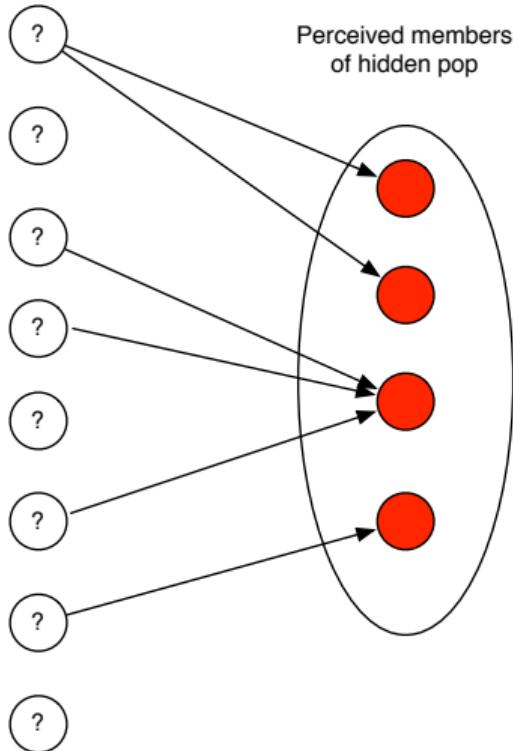
total out-reports = total in-reports

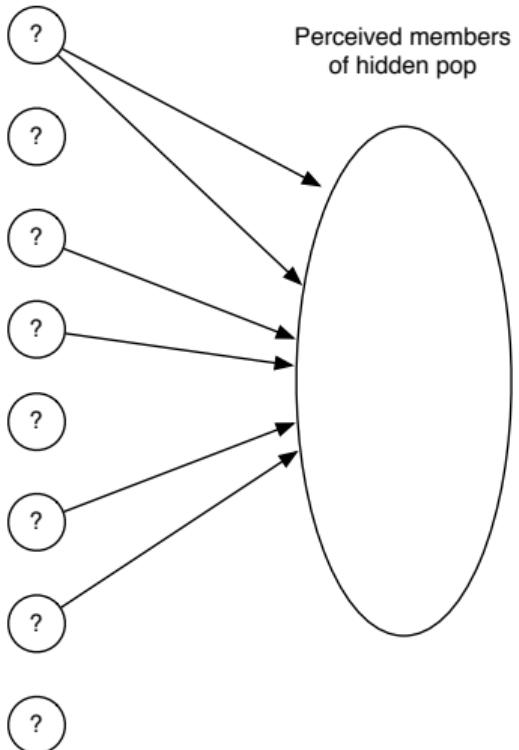
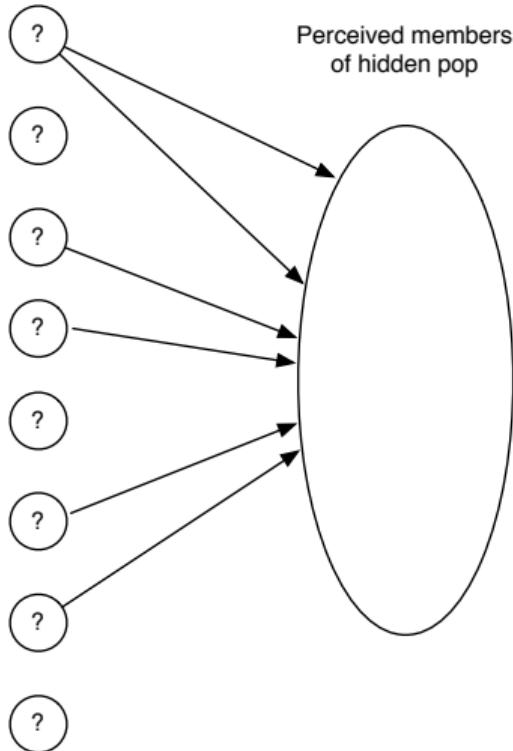
total out-reports = size of hidden pop ×
in-reports per member of hidden pop

total out-reports = total in-reports

total out-reports = size of hidden pop ×
in-reports per member of hidden pop

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





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

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

If there are no false positives,

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

Generalized scale-up identity

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

Basic scale-up estimator

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

Generalized scale-up identity

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

Basic scale-up estimator

$$\underbrace{\hat{N}_H}_{\text{est size of hidden pop}} = \frac{\overbrace{\sum_{i \in s_F} y_{i,H}}^{\text{total out-reports}}}{\sum_{i \in s_F} \hat{d}_{i,U} / N}$$

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

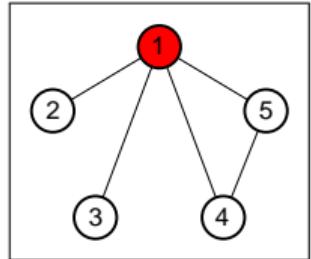
$$N_H = \underbrace{\left(\frac{y_{F,H}}{\bar{d}_{U,F}} \right)}_{\text{basic scale-up}} \times \underbrace{\frac{1}{\bar{d}_{F,F}/\bar{d}_{U,F}}}_{\substack{\text{frame ratio} \\ \phi_F}} \times \underbrace{\frac{1}{\bar{d}_{H,F}/\bar{d}_{F,F}}}_{\substack{\text{degree ratio} \\ \delta_F}} \times \underbrace{\frac{1}{\bar{v}_{H,F}/\bar{d}_{H,F}}}_{\substack{\text{true positive rate} \\ \tau_F}} = \underbrace{\left(\frac{y_{F,H}}{\bar{v}_{H,F}} \right)}_{\text{generalized scale-up}}. \quad (15)$$

adjustment factors

Frame ratio: Less a focus for us. As a first approximation, sampling frame is adults
you don't want to include kids in any of the reports

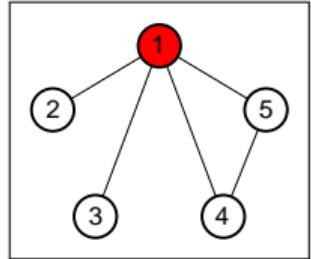
Degree ratio: If the hidden population has smaller network sizes than the general population, the size of the hidden population will be underestimated. Likewise, if the hidden population has larger network sizes than the general population, the size of the hidden population will be overestimated.

Set of egos can be different from set of alters.

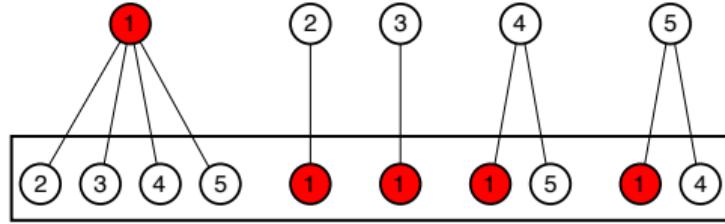


$$p = 0.2$$

Set of egos can be different from set of alters.

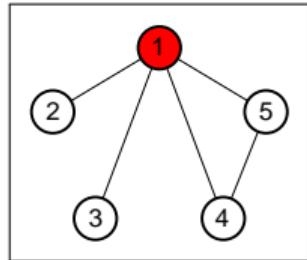


$$p = 0.2$$

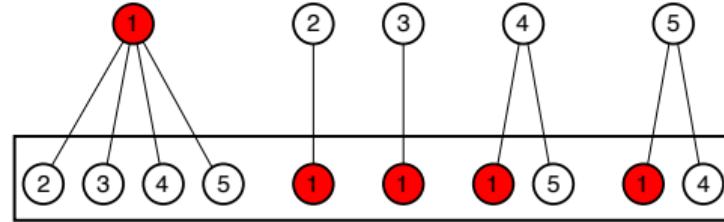


$$p_{alter} = 0.4$$

Set of egos can be different from set of alters.



$$p = 0.2$$



$$p_{alter} = 0.4$$

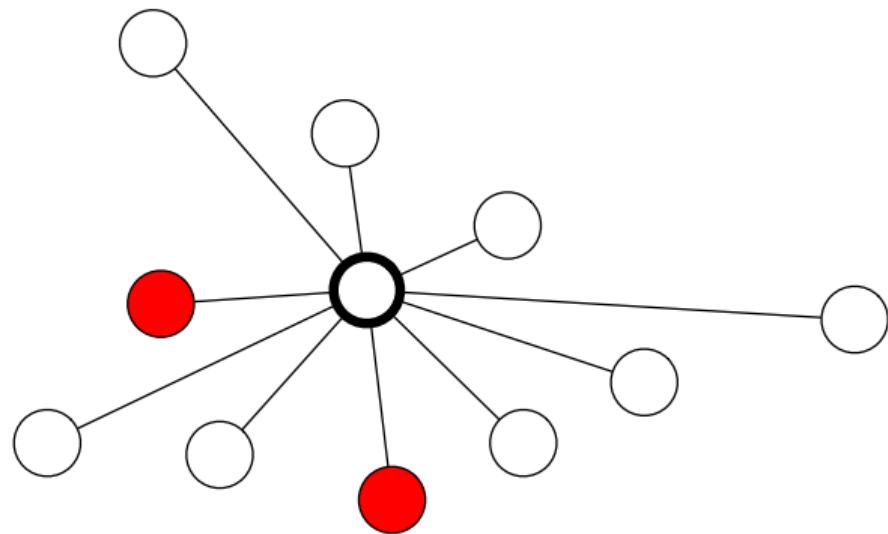
$$p_{alter} = p \times \frac{\text{avg. degree (hidden pop.)}}{\text{avg. degree (general pop.)}} = p\delta$$

Estimates will be biased by a factor of δ_F ("degree ratio")

True positive rate: If people are connected to people in the hidden population but not aware of it, the size of the hidden population will be underestimated

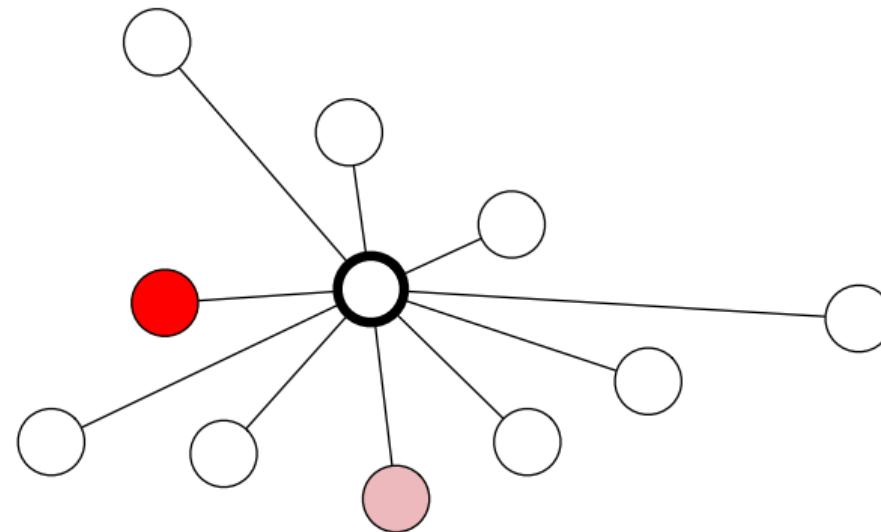
How might imperfect knowledge impact scale-up estimates?

Ego is not aware of everything about all of their alters.



How might imperfect knowledge impact scale-up estimates?

Ego is not aware of everything about all of their alters.



Estimates will be biased by a factor of τ_F ("true positive rate")

More about this in next lecture

$$N_H = \underbrace{\left(\frac{y_{F,H}}{\bar{d}_{U,F}} \right)}_{\text{basic scale-up}} \times \underbrace{\frac{1}{\bar{d}_{F,F}/\bar{d}_{U,F}}}_{\substack{\text{frame ratio} \\ \phi_F}} \times \underbrace{\frac{1}{\bar{d}_{H,F}/\bar{d}_{F,F}}}_{\substack{\text{degree ratio} \\ \delta_F}} \times \underbrace{\frac{1}{\bar{v}_{H,F}/\bar{d}_{H,F}}}_{\substack{\text{true positive rate} \\ \tau_F}} = \underbrace{\left(\frac{y_{F,H}}{\bar{v}_{H,F}} \right)}_{\text{generalized scale-up}}. \quad (15)$$

adjustment factors

From a talk I gave at a UNAIDS workshop

Generalized scale-up approach

- ▶ simple estimators (just addition, subtraction, multiplication, and division)
- ▶ handles incomplete social awareness
- ▶ no assumptions about the underlying social network
- ▶ handles incomplete frames
- ▶ handles complex sample designs

and could still be very wrong in practice!



Assumptions can be put into four broad categories

- ▶ sampling
- ▶ social network structure
- ▶ reporting
- ▶ survey construction

Results for non-sampling assumptions have this form:

Estimator	Imperfect assumptions	Effective estimand
$\widehat{\bar{d}}_{F,F}$ (Result B.3)	(i) $\widehat{N}_{\mathcal{A}} = c_1 N_{\mathcal{A}}$ (ii) $\bar{d}_{\mathcal{A},F} = c_2 \bar{d}_{F,F}$ (iii) $y_{F,\mathcal{A}} = c_3 d_{F,\mathcal{A}}$	$\frac{c_2 c_3}{c_1} \bar{d}_{F,F}$

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



Fails when:

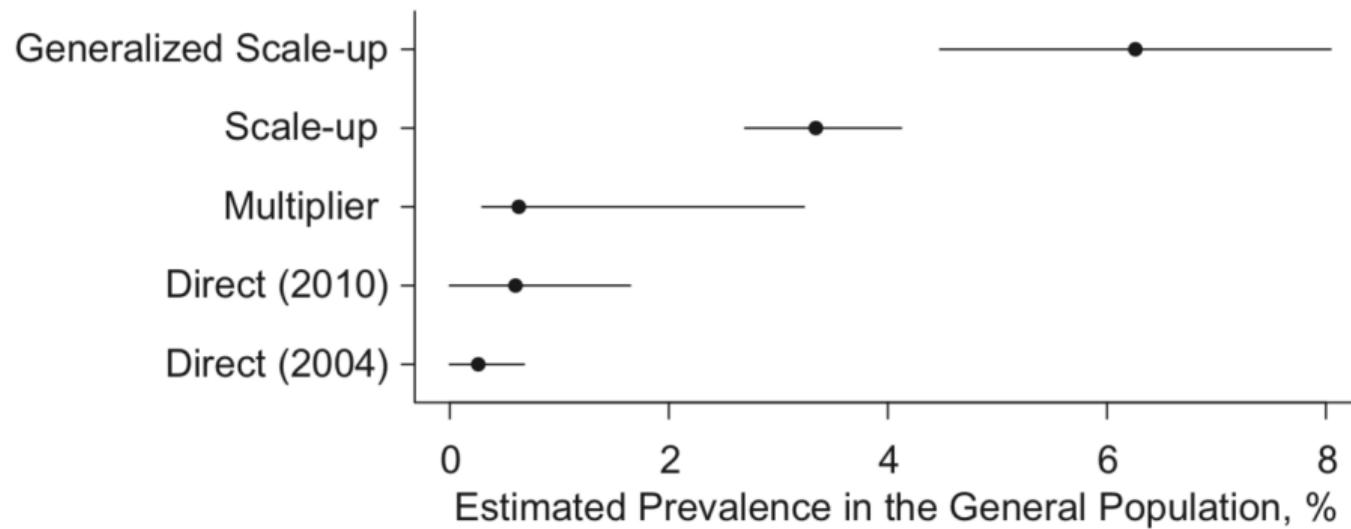
- ▶ adjustment factors not measured
- ▶ adjustment factors measured poorly
- ▶ reporting not consistent with awareness
- ▶ variance estimation fails
- ▶ . . .

Modeling
counting with multiplicity



Empirical
Rwanda (this lecture)
Brazil (next lecture)

How can we improve
if we don't know how we are doing?



Source: Salganik, M.J., Fazito, D., Bertoni, N., Abdo, A.H., Mello, M.B., and Bastos, F.I. (2011) *American Journal of Epidemiology*.

How can we reliably make progress in our ability to estimate quantities that will never be known?

How can we reliably make progress in our ability to estimate quantities that will never be known?

- ▶ Theoretical
- ▶ Empirical

Network scale-up study in Rwanda

Study in Rwanda was designed to estimate the number of

- ▶ men who have sex with men
- ▶ female sex workers
- ▶ clients of female sex workers
- ▶ injection drug users

AND

to produce generalizable knowledge about the scale-up method

		Consideration of use?	
		Yes	No
Quest for fundamental understanding?	Yes	Pure basic research (Bohr)	Use-inspired basic research (Pasteur)
	No		Pure applied research (Edison)

- ▶ Why was UNAIDS excited about this study?

- ▶ Why was UNAIDS excited about this study?
- ▶ Why was the Rwandan National AIDS program excited about this study?

- ▶ Why was UNAIDS excited about this study?
- ▶ Why was the Rwandan National AIDS program excited about this study?
- ▶ Why was I excited about this study?



NUR / SPH

NATIONAL UNIVERSITY
OF RWANDA

SCHOOL OF PUBLIC HEALTH

P.O.BOX: 5229 KIGALI TEL: (250)585166 FAX: (250)500014

E-mail: infospnur@yahoo.com

WEBSITE: www.espnur.org

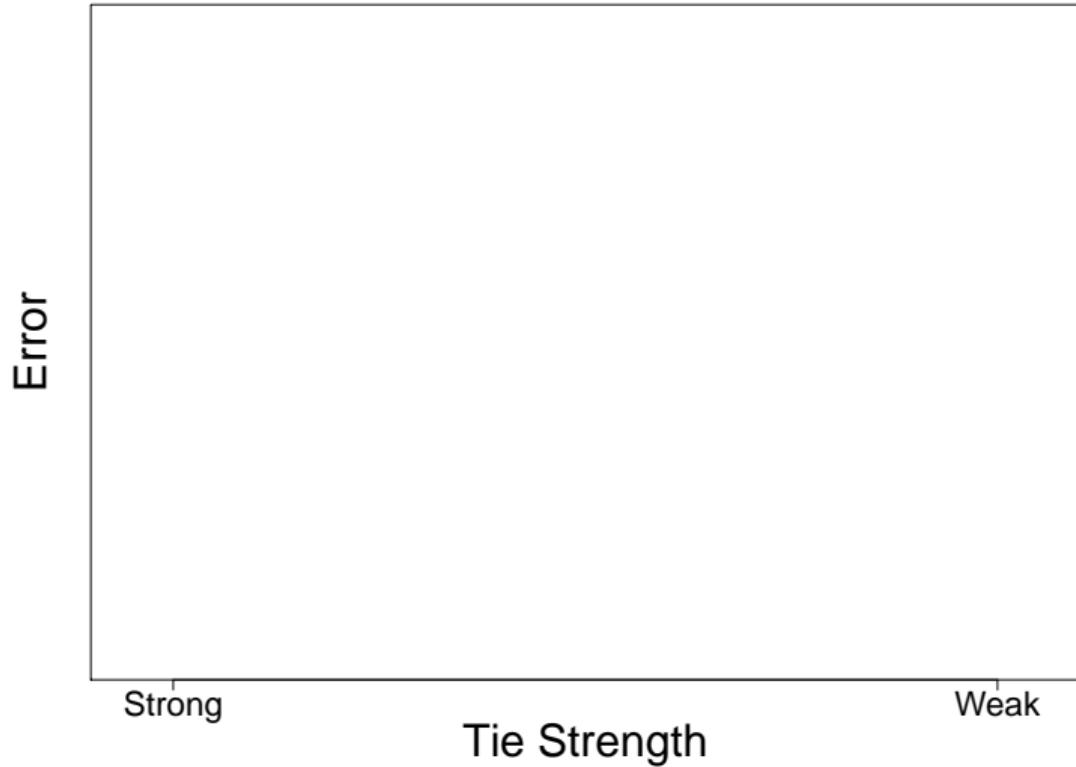


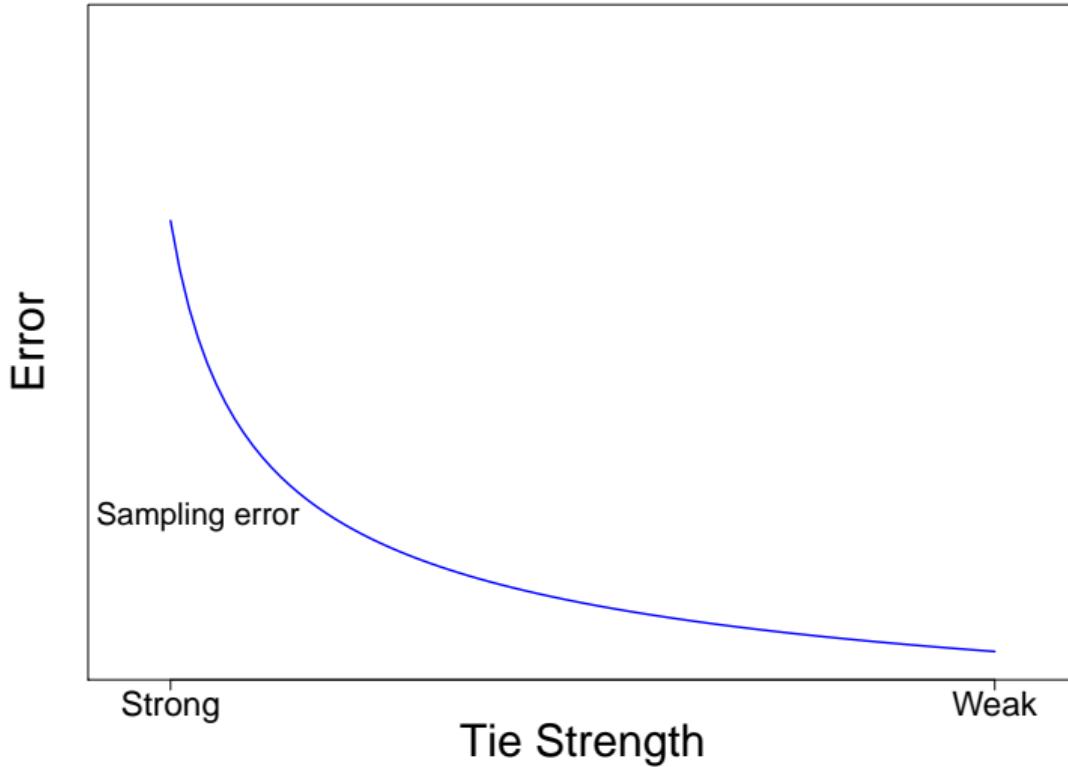


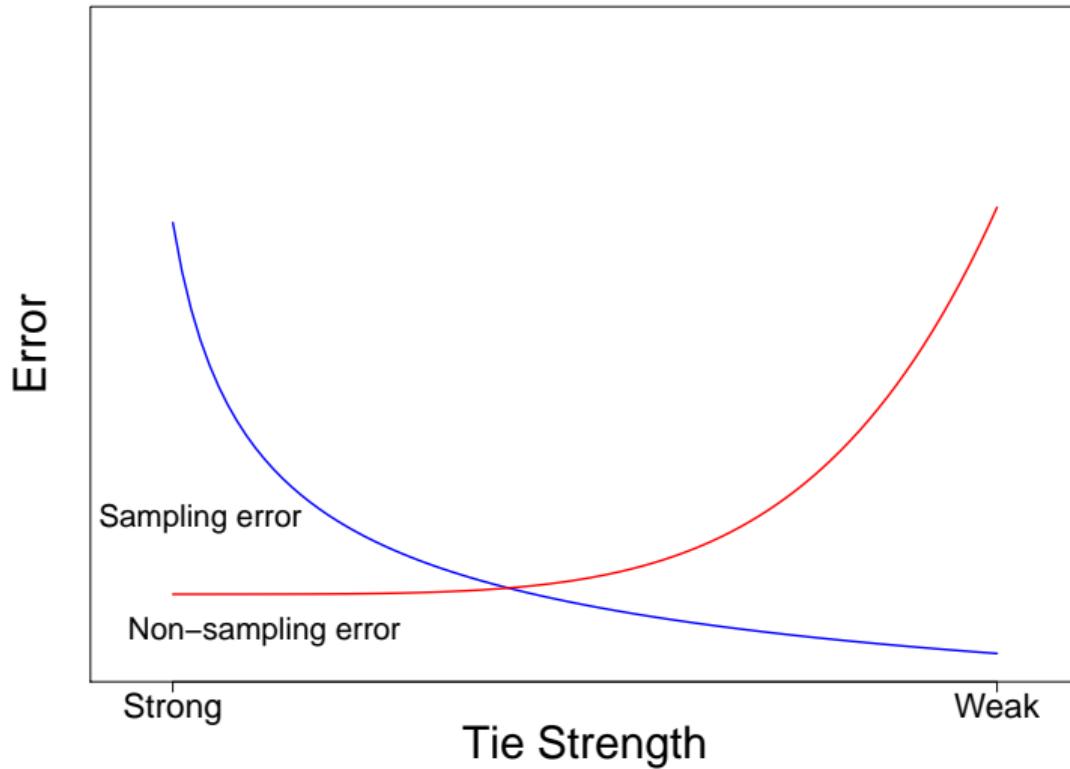


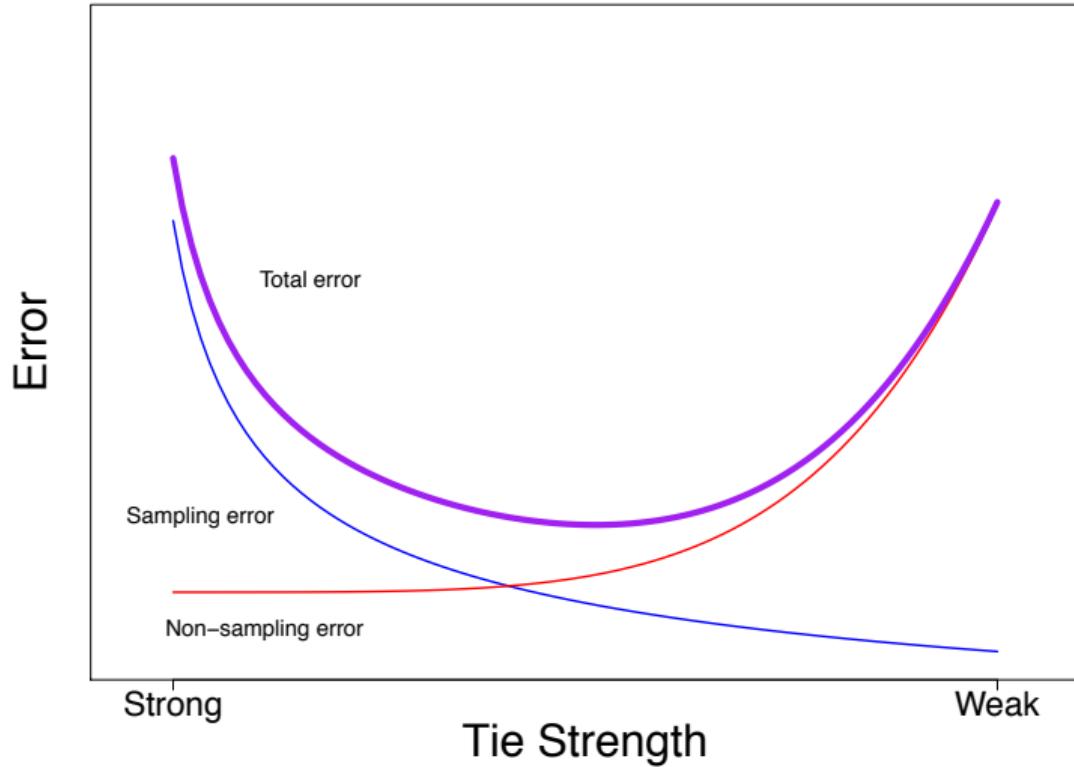


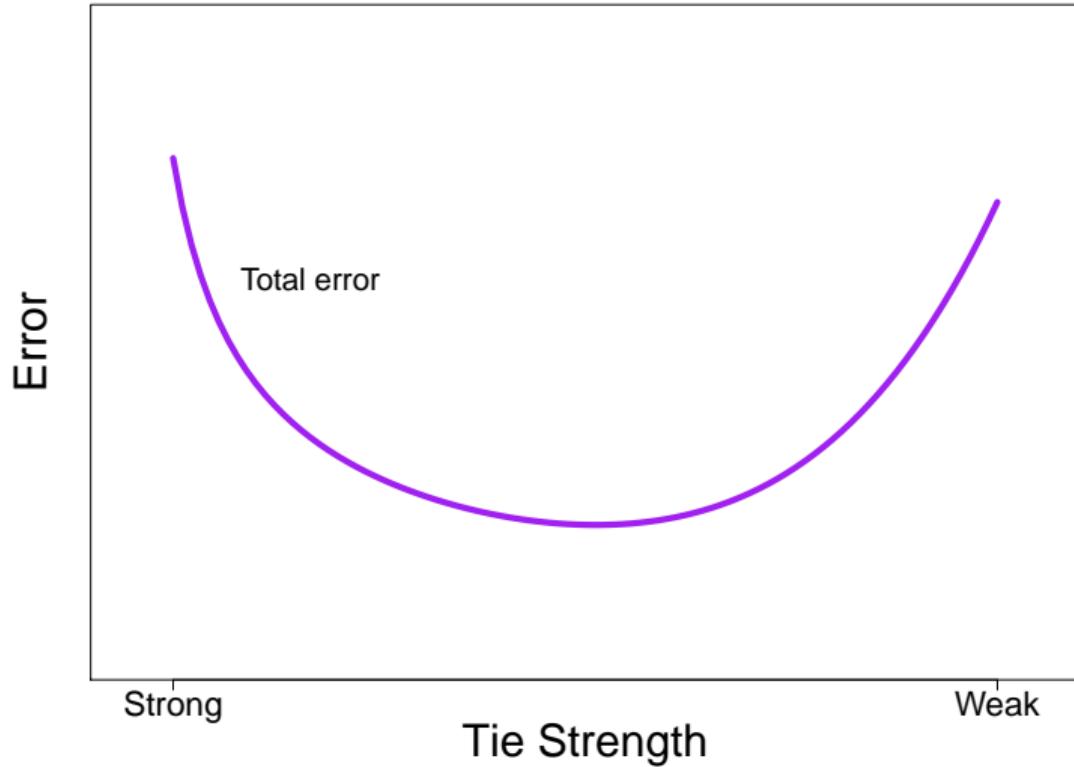












Basic definition ($n = 2,500$)

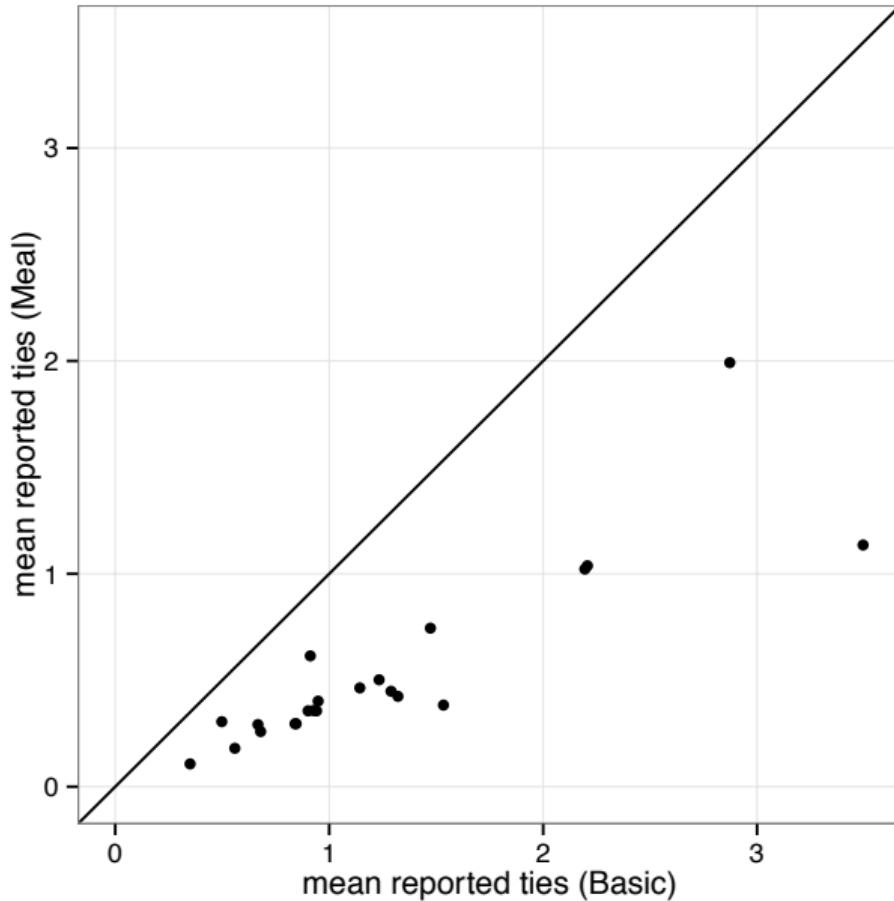
- ▶ people you know by sight and name and who also know you by sight and name
- ▶ people you have **had some contact with** in the past 12 months
- ▶ people of all ages who live in Rwanda

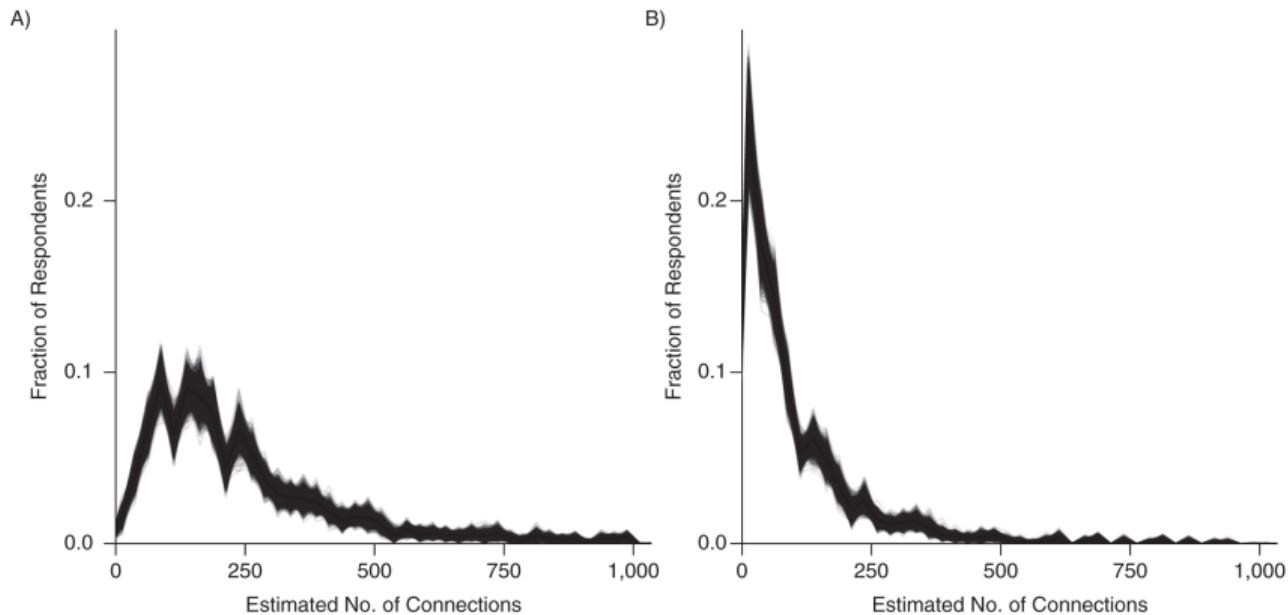
Meal definition ($n = 2,500$)

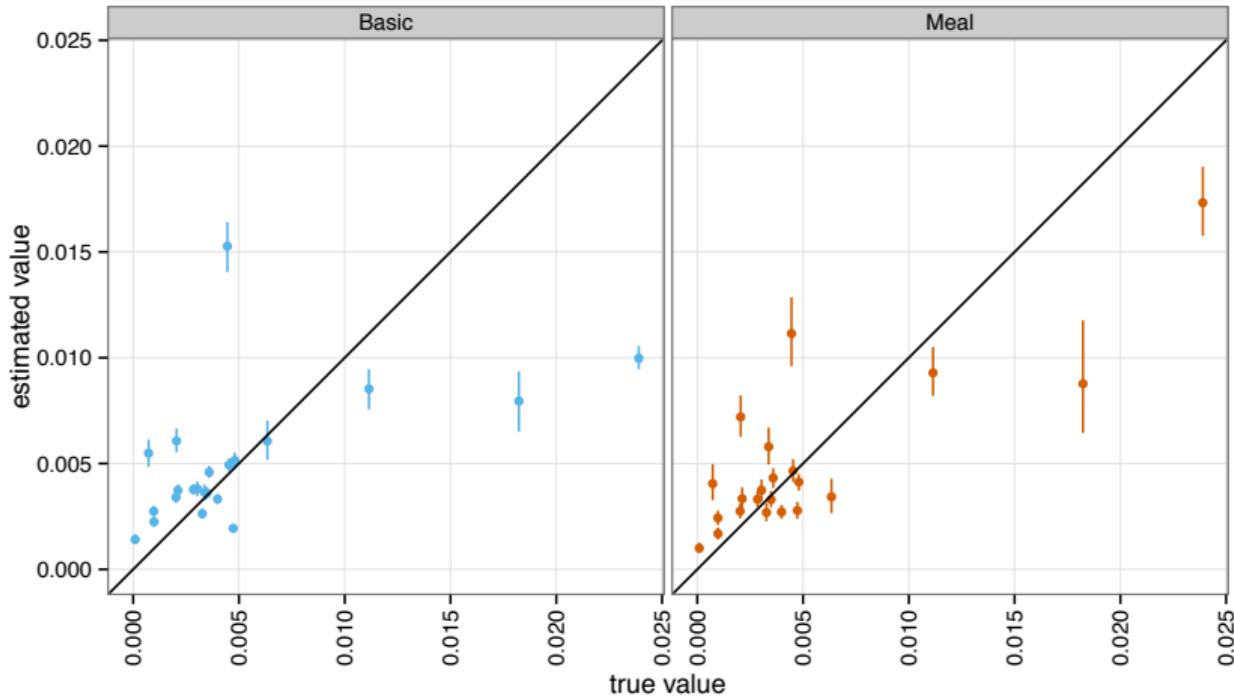
- ▶ people you know by sight and name and who also know you by sight and name
- ▶ people you have **shared a meal or drink with** in the past 12 months
- ▶ people of all ages who live in Rwanda

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.

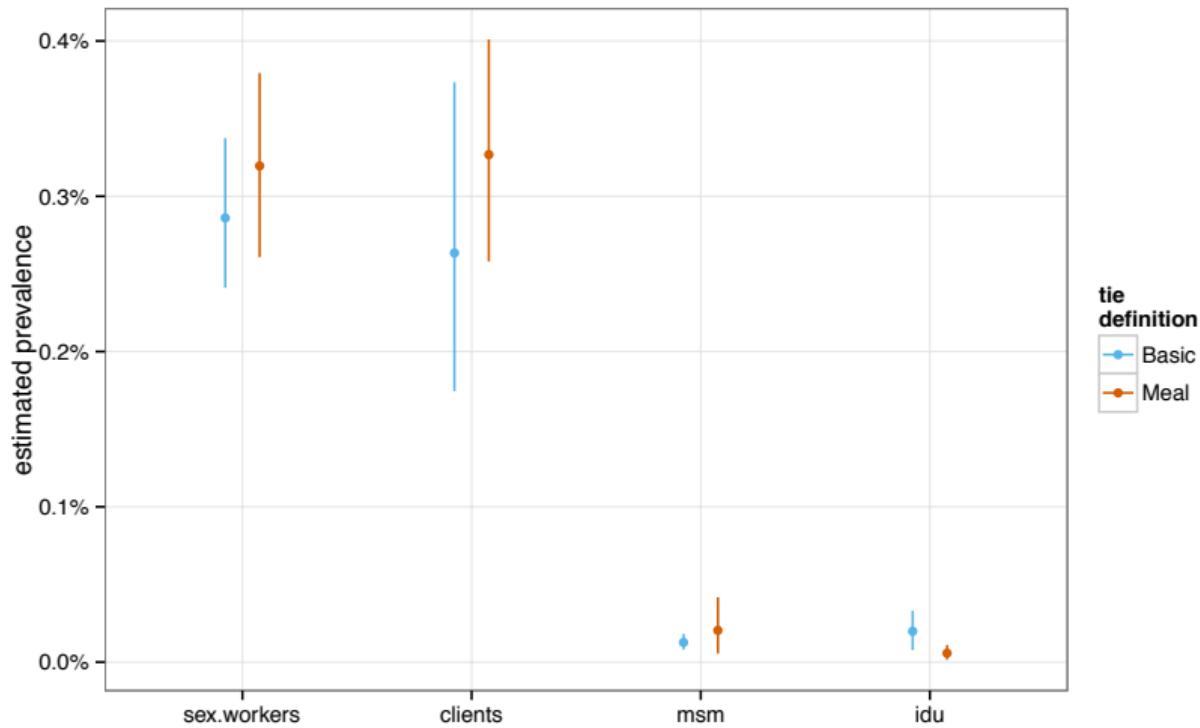
Twahirwa
Mukandekezi
Nyiraneza
Ndayambaje
Murekatete
Nsengimana
Mukandayisenga
Ndagijimana
Bizimana
Nyirahabimana
Nsabimana
Mukamana







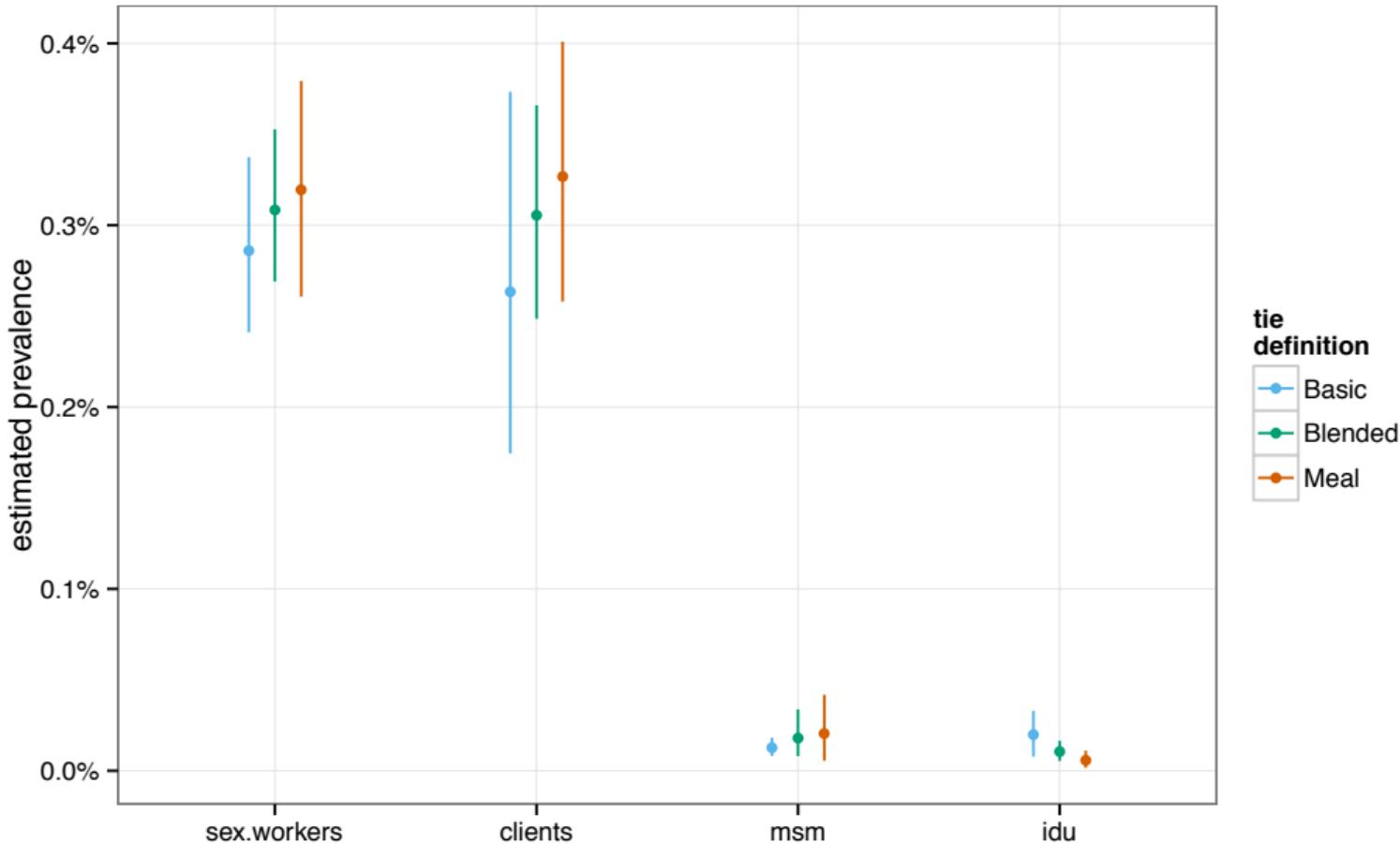
Meal definition has lower error (RMSE, MAE, MRE)



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

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

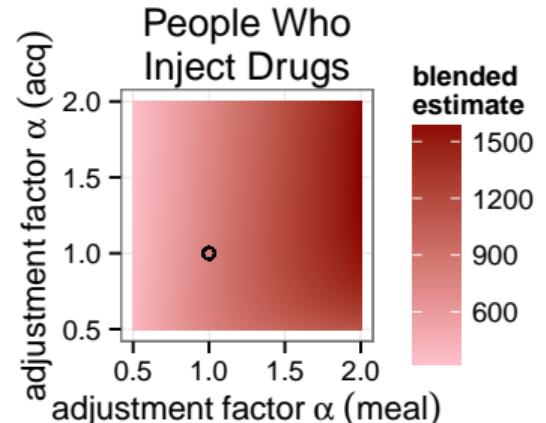
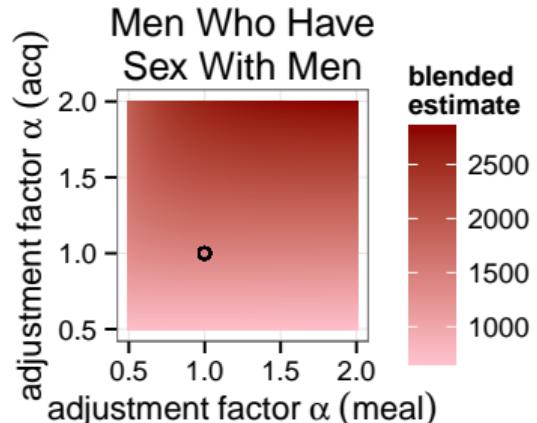
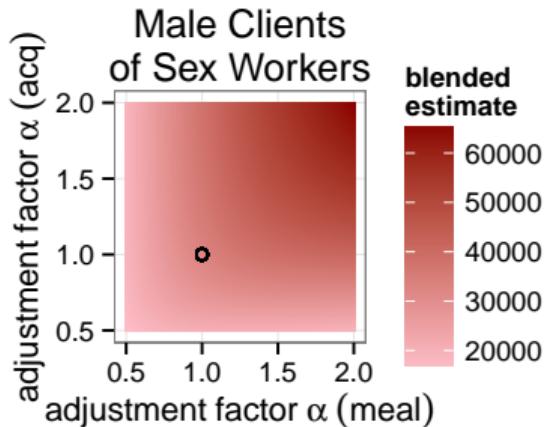
$$w = \frac{\hat{\sigma}^2_{basic}}{\hat{\sigma}^2_{basic} + \hat{\sigma}^2_{meal}}$$

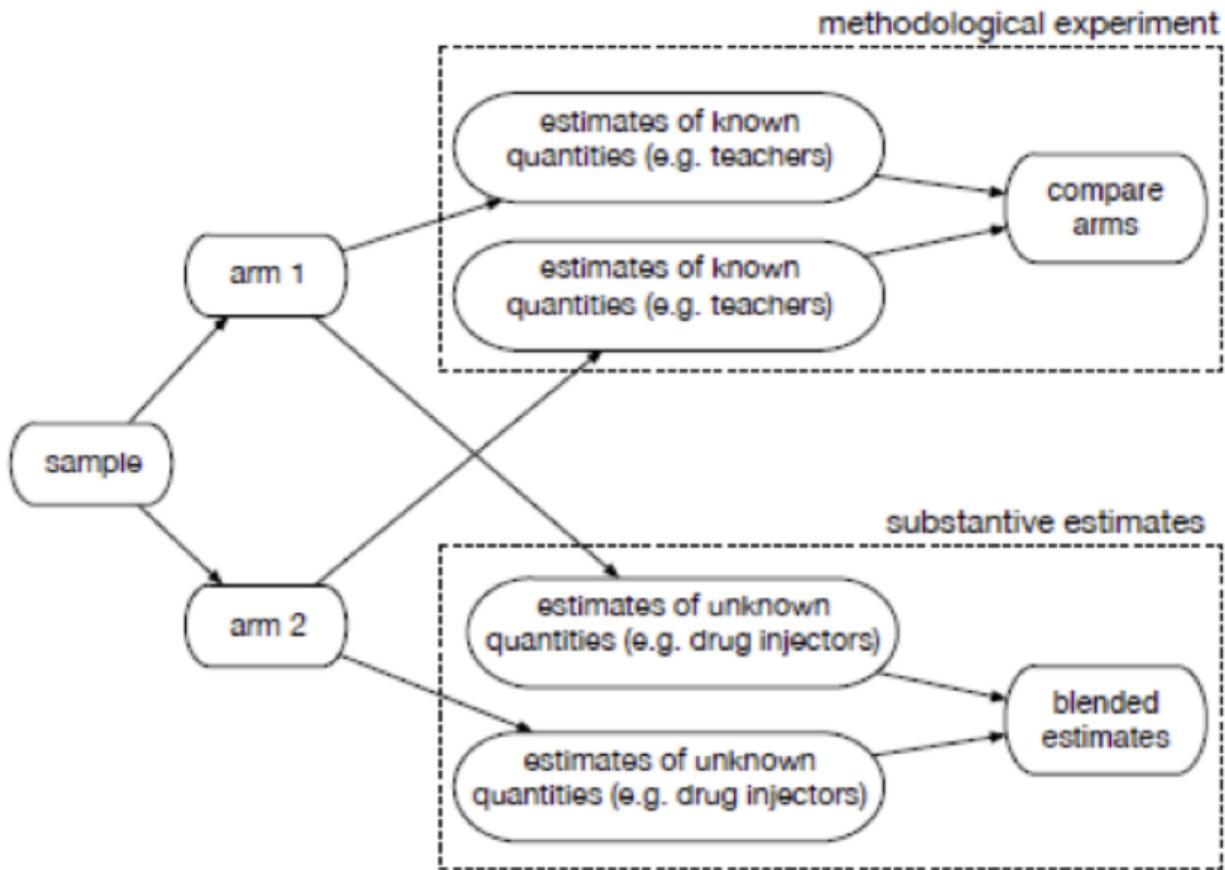


$$N_H = \alpha \hat{N}_H$$

where

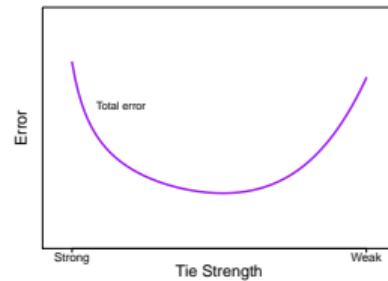
$$\alpha = \underbrace{\left(\frac{\eta_F}{\tau_F} \right)}_{\text{reporting distortions}} \times \underbrace{\left(\frac{1}{\phi_F \delta_F} \right)}_{\text{structural distortions}}$$



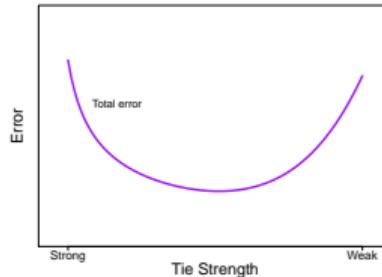


Extensions

With only two arms, we cannot demonstrate U-shaped relationship!



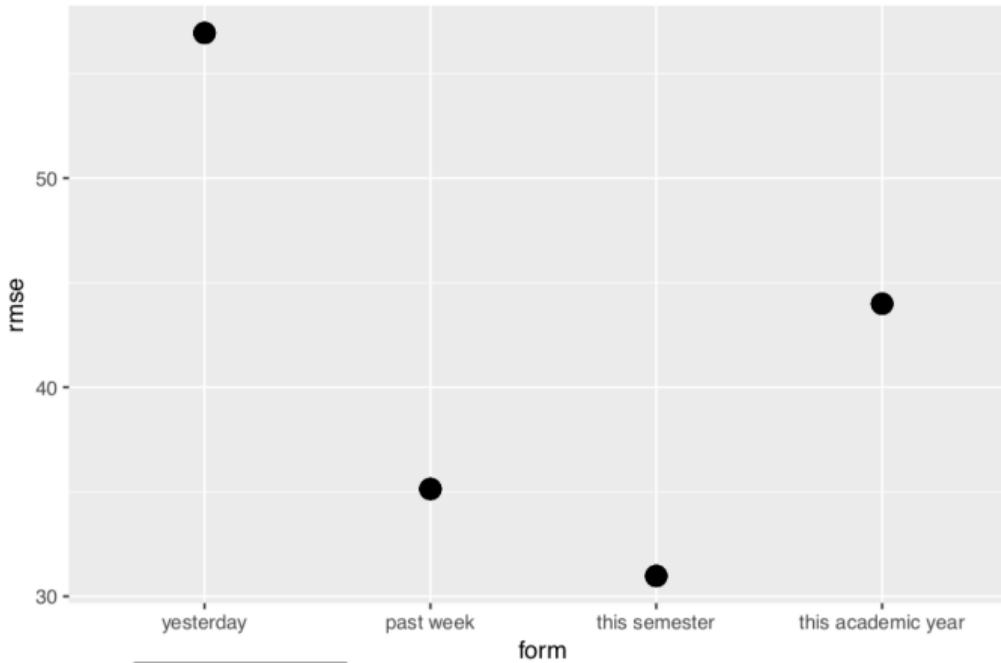
With only two arms, we cannot demonstrate U-shaped relationship!



In previous classes, we tried a survey experiment as a homework assignment in this class.

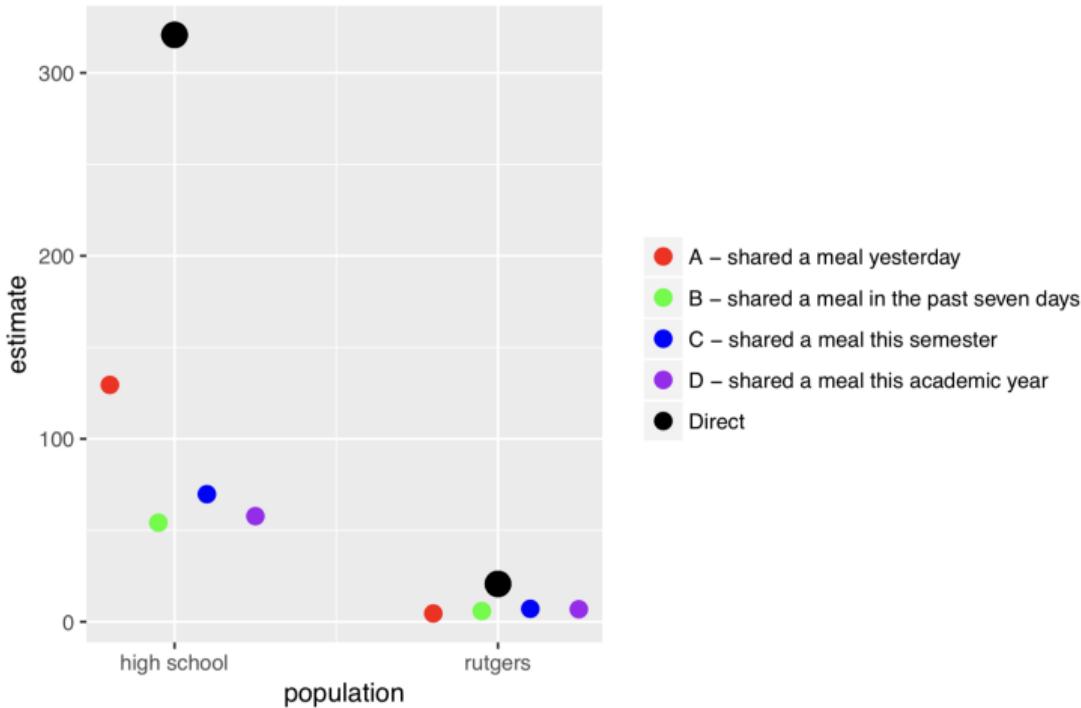
- ▶ We wanted to estimate number of people dating someone from their high school and number of people dating someone from Rutgers.
- ▶ We asked about connections to 9 groups of known size (e.g., sociology majors).
- ▶ We used 4 definitions of to know:
 1. shared a meal with yesterday
 2. shared a meal with in the past seven days
 3. shared a meal with this semester
 4. shared a meal with this academic year

Figure 8: Root Mean Squared Error, by Form



$$RMSE = \sqrt{\frac{\sum_{i=1}^n (estimate_i - actual_i)^2}{n}}$$

Figure 11: Direct and Scale–Up Estimates, by Form



- ▶ scale-up estimates might be too low because of imperfect visibility

Lecture 23: Who knows what about who?

- ▶ Salganik, M.J. et al. (2011). The game of contacts: Estimating the social visibility of groups. *Social Networks*.
- ▶ Cowan, S. (2014). Secrets and Misperceptions: The Creation of Self-Fulfilling Illusions. *Sociological Science*.