

# Using sampled social network data to estimate the size of hidden populations

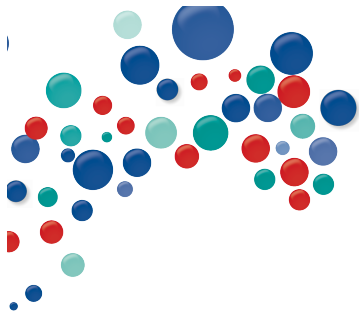
Dennis M. Feehan  
Dept of Demography  
UC Berkeley

Joint with: Matthew J. Salganik (Princeton), Mary Mahy (UNAIDS), Aline Umubyeyi (U. of Rwanda), Wolfgang Hladik (CDC)

WIND workshop, March 22, 2016

# The problem: estimating the size of hidden populations

- ▶ people who inject drugs
- ▶ sex workers
- ▶ clients of sex workers
- ▶ men who have sex with men



UNAIDS/WHO Working Group  
on Global HIV/AIDS and STI Surveillance

Guidelines  
on Estimating the Size of  
Populations Most at Risk to HIV



## Network scale-up method: the idea

Survey respondents have useful information about the people they are connected to in their personal networks

## Network scale-up method: the idea

Survey respondents have useful information about the people they are connected to in their personal networks

We can ask them: “how many drug injectors do you know?”

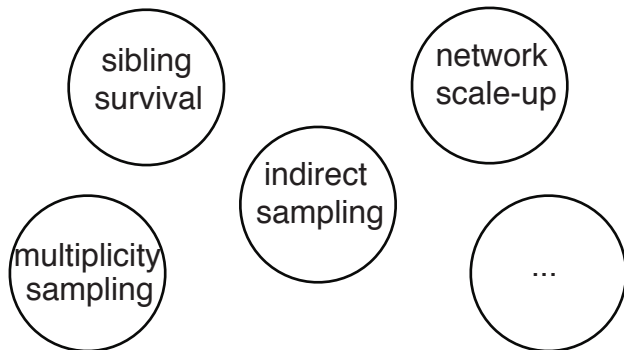
## Network scale-up method: the idea

Survey respondents have useful information about the people they are connected to in their personal networks

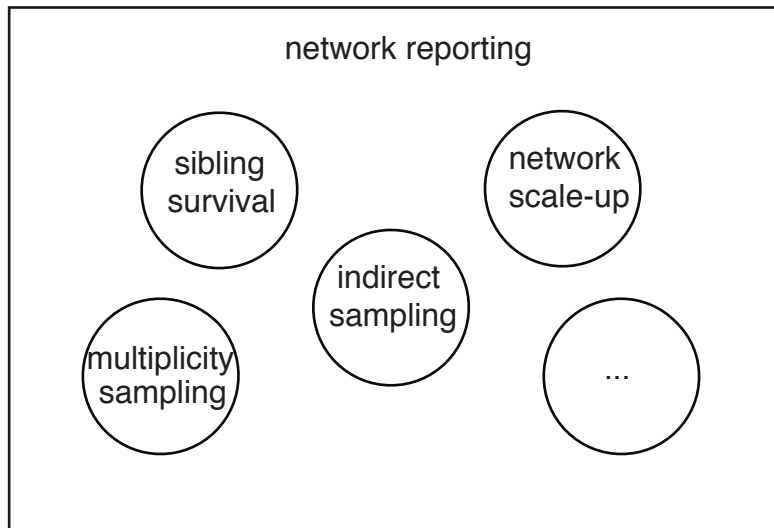
We can ask them: “how many drug injectors do you know?”

Bernard et al (1989); Killworth et al (1998b)

## Network scale-up and network reporting

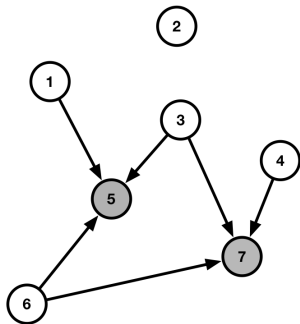


## Network scale-up and network reporting

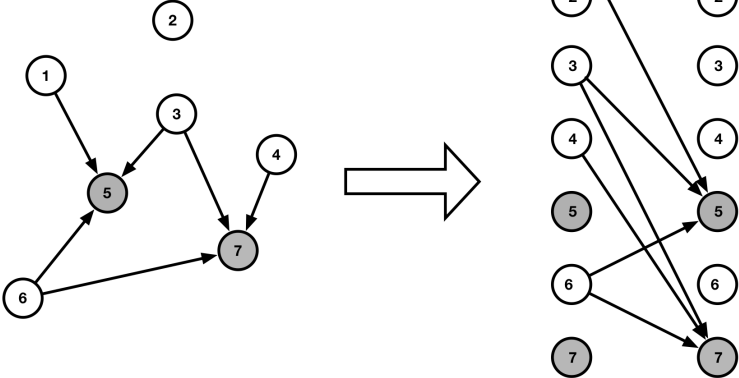


# Network reporting

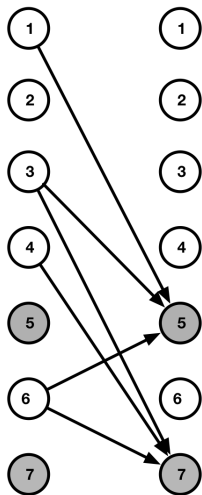
# Network reporting



# Network reporting

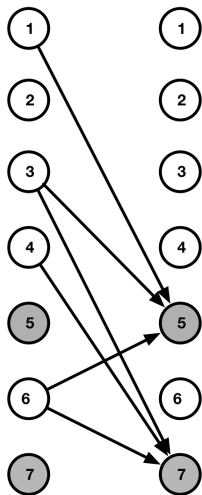


## Network reporting: deriving generalized scale-up



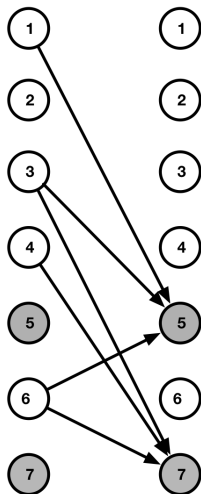
(1) total out-reports = total in-reports

## Network reporting: deriving generalized scale-up



- (1) total out-reports = total in-reports
- (2) total out-reports = number of drug inj.  $\times$  in-reports per drug inj.

## Network reporting: deriving generalized scale-up

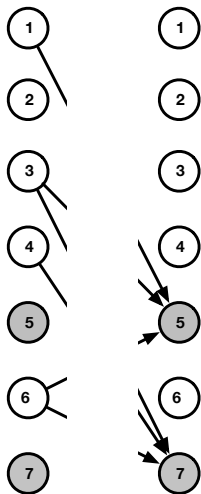


(1) total out-reports = total in-reports

(2) total out-reports = number of drug inj.  $\times$   
in-reports per drug inj.

(3) number of drug inj. =  $\frac{\text{total out-reports}}{\text{in-reports per drug inj.}}$

## Network reporting: deriving generalized scale-up



(1) total out-reports = total in-reports

(2) total out-reports = number of drug inj.  $\times$   
in-reports per drug inj.

(3) number of drug inj. =  $\frac{\text{total out-reports}}{\text{in-reports per drug inj.}}$

## Out-reports: Connections to hidden population members

$$\text{number of drug injectors} = \frac{\text{total out-reports}}{\text{in-reports per drug injector}}$$

## Out-reports: Connections to hidden population members

How many people do you know who inject drugs?

## Visibility: Number of in-reports per drug injector

$$\text{number of drug injectors} = \frac{\text{total out-reports}}{\text{in-reports per drug injector}}$$

# Visibility: Number of in-reports per drug injector

Lots of potential strategies for estimating visibility.

- ▶ basic scale-up
- ▶ generalized scale-up

## Visibility: Number of in-reports per drug injector

Use survey respondents' personal networks to approximate hidden population visibility.

## Visibility: Number of in-reports per drug injector

Use survey respondents' personal networks to approximate hidden population visibility.

Basic network scale-up:

- ▶ assume reports are the same as underlying network
- ▶ assume the average personal network size of drug injectors is the same as the general population
- ▶ assume respondents are perfectly aware of who is a drug injector

## Visibility: Number of in-reports per drug injector

Use survey respondents' personal networks to approximate hidden population visibility.

Basic network scale-up:

- ▶ assume reports are the same as underlying network
- ▶ assume the average personal network size of drug injectors is the same as the general population
- ▶ assume respondents are perfectly aware of who is a drug injector

For example, if our survey results tell us that adults in the general population have an average network size of 200

... then we assume that drug injectors have an average visibility of 200.

## Estimating visibility

To estimate network size, we ask questions about ties to populations of **known** size (Killworth et al, 1998).

# Estimating visibility

Suppose that there are

50 thousand people named Nsabimana in Rwanda,

# Estimating visibility

Suppose that there are

50 thousand people named Nsabimana in Rwanda,

and a respondent reports having

connections to 2 people named Nsabimana.

## Estimating visibility

Suppose that there are  
50 thousand people named Nsabimana in Rwanda,

and a respondent reports having  
connections to 2 people named Nsabimana.

Then we could estimate the respondent's network size with:

$$\begin{aligned}\hat{d} &= \frac{\overbrace{\text{number of connections to Nsabimanas}}^{\text{proportion of Nsabimanas respondent is connected to}}}{\text{total number of Nsabimanas in Rwanda}} \times \text{size of Rw's pop.} \\ &= \frac{2}{50,000} \times 10,000,000 \\ &= 400 \text{ people}\end{aligned}$$

## Estimating visibility

In practice, we ask about many known populations – usually about 20 – to get a better estimate:

$$\hat{d}_i = \frac{\sum_j y_{ij}}{\sum_j N_j} \cdot N$$

- ▶  $\hat{d}_i$  - the estimate of respondent  $i$ 's network size
- ▶  $\hat{y}_{ij}$  - the number of ties that respondent  $i$  reports to members of subpopulation  $j$
- ▶  $N_j$  - the size of subpopulation  $j$
- ▶  $N$  - the size of the entire population

Feehan and Salganik (2016) has the precise conditions that need to hold for this to produce unbiased estimates.

# Data: household survey in Rwanda



Map source: Wikipedia

## Data: household survey in Rwanda

- ▶ Intended to mimic a Demographic and Health Survey
- ▶ Stratified, two-stage cluster sample of approximately 5,000 Rwandans aged 15 and older (oversampled Kigali)

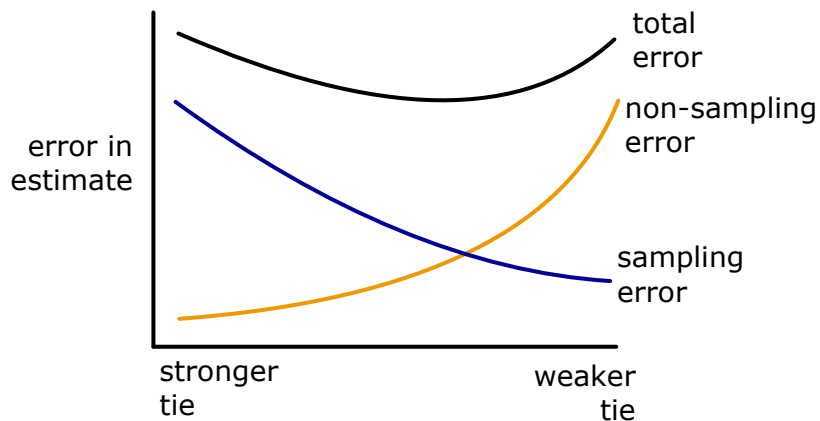
# Tie definition - survey experiment

## Acquaintance ( $n = 2,236$ )

- ▶ people of all ages who live in Rwanda
- ▶ people the respondent knows, by sight AND name, and who also know the respondent by sight and name
- ▶ people the respondent has had some contact with – either in person, over the phone, or on the computer in the previous 12 months

## Meal ( $n = 2,433$ )

# Framework for tie definitions



# Tie definition - survey experiment

## Acquaintance ( $n = 2,236$ )

- ▶ people of all ages who live in Rwanda
- ▶ people the respondent knows, by sight AND name, and who also know the respondent by sight and name
- ▶ people the respondent has had some contact with – either in person, over the phone, or on the computer in the previous 12 months

## Meal ( $n = 2,433$ )

# Tie definition - survey experiment

## Acquaintance ( $n = 2,236$ )

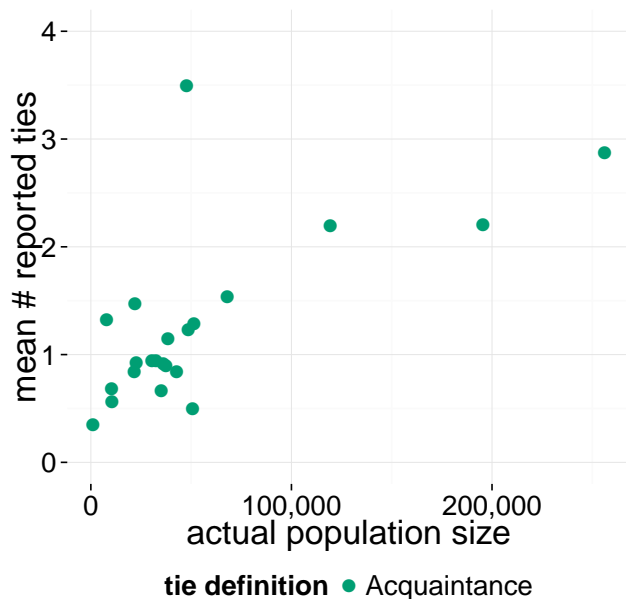
- ▶ people of all ages who live in Rwanda
- ▶ people the respondent knows, by sight AND name, and who also know the respondent by sight and name
- ▶ people the respondent has had some contact with – either in person, over the phone, or on the computer in the previous 12 months

## Meal ( $n = 2,433$ )

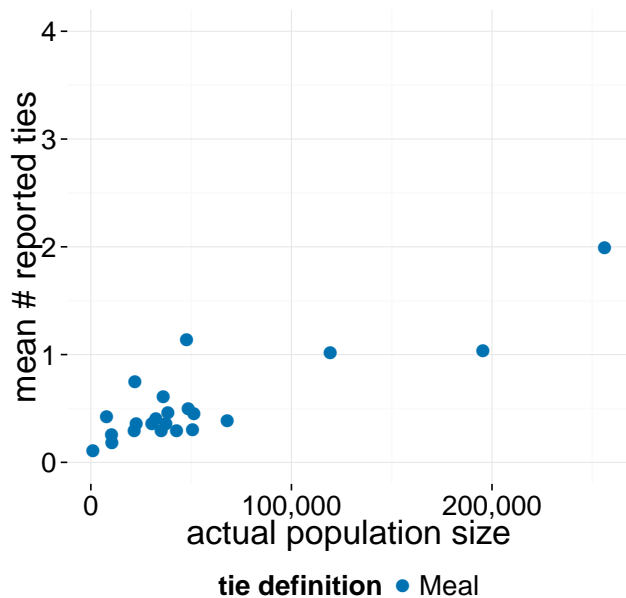
- ▶ people of all ages who live in Rwanda
- ▶ people the respondent knows, by sight AND name, and who also know the respondent by sight and name
- ▶ people the respondent has shared a meal or drink with in the past 12 months, including family members, friends, co-workers, or neighbors, as well as meals or drinks taken at any location, such as at home, at work, or in a restaurant.

# Reported connections to known populations

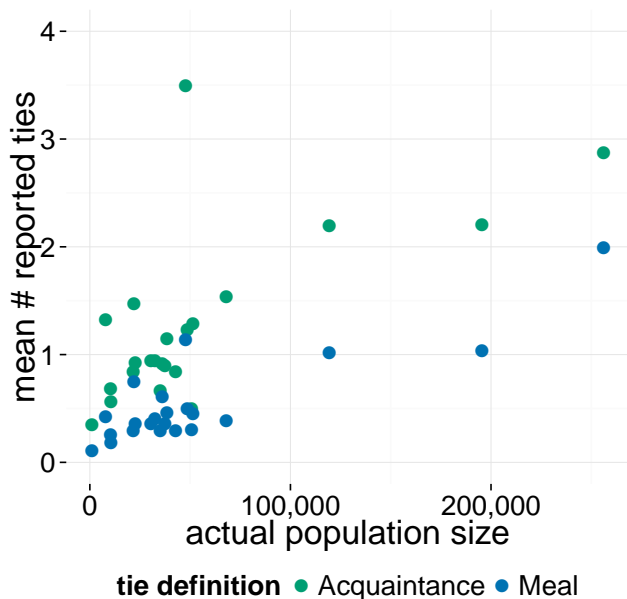
## Reported connections to known populations



## Reported connections to known populations



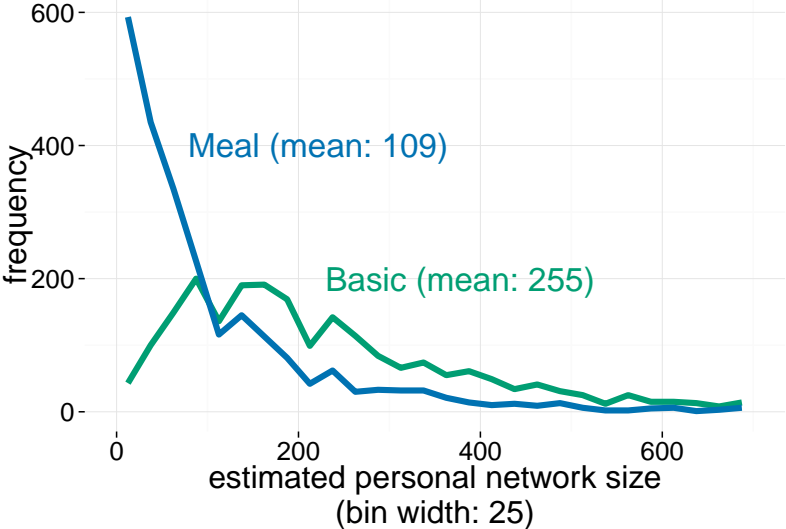
## Reported connections to known populations



# Estimated network size distributions

# Estimated network size distributions

## Distribution of estimated personal network size



# Internal consistency

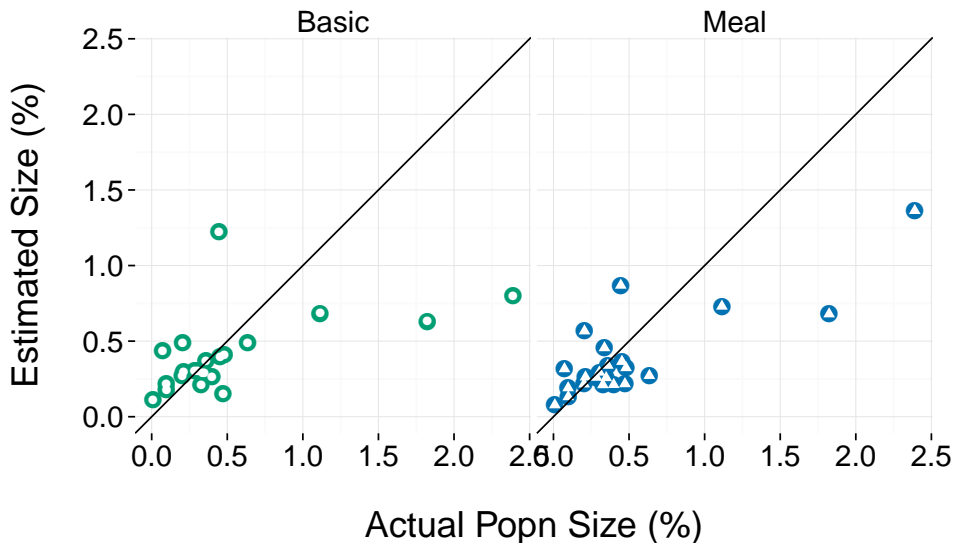
# Internal consistency

Internal consistency:

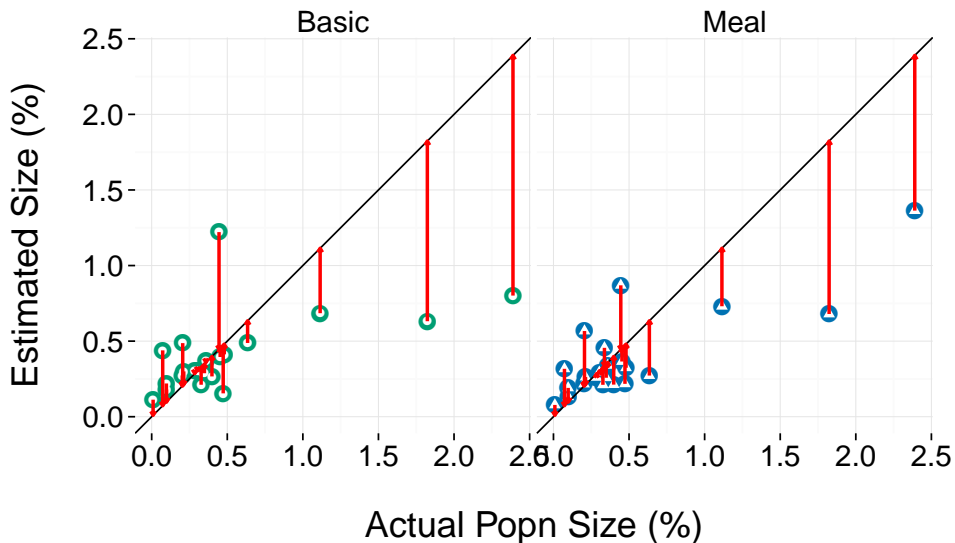
- ▶ pretend that we don't know how many Nsabimanas there are in Rwanda
- ▶ estimate network sizes using all of the known populations *except* Nsabimanas
- ▶ use the number of connections reported to Nsabimanas to estimate the total number of Nsabimanas
- ▶ compare our estimate to the known value

# Internal consistency

## Internal consistency

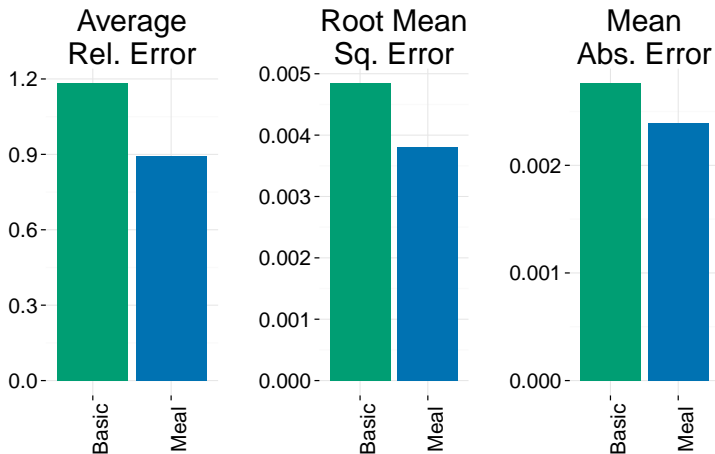


## Internal consistency



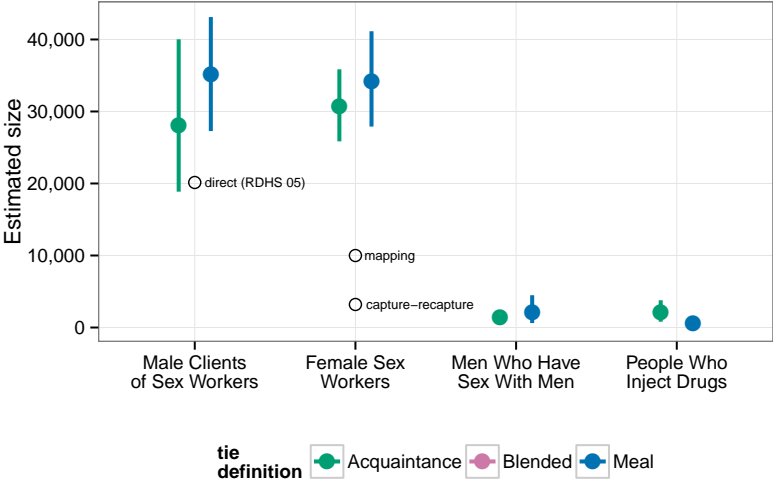
# Tie definition accuracy

# Tie definition accuracy



# Rwanda: estimates

# Rwanda: estimates



# Framework for sensitivity analysis

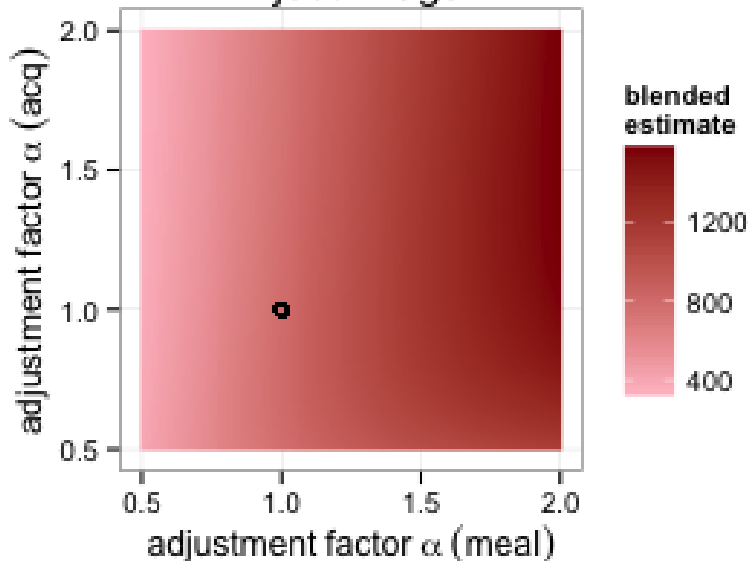
# Framework for sensitivity analysis

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

adjustment factors



## People Who Inject Drugs



# Thanks!

- ▶ thanks to my collaborators: Matthew J. Salganik (Princeton), Mary Mahy (UNAIDS), Aline Umubyeyi (U. of Rwanda), Wolfgang Hladik (CDC)
- ▶ thanks to funders: UNAIDS, USAID, NIH

# Thanks!

- ▶ Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. (1989). Estimating the size of an average personal network and of an event subpopulation. In Kochen, M., editor, *The Small World*, pages 159-175. Ablex Publishing.
- ▶ Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A., and Johnsen, E. C. (1998b). Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review*, 22(2):289-308.
- ▶ Salganik, M. J., Mello, M. B., Abdo, A. H., Bertoni, N., Fazito, D., and Bastos, F. I. (2011b). The game of contacts: Estimating the social visibility of groups. *Social Networks*, 33(1):70-78.

# Thanks!

- ▶ R package, networkreporting, is available on CRAN
- ▶ Feehan and Salganik “Generalizing the network scaleup method,” *Sociological Methodology* (in press).
- ▶ Feehan, Umubyeyi, Mahy, Hladik, and Salganik “Quantity vs quality: a survey experiment to improve the network scale-up method,” *American Journal of Epidemiology* (in press).
- ▶ Feehan, Mahy, and Salganik “The network survival estimator for adult mortality: evidence from Rwanda” (working paper)
- ▶ Feehan (2015) “Network Reporting Methods” (my dissertation)
- ▶ see <http://www.dennisfeehan.org> for more information

