Class 13: Respondent-driven sampling

Matthew J. Salganik

Sociology 204: Social Networks
Princeton University

2/3 Estimation

Respondent-driven sampling describes both:

- a method of data collection
- a method of estimation

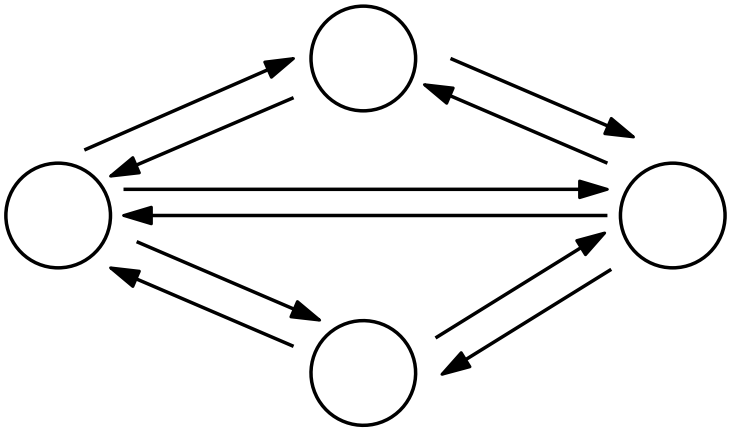
To summarize years of work:
Under certain assumptions about the recruitment process, participants are selected with probability proportional to their degree. For example, someone with 10 friends in the hidden population is twice as likely to be selected as someone with 5 friends.

Estimation: Assumptions

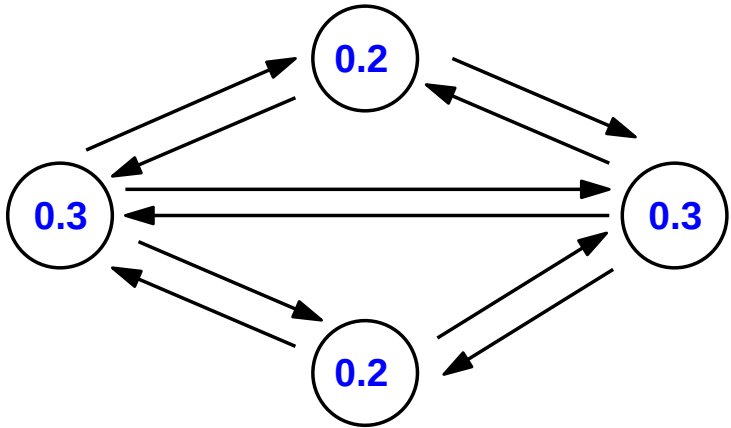
4 key assumptions (but really 3)

- Population forms one connected component and ties are reciprocal
- Sampling with replacement
- People recruit randomly from their friends
- Seed selected with probability proportion to their degree
  - This assumption can be relaxed if the sample size is "large"

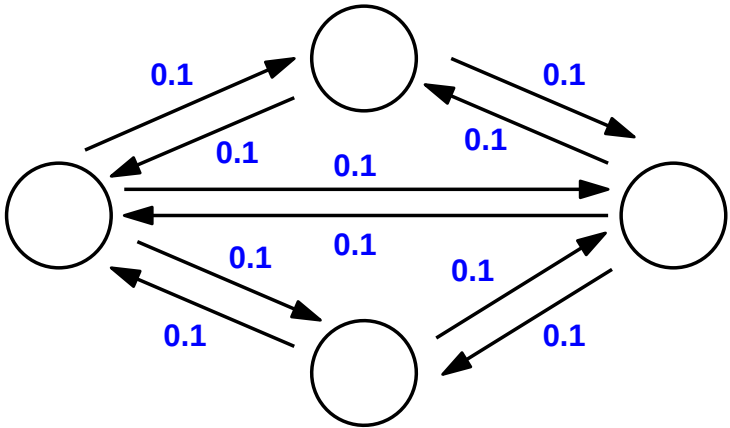
Estimation: Consequences of assumptions



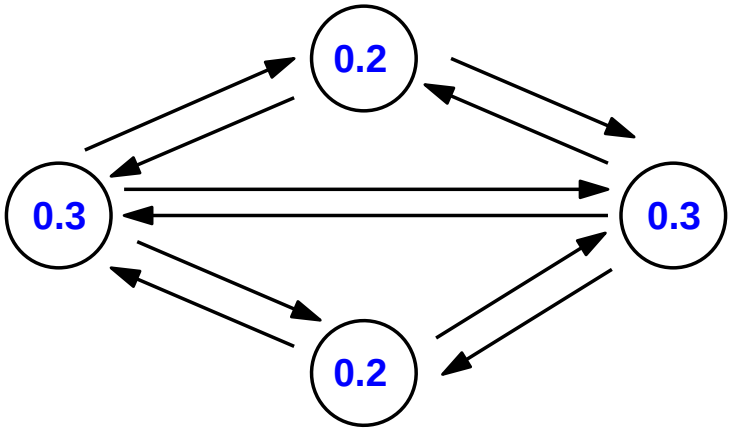
Estimation: Consequences of assumptions



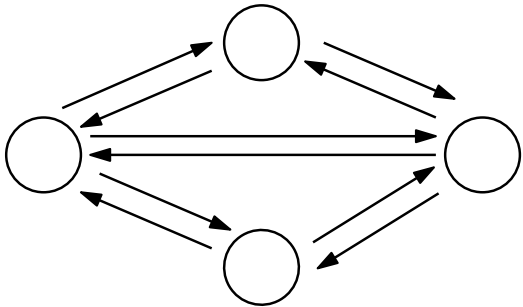
Estimation: Consequences of assumptions



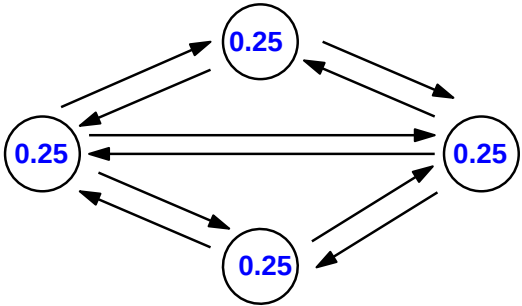
Estimation: Consequences of assumptions



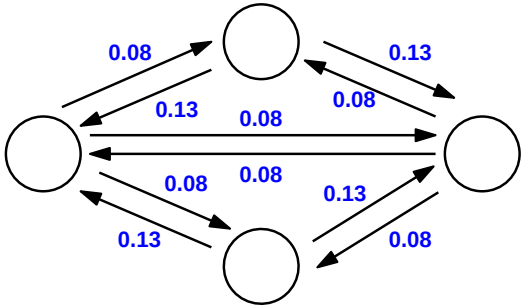
Seeds not drawn correctly



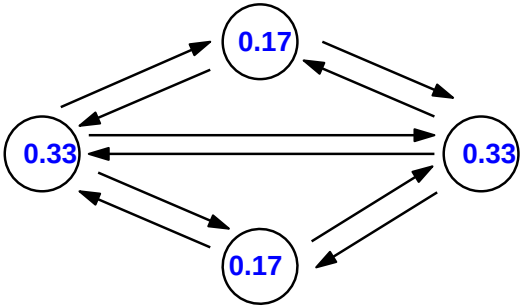
Seeds not drawn correctly



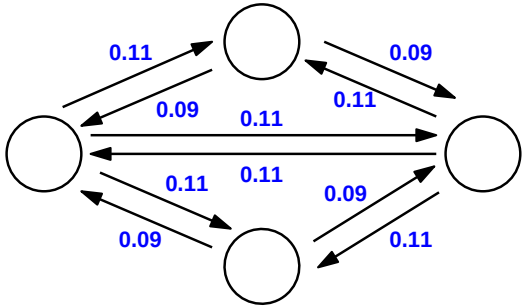
Seeds not drawn correctly



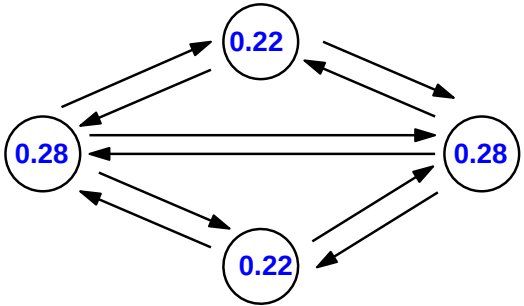
Seeds not drawn correctly



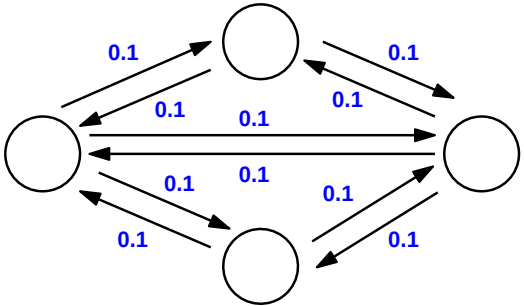
Seeds not drawn correctly



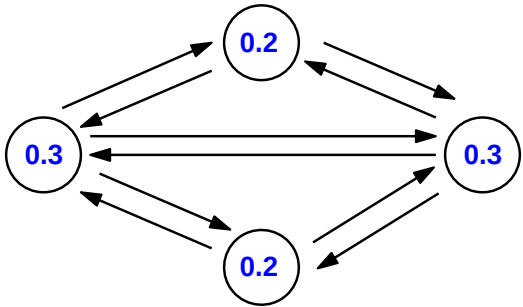
Seeds not drawn correctly



Seeds not drawn correctly



Seeds not drawn correctly



If people are selected with probability proportional to degree (and we can measure their degree), then we can weight people by the inverse of their degree.

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{n} y_i/d_i}{\sum_{i=1}^{n} 1/d_i}$$

where

- $\hat{\bar{y}}$: estimated average infection rate
- y_i: infection status of person i
- d_i: degree of person i

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{n} y_i/d_i}{\sum_{i=1}^{n} 1/d_i}$$

Infected (y_i)	Degree (d_i)
0	10
1	20
1	35
0	5
0	10
1	30
0	5
1	40
0	10
1	20

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{n} y_i/d_i}{\sum_{i=1}^{n} 1/d_i}$$

Infected (y_i)	Degree (d_i)
0	10
1	20
1	35
0	5
0	10
1	30
0	5
1	40
0	10
1	20

- sample mean $= 0.5$

$$\hat{\hat{y}} = \frac{\sum_{i=1}^{n} y_i/d_i}{\sum_{i=1}^{n} 1/d_i}$$

Infected (y_i)	Degree (d_i)
0	10
1	20
1	35
0	5
0	10
1	30
0	5
1	40
0	10
1	20

- sample mean $= 0.5$
- $\hat{\hat{y}} = \frac{0+1/20+1/35+0+0+1/30+0+1/40+0+1/20}{1/10+1/20+1/35+1/5+1/10+1/30+1/5+1/40+1/10+1/20} \approx 0.19/0.89 \approx 0.21$

But . . .

To summarize years of work:
The structure of real social networks may be unfavorable to RDS. In particular, it may lead to estimates with high sample-to-sample variability, which makes the estimates less useful.

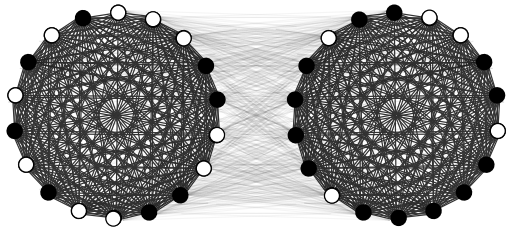
RDS as MCMC Importance Sampling

Goel and Salganik (2009) introduced a connection between respondent-driven sampling and Markov Chain Monte Carlo importance sampling. By making this connection we can establish two things:

- Community structure (e.g., cohesive subgroups) in the social network increases the variance of RDS estimates. In particular, "bottlenecks" anywhere in the network may degrade estimates—bottlenecks need not be directly related to the characteristics being studied.
- A design that incorporates multiple recruitment increases the variance of RDS estimates.

An Example: Bottlenecks

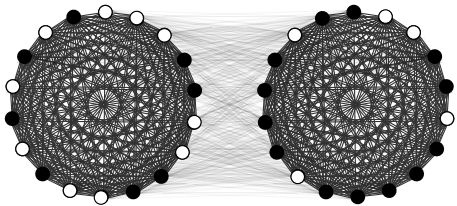
Consider a population consisting of two equal-sized groups where *within-group* edges are more likely than *between-group* edges. For example, street-based and agency-based sex workers in Belgrade (Simic, et al 2007).



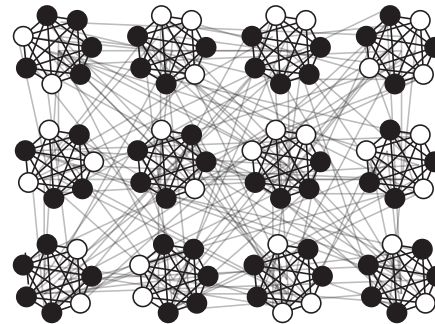
White nodes are infected; black nodes are healthy.

Focusing solely on the connection between healthy and infected overlooks the key structural feature of this network. Bottlenecks need not be directly related to the characteristic being studied.

An Example: Bottlenecks



(a)

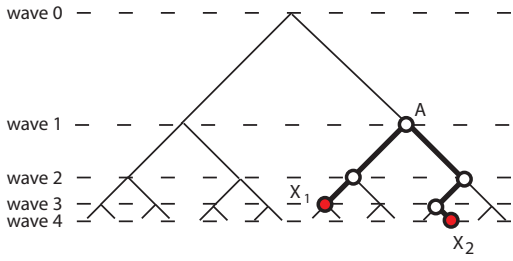


(b)

More subtle bottlenecks can cause the same problem. These two networks are the same in terms of RDS variance.

An Example: Multiple recruitment

Multiple recruitment is needed for the sampling process to continue in practice, but it increases the variance of the estimates by increasing the dependence between sample observations.



The effective sample size is smaller than the observed sample size. In other words, you have less information than you think.

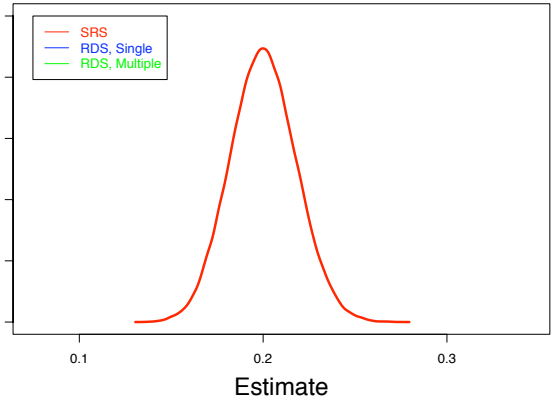
Toy example

To summarize our findings for this hypothetical population, we compare three sampling situations: simple random sampling, RDS with single recruitment, and RDS with multiple recruitment. We use parameters $p_A = .1$, $p_B = .3$, $c = .1$, sample size $n = 500$, and 2 seeds chosen independently from the stationary distribution (more general results in our paper). Multiple recruitment is based on a branching process with offspring distribution:

	Number of recruits			
	0	1	2	3
Probability	1/3	1/6	1/6	1/3

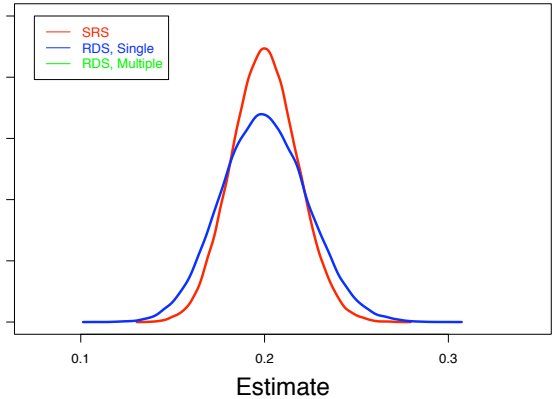
This recruitment distribution is based on RDS data from the Frost, et. al. (2006) study of drug-injectors in Tijuana and Ciudad Juarez.

Comparing sampling schemes (n=500)



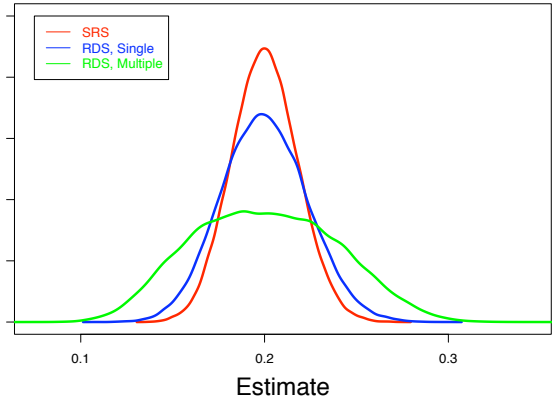
	effective sample size
Simple Random Sampling	500

Comparing sampling schemes (n=500)



	effective sample size
Simple Random Sampling	500
RDS – Single Recruitment	335

Comparing sampling schemes (n=500)



	effective sample size
Simple Random Sampling	500
RDS – Single Recruitment	335
RDS – Multiple Recruitment	136

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2202–2229
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3613

Respondent-driven sampling as Markov chain Monte Carlo

Sharad Goel[1,‡] and Matthew J. Salganik[2,*,†]

[1]*Yahoo! Research, 111 W. 40th Street, New York, NY 10018, U.S.A.*
[2]*Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ 08544, U.S.A.*

https://doi.org/10.1002/sim.3613

What about real networks?

Hopefully the toy example helps to build intuition. But what about in real networks? We quickly run into a data problem:

- Complete social network data for large networks
- Demographic information for each node

What about real networks?

Hopefully the toy example helps to build intuition. But what about in real networks? We quickly run into a data problem:

- Complete social network data for large networks
- Demographic information for each node

We have found 2 such sources:

- Add Health
- Project 90

A note on these simulations

The following simulations will assume that all RDS assumptions are met. This represents a useful case to examine, but of course, many assumptions will not be met in practice.

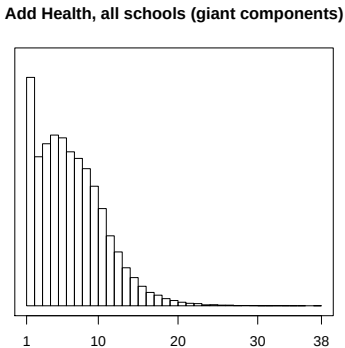
- sample size of 500
- 10 seeds chosen with probability proportional to degree (i.e., from stationary distribution)
- Offspring distribution based on Frost et al. (2006):

	Number of recruits			
	0	1	2	3
Probability	1/3	1/6	1/6	1/3

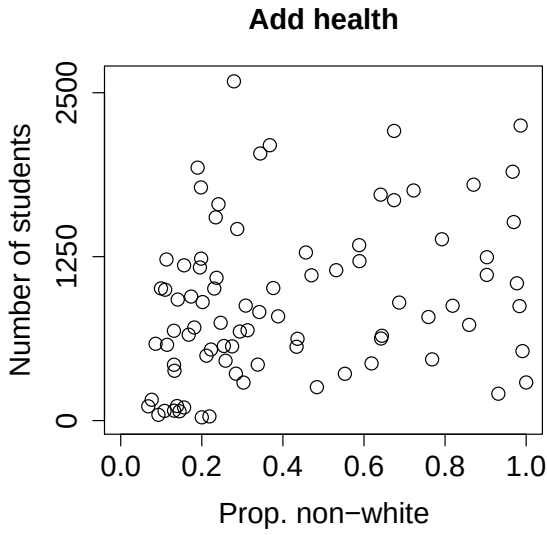
Add Health

National Longitudinal Study of Adolescent Health (Add Health), Source: UNC-Pop. Center

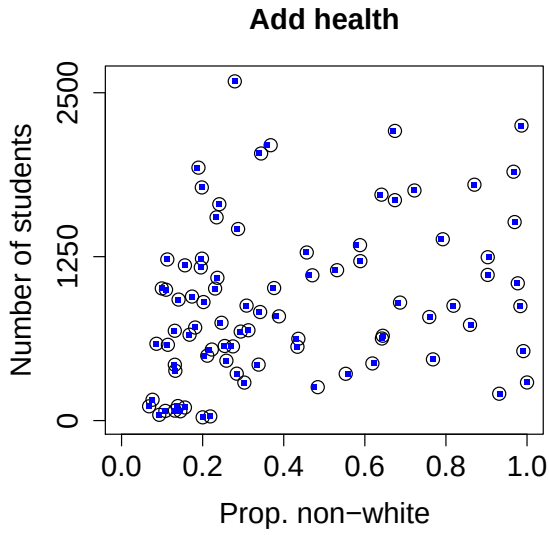
- 84 schools with giant components ranging from 25 to 2,539 (mean=860).
- Respondents were asked to chose from a roster up to 5 male friends and up to 5 female friends. All responses were symmetrized and degrees range from 1 to 38 (mean=7).



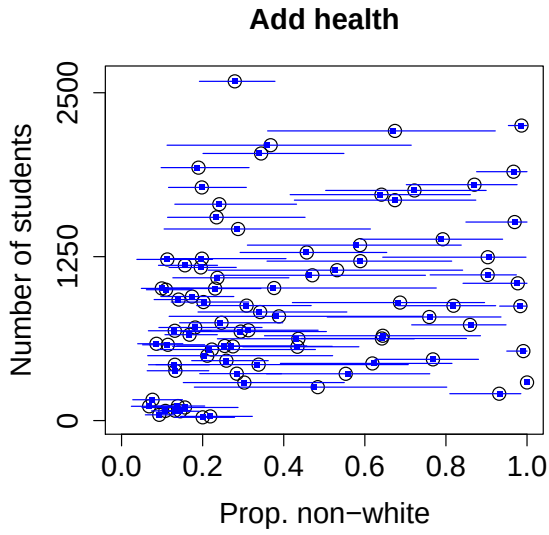
Add Health: Estimates, prop. non-white



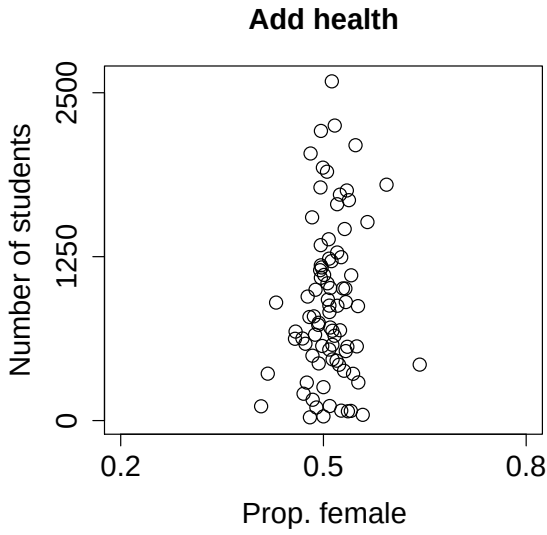
Add Health: Estimates, prop. non-white



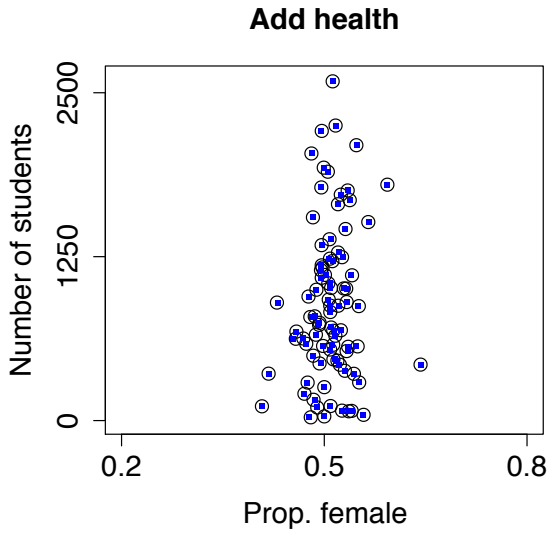
Add Health: Estimates, prop. non-white



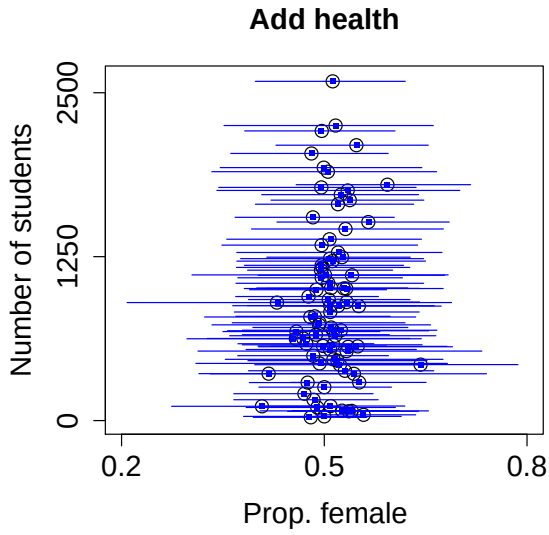
Add Health: Estimates, prop. female



Add Health: Estimates, prop. female



Add Health: Estimates, prop. female



Design effect

Quantify variability with design effect: $deff = \frac{var(\hat{p}_{RDS})}{var(\hat{p}_{SRS})}$

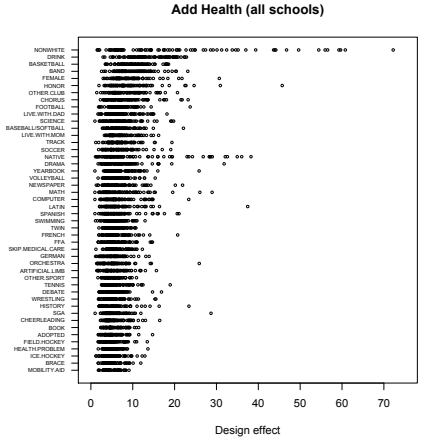
- Can be thought of as a measure of the "relative efficiency" of the sampling methods. Values bigger than 1 mean that RDS is less efficient than simple random sampling.

Design effect

Quantify variability with design effect: $deff = \frac{var(\hat{p}_{RDS})}{var(\hat{p}_{SRS})}$

- Can be thought of as a measure of the "relative efficiency" of the sampling methods. Values bigger than 1 mean that RDS is less efficient than simple random sampling.
- Note that design effect is a property the estimator for a trait in a population, not just a population. For example, the design effect for prop. female may be different from design effect for prop. non-white.

Add Health: Design effects



The median design effect is about 5 and socially salient characteristics are even higher.

Add Health: Design effects

Design effect of 5 means:

- RDS sample of 500 have same variance as a simple random sample of size 100.
- RDS confidence intervals should be about 2 times ($\sqrt{5}$) wider.
- It will be very difficult to reliably detect change over time (from 40% to 30% requires $n = 1750$)

Limitations of Add Health data

- Some schools were very small.

Limitations of Add Health data

- Some schools were very small.
- Fixed choice design limited out-degree.

Limitations of Add Health data

- Some schools were very small.
- Fixed choice design limited out-degree.
- Networks of high school students might be different from networks of drug injectors, sex workers, or other hidden populations.

An Overview of Project 90

Project 90 was a multi-year study that mapped the connections between sex workers, drug injectors, and their sexual partners, beginning in Colorado Springs from 1988-1992.

An Overview of Project 90

Project 90 was a multi-year study that mapped the connections between sex workers, drug injectors, and their sexual partners, beginning in Colorado Springs from 1988-1992.

- The entire Project 90 network contains 5,492 individuals and 21,644 edges, representing social, sexual, and/or drug affiliation.

An Overview of Project 90

Project 90 was a multi-year study that mapped the connections between sex workers, drug injectors, and their sexual partners, beginning in Colorado Springs from 1988-1992.

- The entire Project 90 network contains 5,492 individuals and 21,644 edges, representing social, sexual, and/or drug affiliation.
- We restrict attention to the giant component of this network, consisting of 4,430 nodes and 18,407 edges.

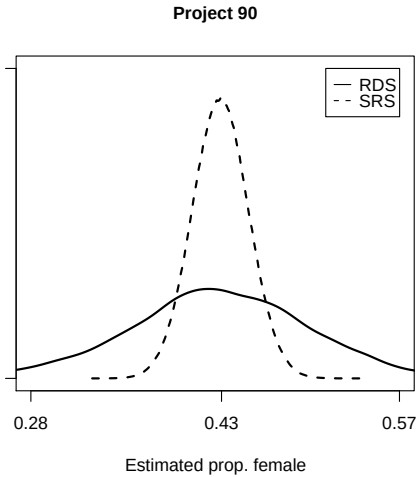
An Overview of Project 90

Project 90 was a multi-year study that mapped the connections between sex workers, drug injectors, and their sexual partners, beginning in Colorado Springs from 1988-1992.

- The entire Project 90 network contains 5,492 individuals and 21,644 edges, representing social, sexual, and/or drug affiliation.
- We restrict attention to the giant component of this network, consisting of 4,430 nodes and 18,407 edges.
- The median degree of an individual in the giant component is 6, with a degree range of 1 to 159.

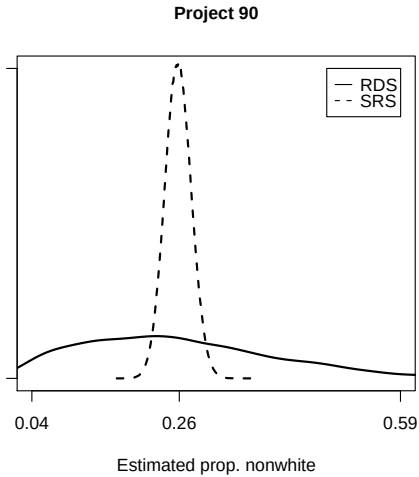
We thank the Project 90 Team, especially Steve Muth and John Potterat, for sharing the data.

Project 90, prop. female



Estimating of prop. female are unbiased, but the design effect is about 11.

Project 90, prop. non-white



Estimates of prop. non-white are unbiased, but the design effect is about 57.

Project 90, Why are the estimates so variable?

There is considerable community structure in this dataset: The giant component partitions into two groups of sizes 1,076 and 3,354 individuals, with few between-group edges (the conductance is 0.13), and such that the percentage of non-Whites in the former group is 55.6% and in the latter 12.2%.

Project 90, Why are the estimates so variable?

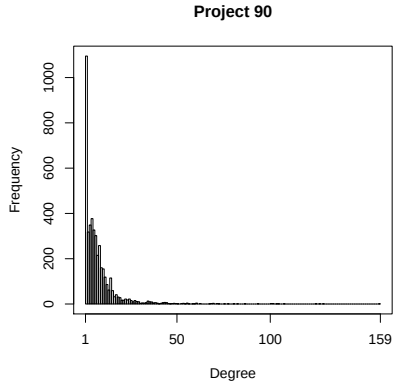
Even worse, the source of this community structure is not clear (to us), so it may be hard to detect just with sample information.
Analogous to our toy example, it is not segregation directly between Whites and non-Whites, for example, that leads to poor RDS estimates. To the contrary, Whites and non-Whites, and men and women, on the whole are well-connected in this network.

This is what you could see in a sample:

	White	Non-White
White	282	63
Non-White	77	68

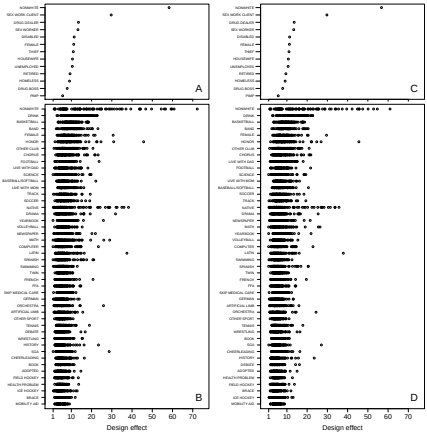
Problems with Project 90

- Missing edges
- Missing nodes
- Many nodes with degree 1 (i.e., very “leafy”)



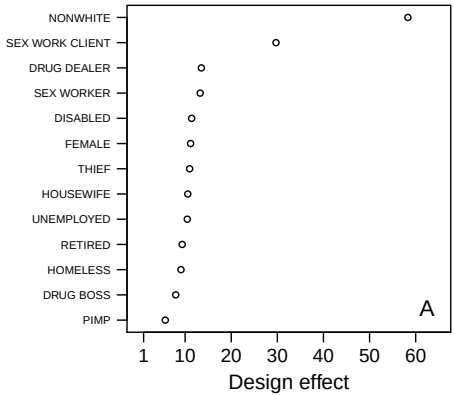
But what about. . . . . ?
We can address these questions with robustness checks.

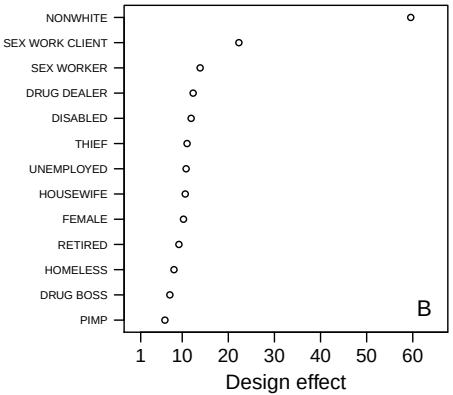
Robustness checks: RDS I and RDS II



$\rightarrow$ Similar results for RDS I and RDS II estimators.

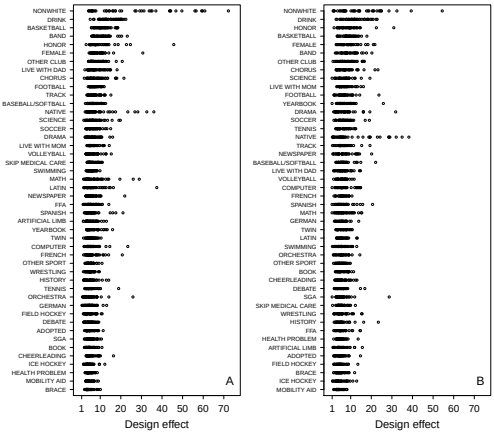
Robustness checks: Leaves in Project 90





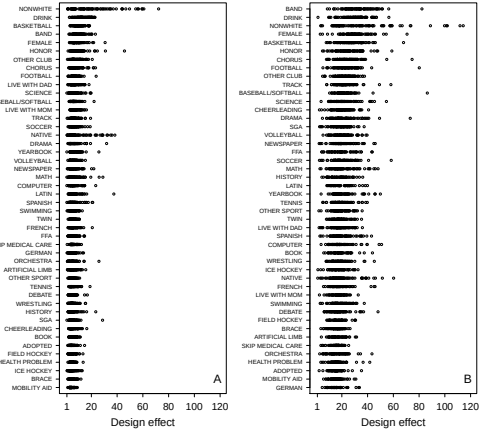
$\rightarrow$ Similar results with and without leaves.

Robustness check: High schools & joint middle/high schools



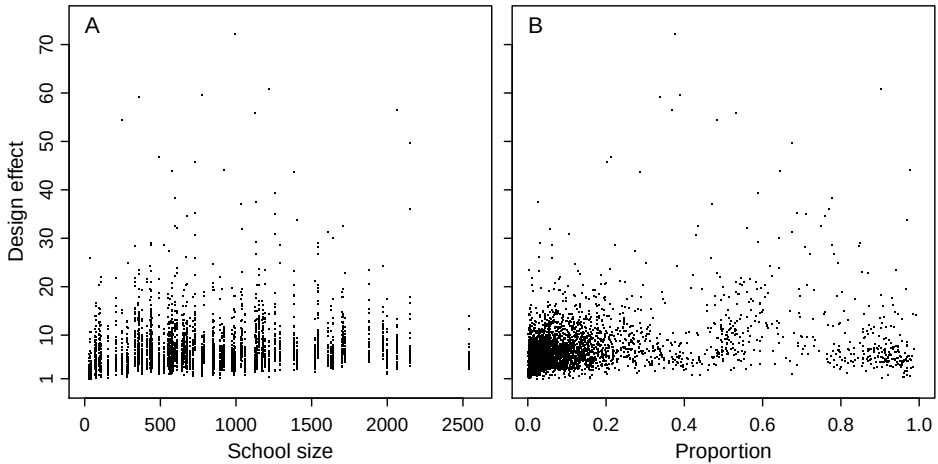
$\rightarrow$ Similar results in high schools and joint middle and high schools.

Robustness check: Reciprocity in constructing network



$\rightarrow$ Median design effect increases to about 19 when using fully reciprocal network.

Robustness check: Not just small schools or rare traits



→ High design effects are not just in small schools or for rare traits.

Robustness checks

Based on:

- Two distinct sources of data
- More than 3,000 network/trait combinations (84 schools $\times$ 46 traits $+$ 1 Project 90 $\times$ 13 traits)
- Several robustness checks

it seems that:

- RDS estimates are much less precise than previously believed and this result seems pretty robust.

Robustness checks

Based on:

- Two distinct sources of data
- More than 3,000 network/trait combinations (84 schools $\times$ 46 traits + 1 Project 90 $\times$ 13 traits)
- Several robustness checks

it seems that:

- RDS estimates are much less precise than previously believed and this result seems pretty robust.
- Estimates are so imprecise that it seems difficult to reliably detect change over time. For example, if we assume a design effect of 5, to reliably detect a drop in risk behavior from 40% prevalence to 30% prevalence ($\alpha = 0.05, \beta = 0.8$) would require samples of about 1750 at both time points.

More details

Assessing respondent-driven sampling

Sharad Goel[a,1] **and Matthew J. Salganik**[b,1]

[a]Microeconomics and Social Systems, Yahoo! Research, 111 West 40th Street, New York, NY, 10018; and [b]Department of Sociology and Office of Population Research, Princeton University, Wallace Hall, Princeton, NJ 08544

https://doi.org/10.1073/pnas.1000261107

How can we improve respondent-driven sampling?