Respondent-Driven Sampling as Markov Chain Monte Carlo*

Sharad Goel[1] and Matthew J. Salganik[2,3]
1: Yahoo! Research
2: Department of Sociology, Princeton University
3: Office of Population Research, Princeton University

UCSF, Center for AIDS Prevention Studies
October 30, 2007

*We thank Andrew Gelman, Doug Heckathorn, Steve Muth and Erik Volz for helpful conversations and comments. This research was supported by the National Science Foundation, the Institute for Social and Economic Research and Policy at Columbia University, and the Applied Statistics Center at Columbia University.

Some motivating questions

- What percentage of drug injectors in San Francisco have HIV?
- What is the average age of sex workers in Moscow?

Hidden populations

These "hidden" populations are hard to sample because:

- No sampling frame
- Small proportion of the general population
- In some cases, desire to remain anonymous

Examples include: drug injectors, sex workers, men who have sex with men, undocumented workers, jazz musicians, and members of some social movements

Previous approaches to the study of hidden populations

- Institutional sampling
  Clearly not representative
- Targeted sampling (Watters and Biernacki, 1989)
  Better, but hard to interpret because probability of selection is not known
- Time-location sampling
  Even better, but not useful for all hidden populations and very expensive

Another approach: snowball sampling

Instead of thinking of people as atomized units on a sampling frame, think of people as embedded in networks. Friends recruit friends and the sample progresses through the social network. But, conventional wisdom was that it was not possible to make unbiased estimates from snowball samples because they:

- oversample popular people
- non-independence of observations (people are similar to their friends)
- depended on the choice of seeds

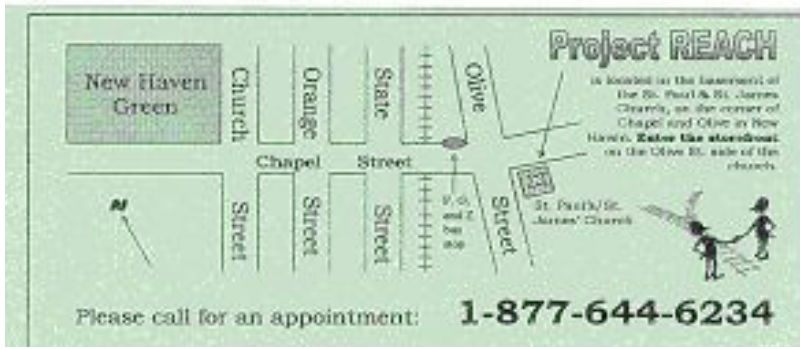
Respondent-driven sampling

It turns out that unbiased estimation is possible under certain general conditions (Salganik and Heckathorn, 2004), and the process is cheaper and faster than existing methods.
Respondent-driven sampling is now being used:

- CDC for studies of drug injectors in the 25 largest U.S. cities (samples of 500 in each city)
- CDC Global AIDS Program to study drug injectors and sex workers in Thailand, Vietnam, Brazil
- Russell Sage Foundation funded a study of undocumented workers in New York, Chicago, and Los Angeles

Sampling

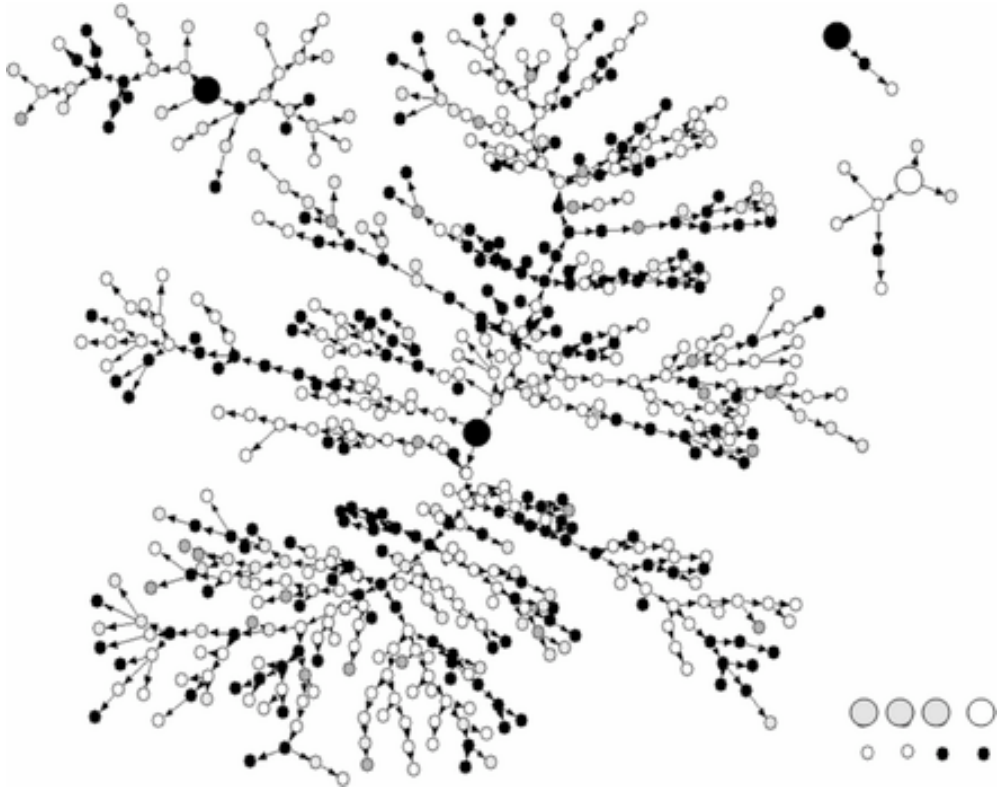
Sample progresses using dual-incentive system (respondents are paid to participant and to recruit others). Participants come to store-front location with coupon (Heckathorn 1997).



Other problems:

- Non-duplication
- Population verification

Sampling



Recruitment network from a study of drug users in New York City (Abdul-Quader et al., 2006)

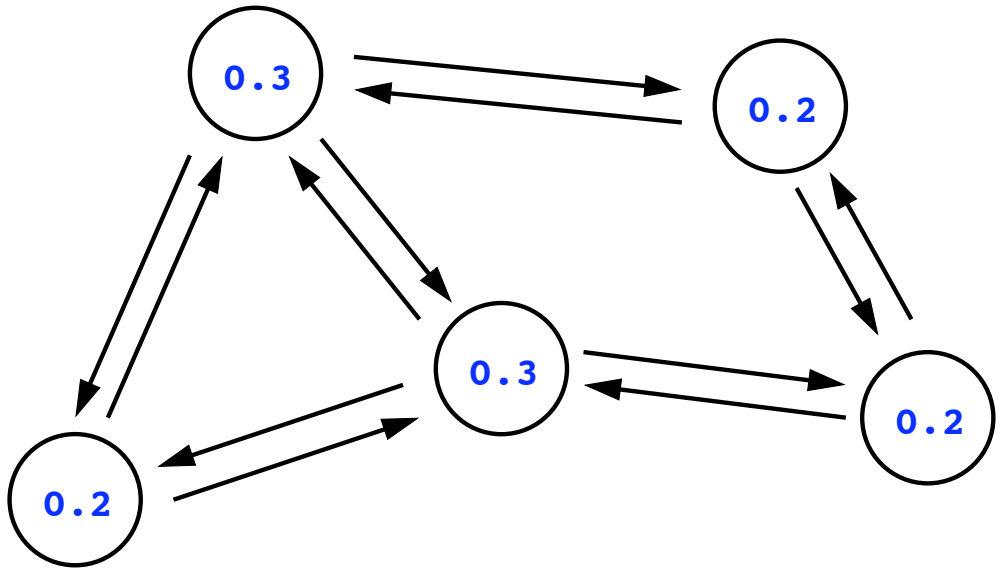
- 8 seeds $\longrightarrow$ 618 drug users
- 13 weeks

Estimation: Assumptions

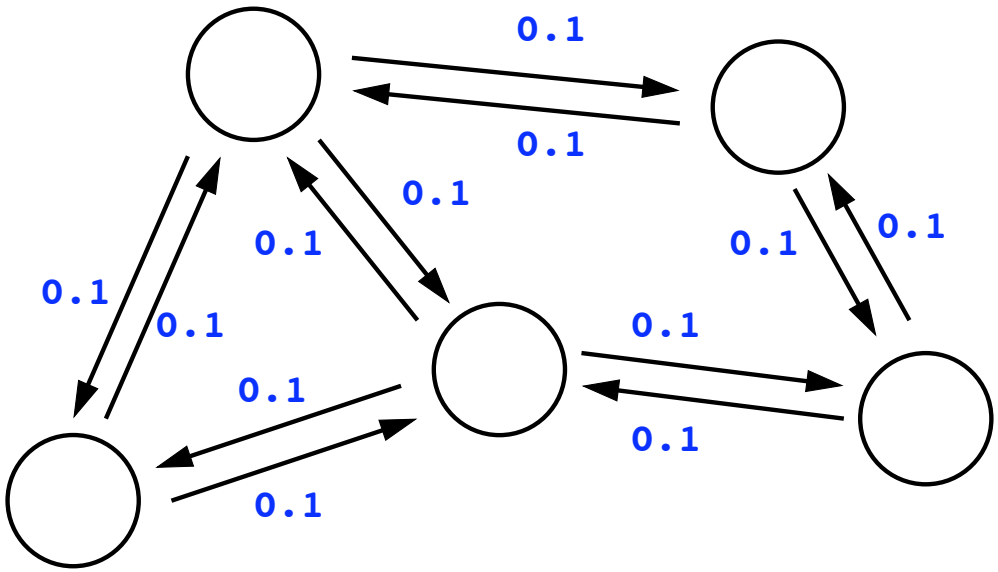
4 key assumptions (but really 3)

- 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*)

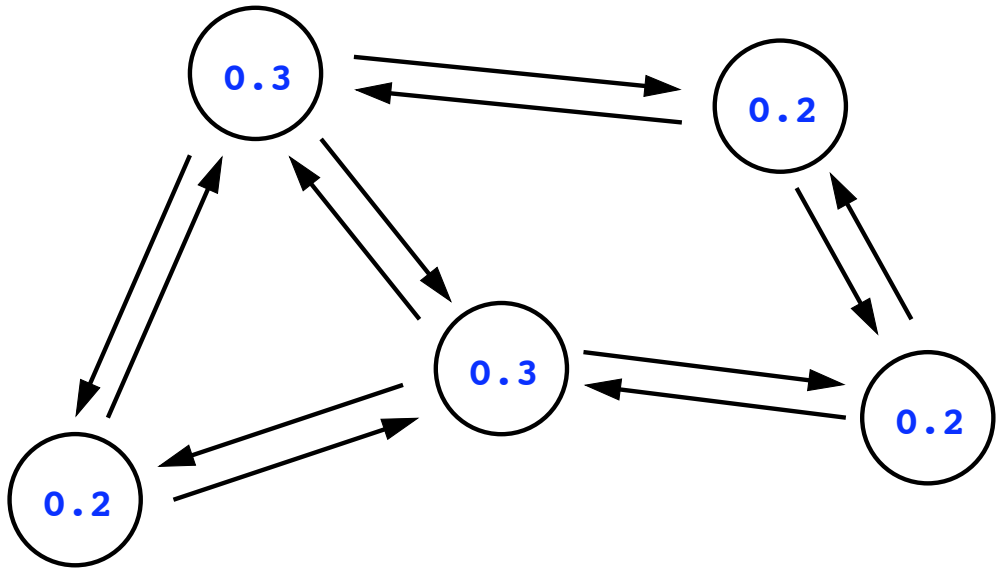
Estimation: Consequences of assumptions



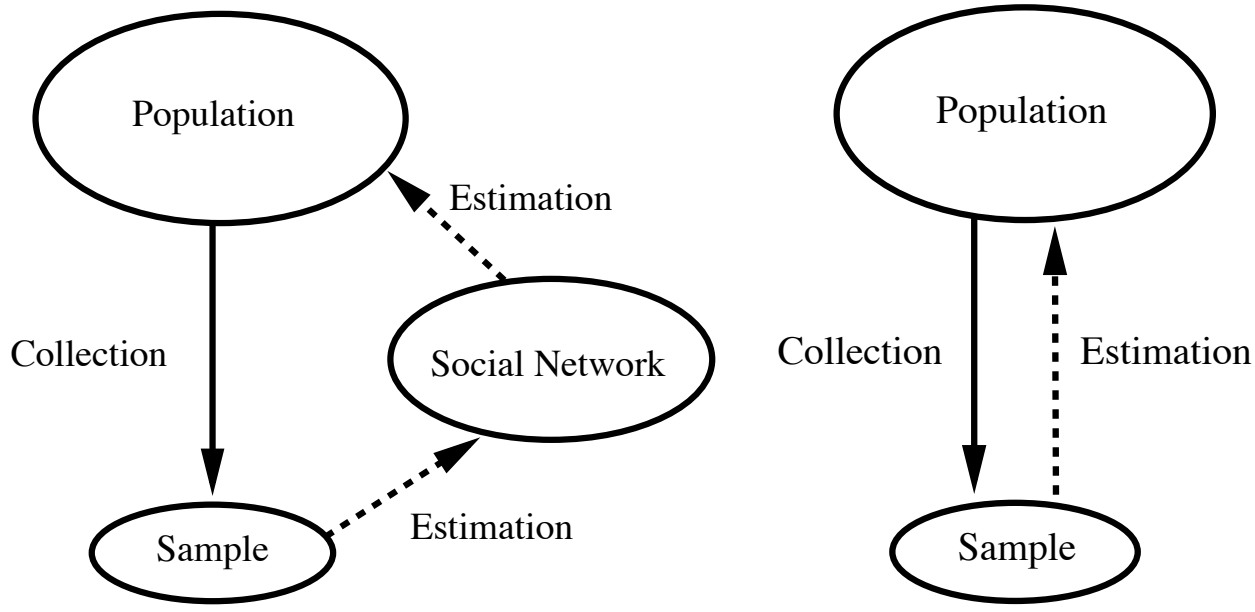
Estimation: Consequences of assumptions



Estimation: Consequences of assumptions



Old approach



(a) Old RDS (b) New RDS

- Estimation based on custom procedure (Salganik and Heckathorn 2004)
- Variance estimation based on modified bootstrap (Salganik 2006). Intuition was that network structure, particularly connections between infected and uninfected group, determine variability of estimates. Note that confidence intervals are only approximate and so p-values are also only approximate.

Respondent-driven sampling and Markov chain Monte Carlo

Previous work hints at the connection between RDS and Markov chain Monte Carlo importance sampling: Salganik & Heckathorn (2004), Volz & Heckathorn (in press), and Thompson (2006). Markov chain Monte Carlo was popularized by the Metropolis algorithm in 1953, and has been extensively applied in physics, chemistry, biology and statistics.

In the context of RDS, a Markov chain on the nodes generates (dependent) samples from a distribution that is proportional to node degree. For example, an individual with twice as many ties is twice as likely to be sampled.

Importance sampling re-weights samples to mimic simple random sampling. Note: samples are still dependent!

RDS as MCMC Importance Sampling

For continuous traits (e.g. age, income), the MCMC estimator is:

$$\hat{\mu} = \frac{1}{\sum_{i=0}^{n-1} 1/\deg(X_i)} \sum_{i=0}^{n-1} f(X_i) \frac{1}{\deg(X_i)}.$$

This was recently introduced in Volz & Heckathorn (in press).

For estimating population proportions (e.g the proportion of infected individuals), this estimator reduces to

$$\hat{p} = \frac{1}{\sum_{i=0}^{n-1} 1/\deg(X_i)} \sum_{X_i \text{ infected}} \frac{1}{\deg(X_i)}.$$

RDS as MCMC Importance Sampling

By establishing this linkage we can extend RDS to estimating continuos variables (age, injection frequency) and contribute two further observations:

- Community structure 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.

RDS Estimators

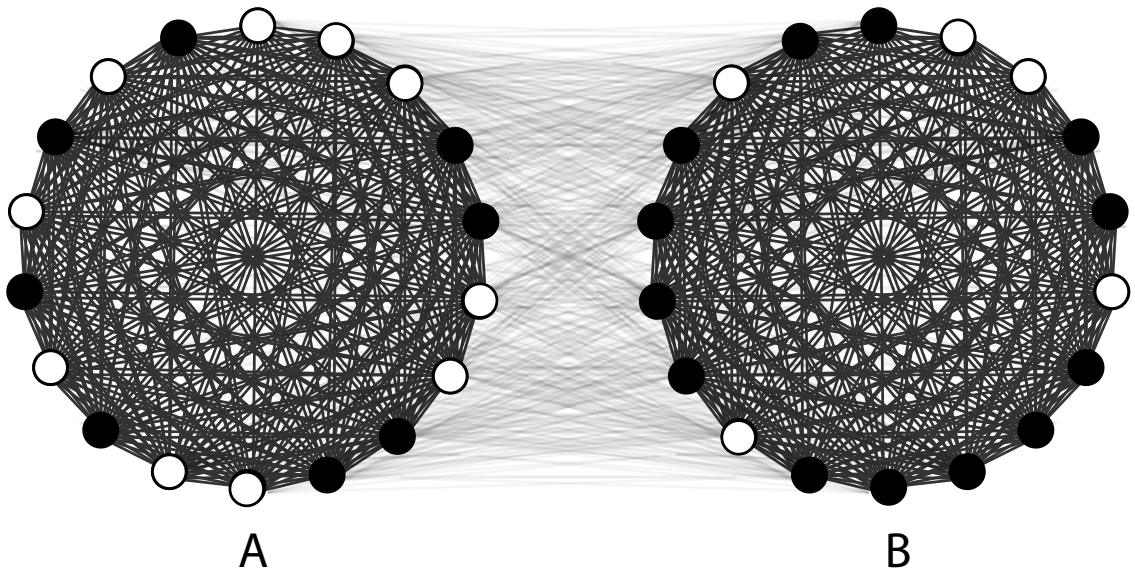
Regardless of the initial distribution of the seeds, the RDS estimators are asymptotically unbiased.

The variance of the estimate $\hat{\mu}$ depends on the variance of f and the autocorrelation structure of the chain, and can be difficult to estimate in practice.

Even though the RDS estimator is asymptotically unbiased, in some cases the variance may be very large.This fact is well understood in the MCMC community, but generally not known among RDS researchers/practioners.

An Example

Consider a population consisting of two equal-sized groups. Edges exist between every pair of individuals, however *within-group* edges have weight $1 - c$ while *between-group* edges have weight c where $0 < c < 1/2$. As c decreases the tendency for within-groups ties becomes stronger.



White nodes are infected; black nodes are healthy.

Conductance

The key parameter in our toy example is the conductance c. It is a measure of the worst bottleneck in the network.

Even though infected and healthy individuals are well connected, the bottleneck between groups A and B is the dominant characteristic of the network.

Focusing solely on infection status overlooks the key structural feature in this network.

An Example – Variance

The samples in an RDS study are dependent. In a segregated population, it is more likely that individuals refer people who are in their same social subgroup, which in turn increases the variance of estimates.

Lemma

Consider the example network, with the seed drawn from the stationary distribution. Then the variance of $\hat{p}$ satisfies

$$Var(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2 \beta_1}{2n(1 - \beta_1)} - \frac{(p_A - p_B)^2 (\beta_1 - \beta_1^{n+1})}{2n^2 (1 - \beta_1)^2}$$

where $\beta_1 = 1 - 2c$ (the second largest eigenvalue of transition matrix) and n is the sample size.

A naive estimate of the variance assumes samples are uncorrelated, yielding only the first term $(p - p^2)/n$. In particular, it does not take into account possible segregation in the hidden population.

An Example – Variance

The simple variance estimate $V(\hat{p}) = (p - p^2)/n$ is related to the true variance by the design effect

$$deff = \frac{\text{Var}(\hat{p})}{(p - p^2)/n} \approx 1 + \frac{(p_A - p_B)^2 \beta_1}{2(p - p^2)(1 - \beta_1)}.$$

In our example, for $c = .1$, $p_A = .3$ and $p_B = .1$, $\text{Var}(\hat{p}) \approx 1.5 \times V(\hat{p})$. Accordingly, confidence intervals determined by the true variance are $\sqrt{1.5} \approx 1.2$ times wider than those suggested by simple random sampling.

Put another way, segregation in our toy network decreases the effective sample size: 500 samples collected via RDS corresponds to 335 independent samples.

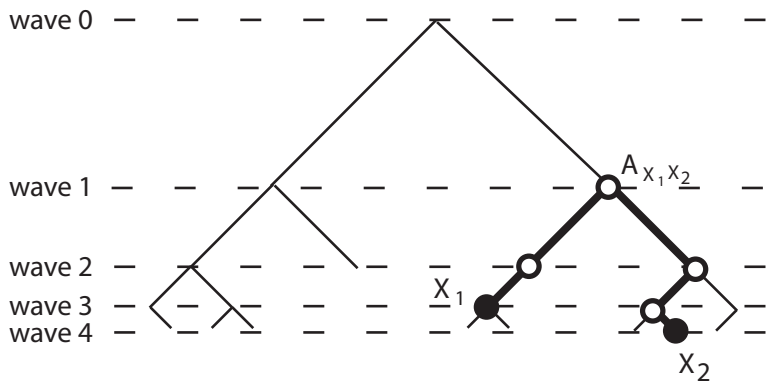
An Example – Multiple Recruitment

We have been assuming that RDS estimates are based on a single, long run of the chain. In practice, this approach is difficult to implement since some sample members do not recruit others, causing the chains to terminate. Instead, in order to ensure that the chains continue, each respondent is allowed to recruit multiple individuals.

With multiple recruitment, samples are the result of relatively short chains, even if total sample size is large. Consequently, there is significant dependence between all samples, increasing the variance of the RDS estimates.

An Example – Multiple Recruitment

The covariance between samples is related to their distance in the recruitment tree and the conductance of the network.



For our toy network, and the recruitment tree above, samples X_1 and X_2 have covariance

$$\mathrm{Cov}(X_1, X_2) = \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{2+3} \quad \text{where } \beta_1 = 1 - 2c.$$

In our example network with parameters $c = .1$, $p_A = .3$ and $p_B = .1$, 500 samples from RDS with multiple recruitment corresponds to approximately 136 independent samples.

An Example – Summary

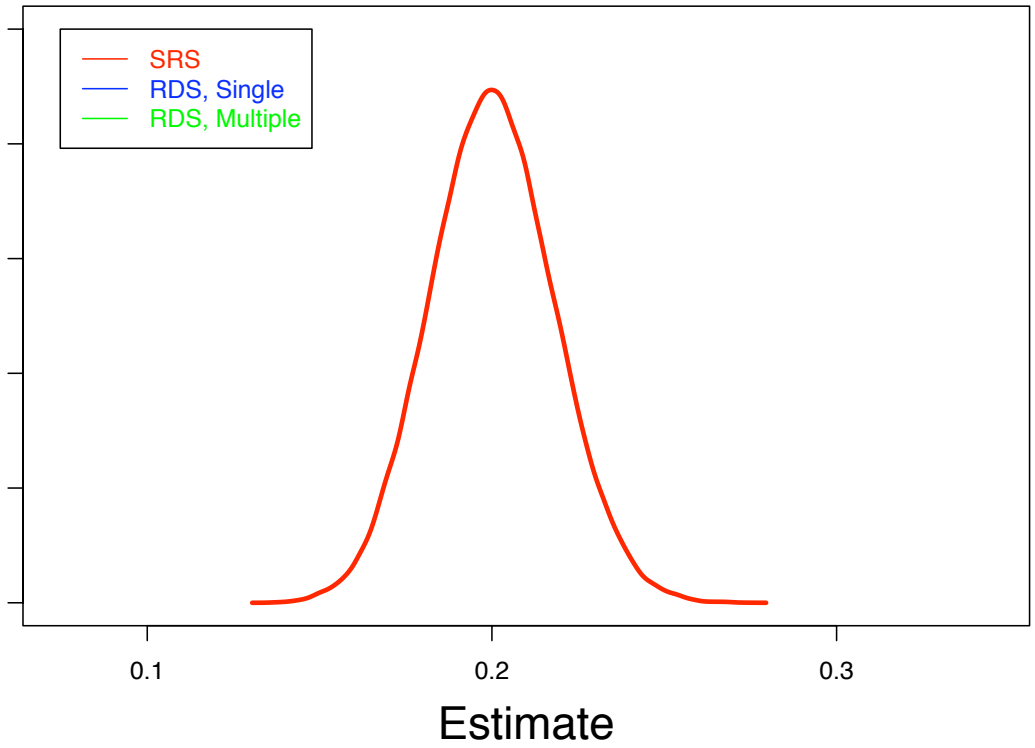
Our toy example shows the effects of both the social network and the recruitment network on RDS estimates in a simplified setting which reasonably maps onto many settings in which RDS may be used.

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. Multiple recruitment is based on a branching process with offspring distribution:

0	1	2	3
1/3	1/6	1/6	1/3

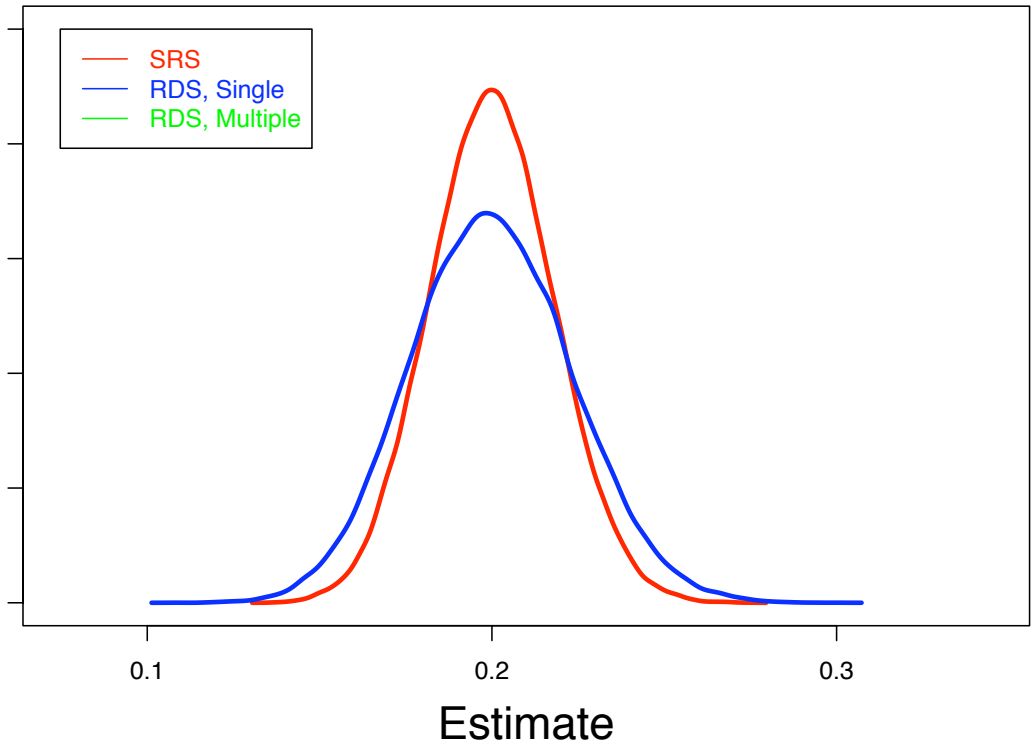
This recruitment distribution is based on RDS data from the Frost, et. al. (2006) study of drug-injectors in Tijuana and Ciudad Juarez.

Comparing sampling schemes (n=500)



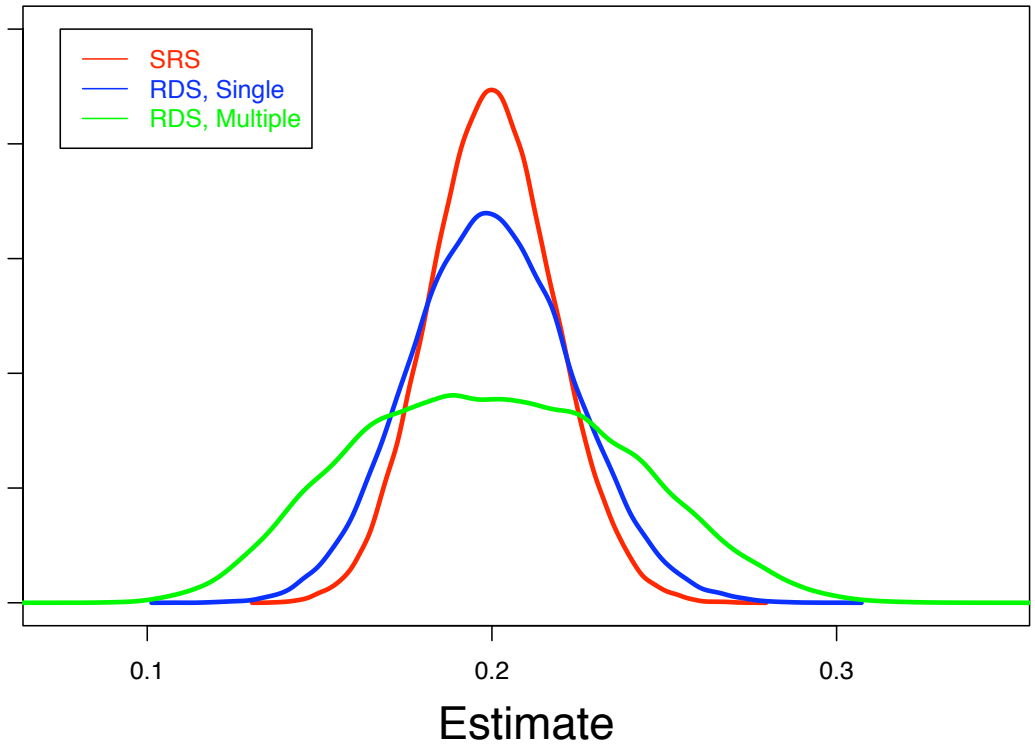
	95% CI	effective sample size
Simple Random Sampling	$\hat{p} \pm 3.6\%$	500

Comparing sampling schemes (n=500)



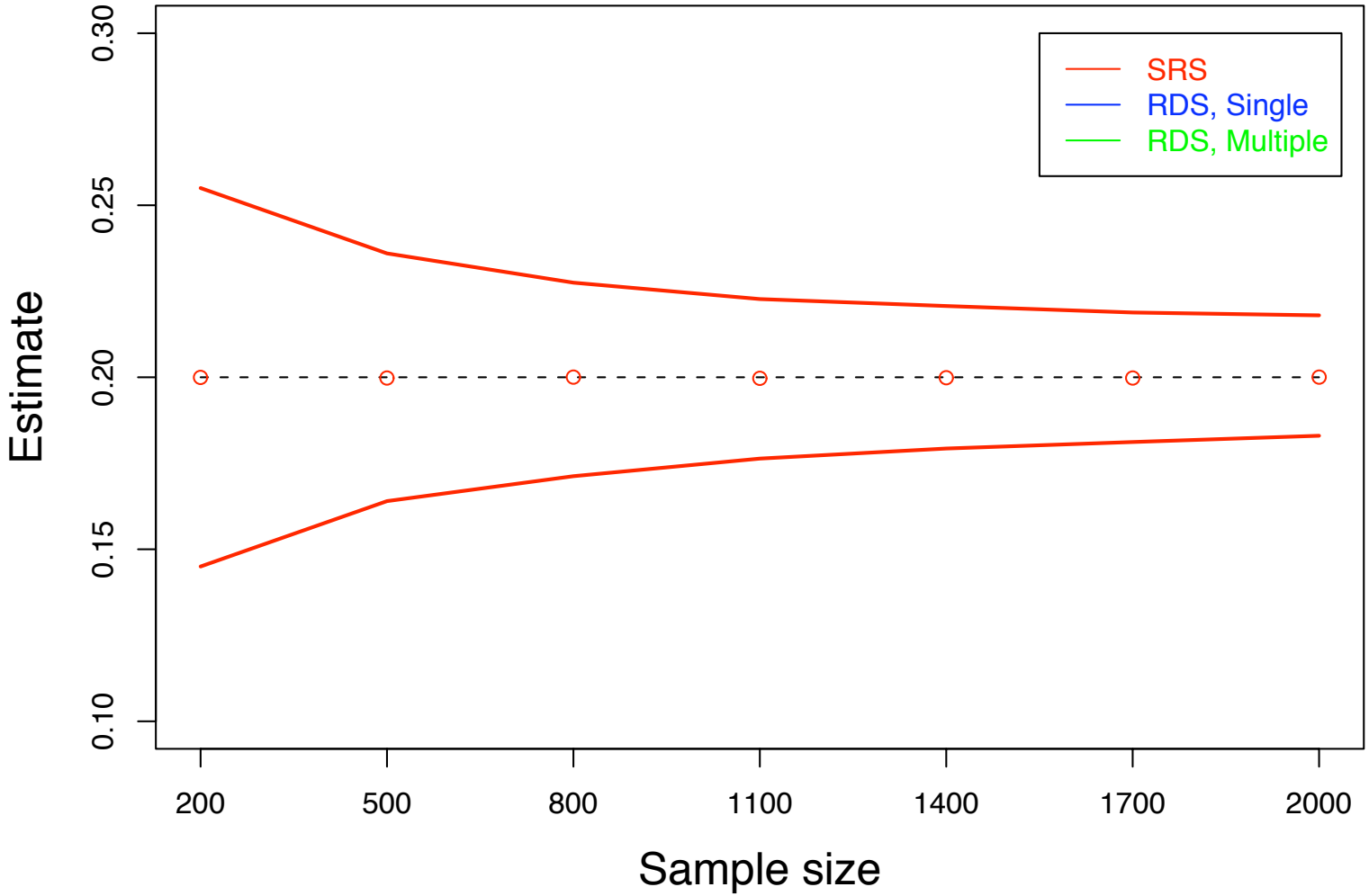
	95% CI	effective sample size
Simple Random Sampling	$\hat{p} \pm 3.6\%$	500
RDS – Single Recruitment	$\hat{p} \pm 4.3\%$	335

Comparing sampling schemes (n=500)

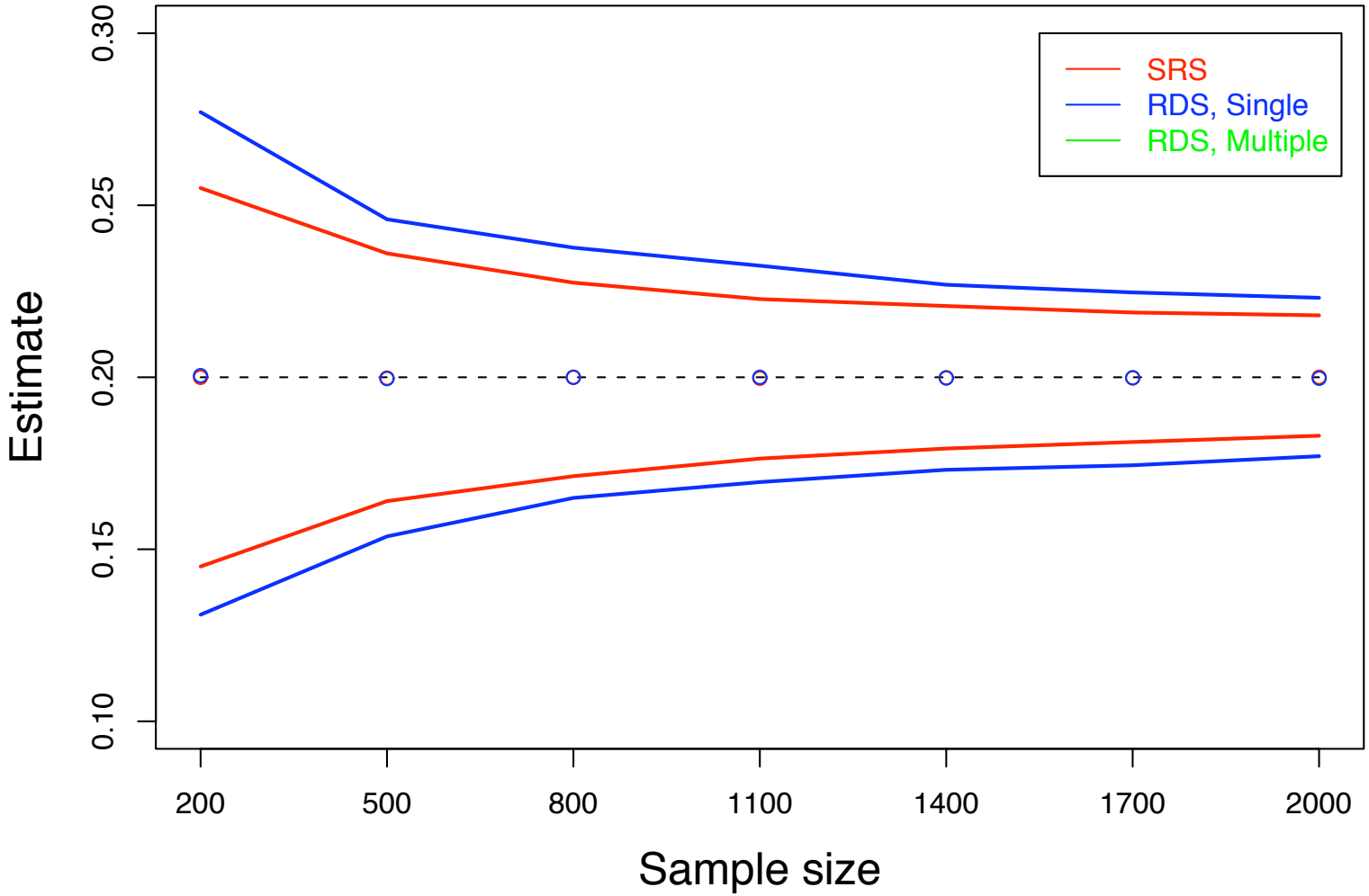


	95% CI	effective sample size
Simple Random Sampling	$\hat{p} \pm 3.6\%$	500
RDS – Single Recruitment	$\hat{p} \pm 4.3\%$	335
RDS – Multiple Recruitment	$\hat{p} \pm 6.7\%$	136

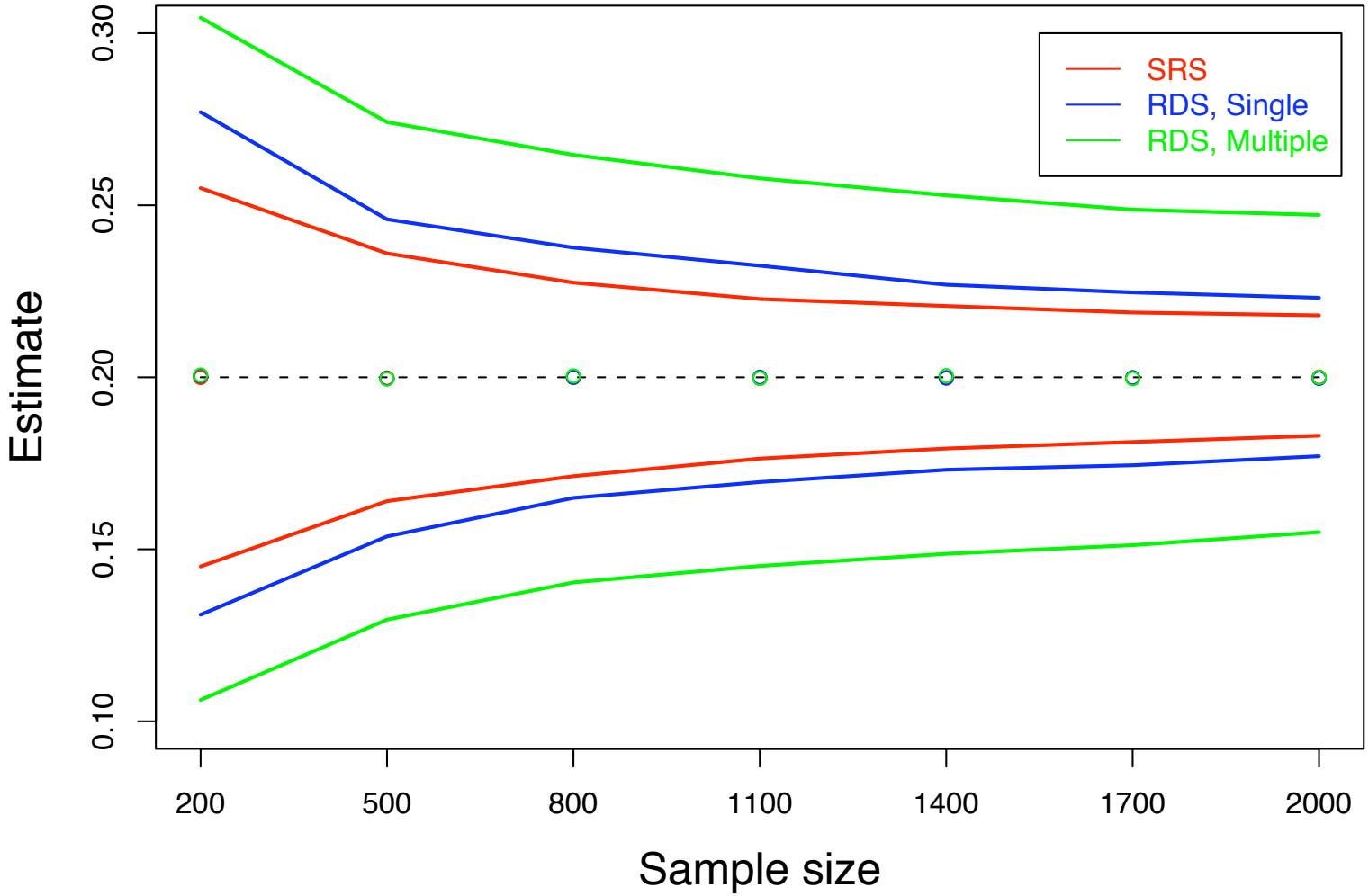
Comparing sampling schemes



Comparing sampling schemes



Comparing sampling schemes



What does this increased variability mean for practice?

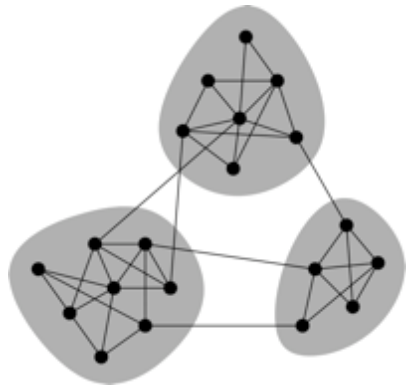
This increased variability of the RDS estimates means that you need a larger sample to reach the same level of precision as you would with simple random sample.

If you don't account for this, your sample may be too small to detect what you are looking for (i.e., low statistical power).

At this time, the best guess is that you should assume a design effect of 2 (Salganik 2006) meaning you will need a sample twice as large as would be needed with simple random sampling. Note that this is just a rule-of-thumb and is subject to change as more data becomes more available.

What kinds of community structure is there?

Moving beyond the toy example, what kind of community structures are there in real networks?



From: Newman (2006)

To find out we need:

- Complete friendship network data for large networks ($n > 1,000$)
- Demographic information each node

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.

Add Health

National Longitudinal Study of Adolescent Health (Add Health): longitudinal school-based survey, 3 waves (1994 to 2002).

- About 100 schools ranging in size from 100 to 2,000 students.
- Respondents were asked to chose from a roster up to 5 male friends and up to 5 female friends.
- Also have information on tie strength (activities done together).

Data is (sort of) publicly available through the UNC-Population Center.

Summary

We have shown that the geometry of both the social network and the recruitment network affects RDS estimates. To summarize:

- Community structure 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. For example, bottlenecks based on race can effect estimates of gender.

- A design that incorporates multiple recruitment increases the variance of RDS estimates.

Implications for practice

Implications for practice:

- In cases where there are likely to be big "bottlenecks," don't use RDS or think of the population as two populations
- Report conductance
- As much as possible, avoid multiple recruitment (3 coupons $\longrightarrow$ 2 coupons)
- Monitor convergence, perhaps using parallel chain method of Gelman and Rubin (1992)
- Allows for relaxing uniform recruitment assumption

Open questions

- How can we use RDS data for regression-type estimates?
- What are the appropriate diagnostics that can be used to test if the sampling is going wrong?
- How robust are the estimates to violations of assumptions?
- What kind of biases exist when sample size is finite?
- Can we improve the confidence intervals around our estimates?
- For what kinds of populations should RDS not be used?
- What is the relationship between participant self-reported degree and probability of selection?

Conclusion

Further information can be found our paper:

- Goel and Salganik, "Respondent-Driven Sampling as Markov Chain Monte Carlo." Under review.

If you have any questions or suggests, feel free to email me (mjs3@princeton.edu) or Sharad, (goel@yahoo-inc.com).