Sampling a hidden population without a sampling frame: A practical application of Network Sampling with Memory

Ted Mouw
M. Giovanna Merli
Ashton Verdery
Jennifer Shen
Jing Lee

Department of Sociology
University of North Carolina and Duke University

April 30, 2014

Overview of this talk:

1) Why would a demographer want to sample from a network?

2) What is NSM?

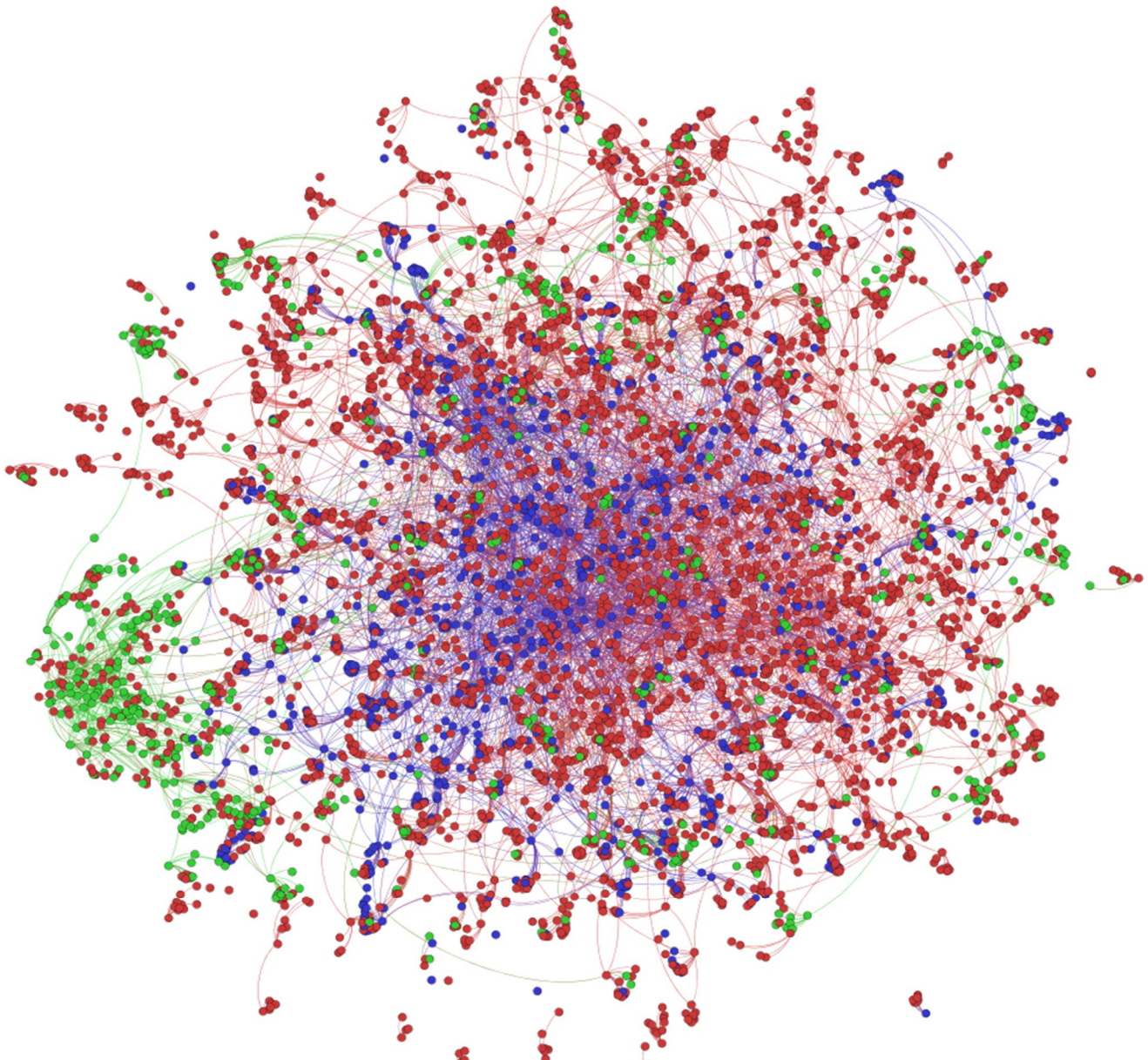
3) forward NSM → do not have to return to original respondents to get contact information for new interviews.

4) Because of the complex sampling process, use a bootstrap resampling approach to get sampling weights.

motivation

- immigration is a network process
- lack of data on social networks and international migration–>collect network data
- mixed method community studies
- vs. large scale seconadary data sets

Figure 3: The binational network of sampled and nominated individuals in the 2010 Network Survey of Immigrant Transnationalism.



HOJA 1: APUNTAR LOS NOMBRES DE AMIGOS O CONOCIDOS MAYORES DE 18 AÑOS QUE VIVEN HOUSTON, TEXAS.

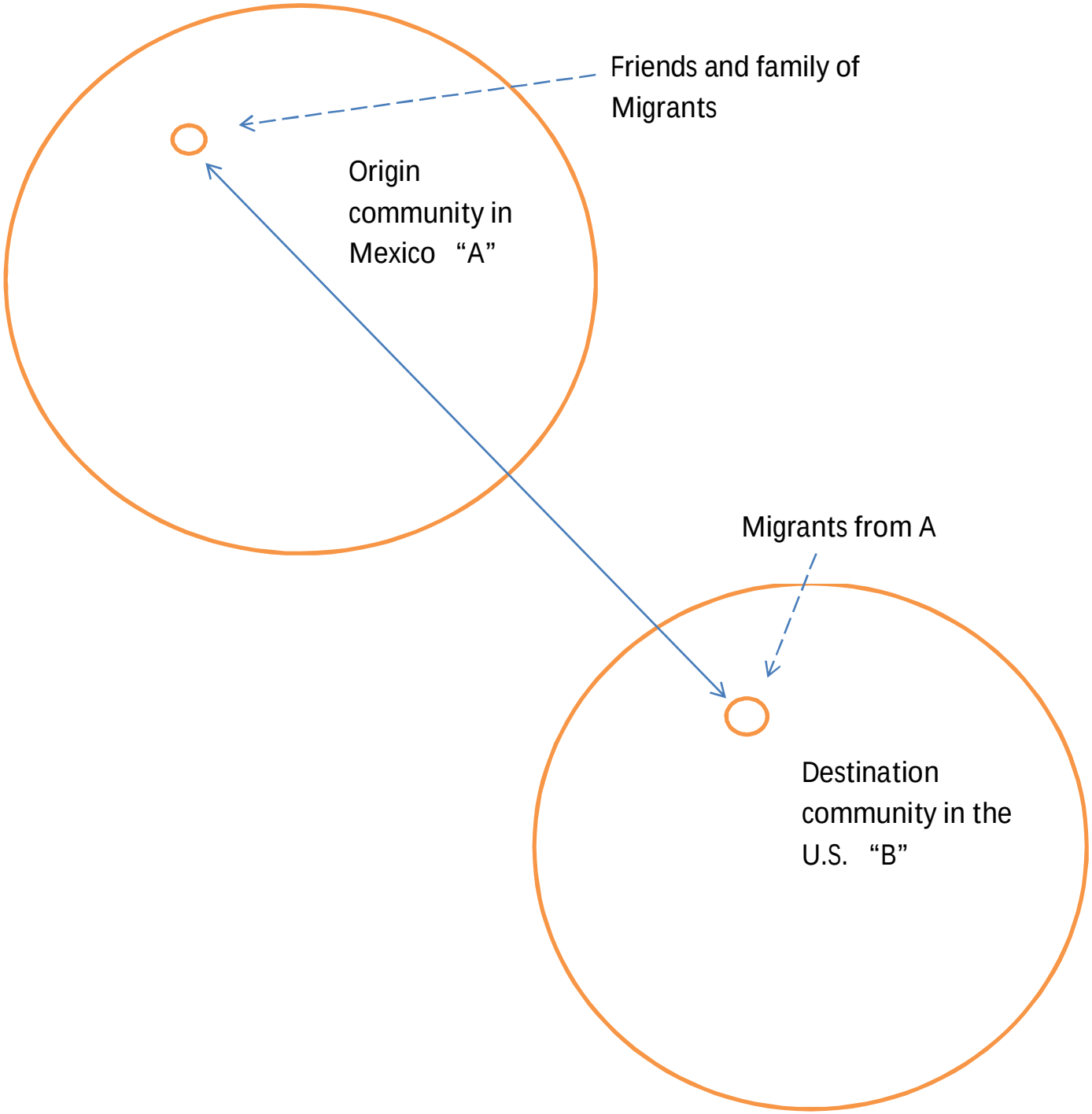
Empezamos. ¿Podría apuntar información básica acerca sus redes sociales en la hoja número uno? Además, es importante que anote los nombres y apodos, la información sobre ocupación, edad, y otras características para evitar confundir personas que tengan nombres similares.

#	Nombre (completo)	Apellido (si lo sabe) (solo primeras 4 letras)	Apodo (si lo sabe)	Ocupación (tipo de trabajo)	Sexo H= hombre M= mujer	Edad	¿Vive aquí con niños propios? (indique cuantos)	Lugar de origen: Estado o ciudad	¿Cuántos años hace que conoció a esta persona?	Aproxima-damente, ¿cuántos años vivió en Houston, Texas?	¿Con que frecuencia usted se comunica con esta persona? 1=cada día 2=cada semana 3=cada mes 4=cada año 5=menos que cada año
A1					H M		0 1 2 3+				
A2					H M		0 1 2 3+				
A3					H M		0 1 2 3+				
A4					H M		0 1 2 3+				
A5					H M		0 1 2 3+				
A6					H M		0 1 2 3+				
A7					H M		0 1 2 3+				
A8					H M		0 1 2 3+				
A9					H M		0 1 2 3+				
A10					H M		0 1 2 3+				

Hidden and rare Populations

Collecting data from a hidden population is difficult because of the absence of a sampling frame

Sampling migrant networks—the needle in a haystack problem



Hidden and rare Populations

Collecting data from a hidden population is difficult because of the absence of a sampling frame

"Respondent Driven Sampling" (RDS)–a random walk (RW) based approach

current respondent gives 1-3 coupons to friends, who become the next wave of respondents

accuracy of network sampling

.

Bias

–RWs are unbiased in large samples

–exhibit bias in finite samples

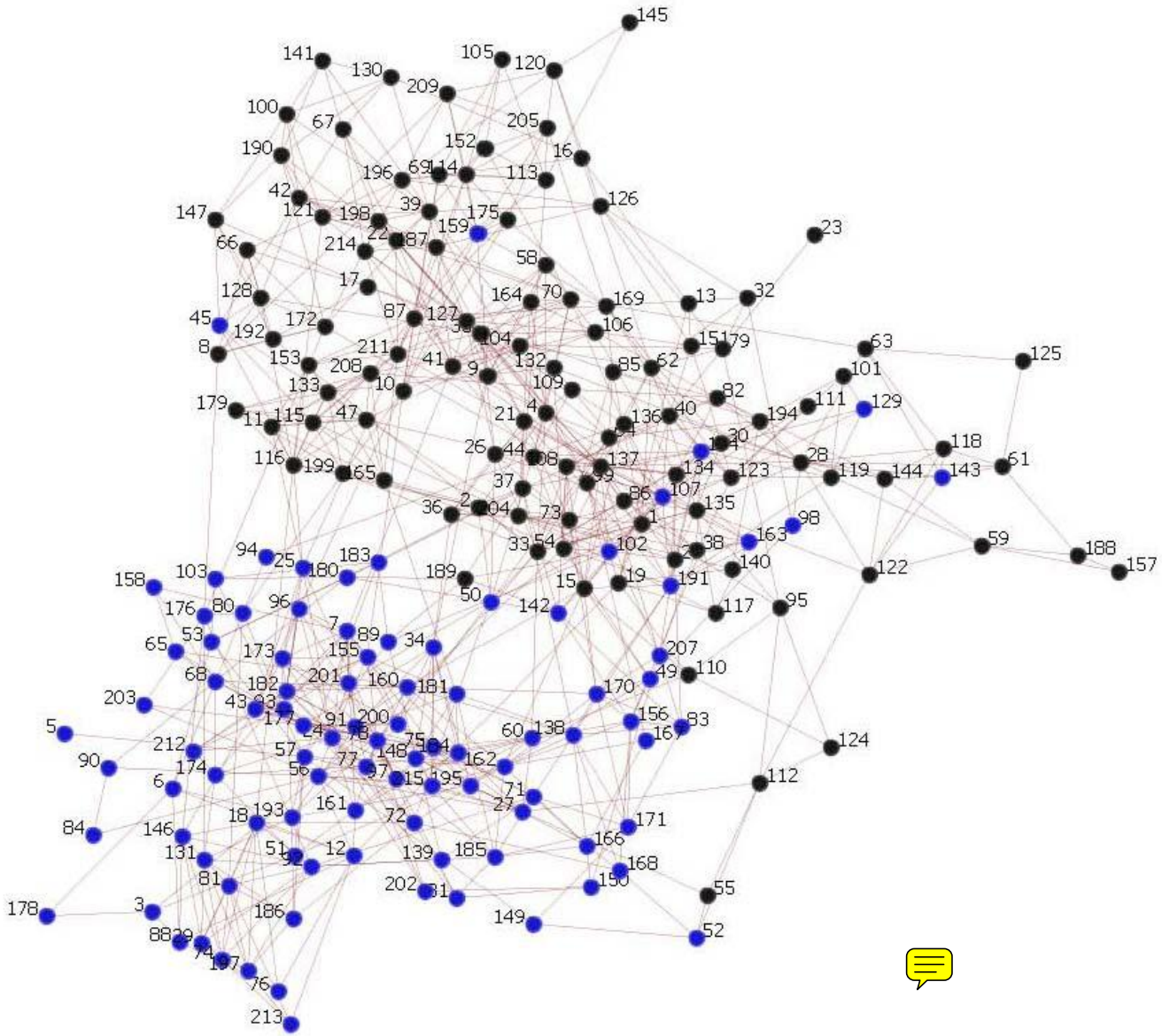
sampling variance

Sampling variance
–Design effect (DE) is the ratio of sampling variance of the network sample to simple random sampling.

–the DE of RWs and RDS is a function of the structure of the network (in addition to sample size)

Dartboard analogy of accuracy (sampling bias) and precision (sampling variance)

Figure 4: Add Health Network # 112.



Notes: Nodes colored by student race. [Black = White , Blue = Non-white]
Node ID numbers have been randomly assigned.

Network Sampling with Memory

.

Mouw and Verdery. 2012 "Network Sampling with Memory: A proposal for more efficient sampling from social networks" Sociological Methodology

Collect network data (example: 2010 NSIT)

Use the network data to sample more efficiently.

Network Sampling

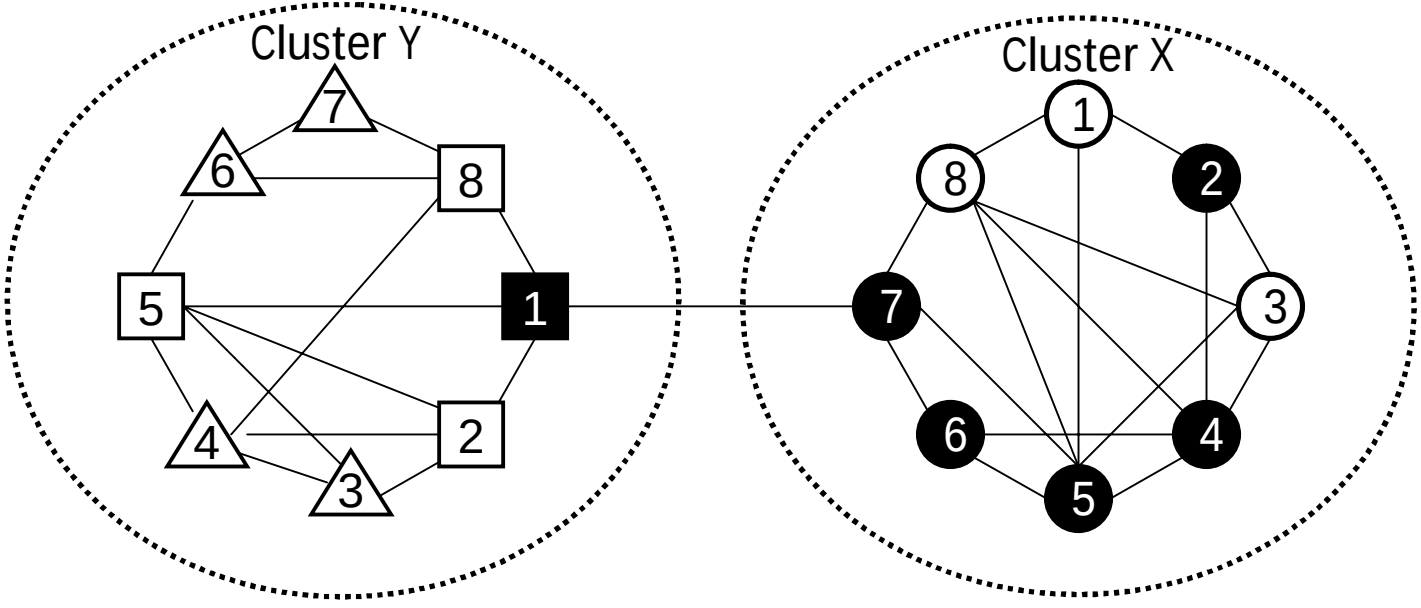
two sampling modes, List and Search

List mode: (a) keep a list, L, of all nominated network members
(b) sample with replacement from L
(c) "Even sampling" of new nodes–sample new nodes at the current cumulative sampling rate
very simple
converges to simple random sampling

search mode

- Search mode: push the sample to explore the network
- Bridge tie: a node that connects two clusters of nodes

Figure 1. Illustrative network with two clusters.



Notes: Hollow nodes are unsampled, dark nodes are sampled. Circles indicate nodes nominated 2+ times, squares indicate nodes nominated 1 time, triangles indicate nodes nominated 0 times.

hybrid approach

- List mode [sample from list L]
- Search mode [sample friends of bridge ties]
- NSM hybrid approach: start in Search mode
- –switch to List mode as the network is explored

Results

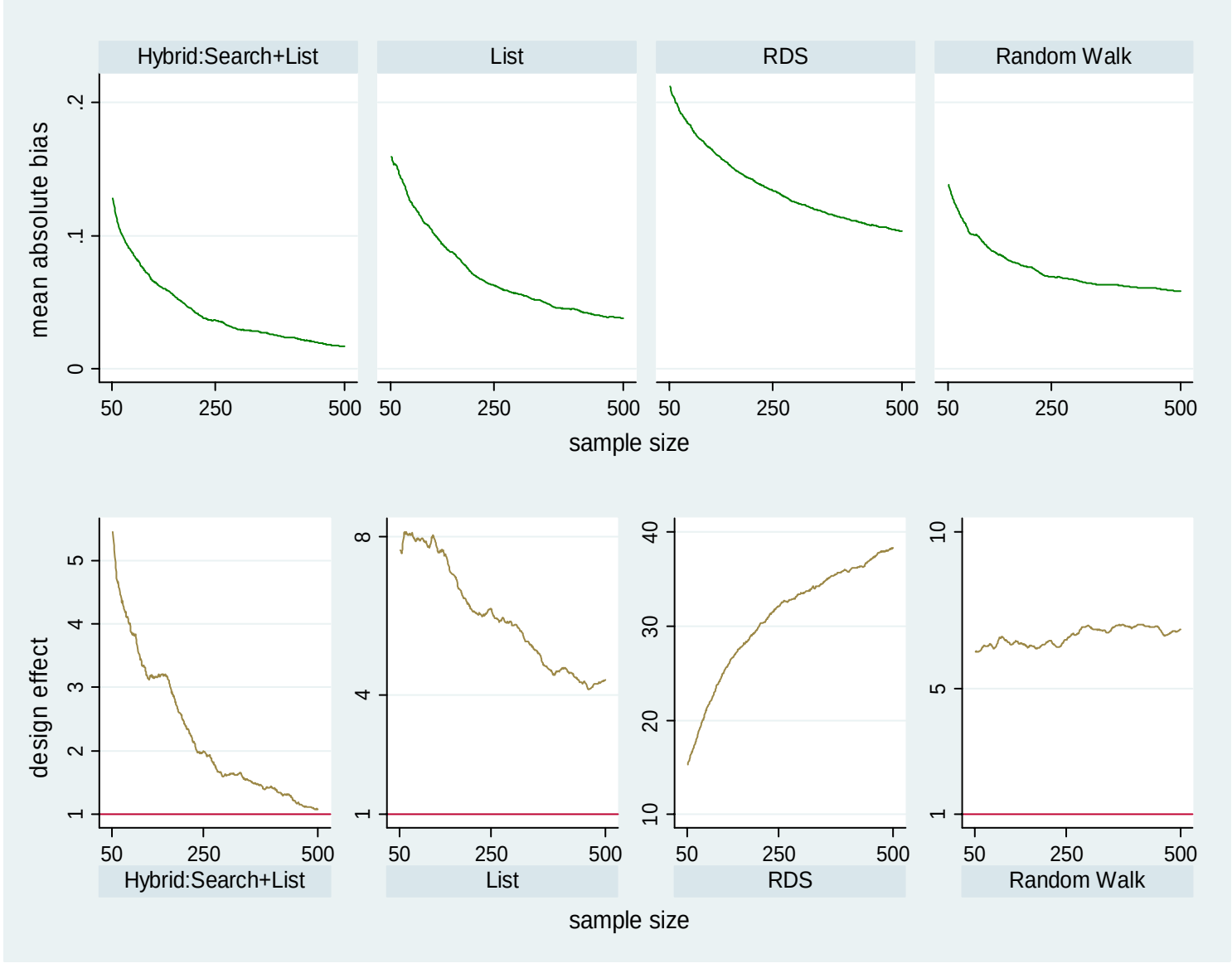
- Test NSM vs RWs and RDS using 162 university and school networks from Facebook and Add Health
- size ranges from 300 to 16,500 nodes
- estimate % white (Add Health) and % first year students (Facebook)
- start from a randomly selected student, repeat 500 times for each network
- calculate bias, design effects, and mean absolute bias

.

Results

- NSM has a 97.5% reduction in the design effects on these 162 networks
- 1.16 for NSM
- 77.38 for RDS

Figure 6: Sampling results the largest Facebook university network (16,280 nodes, dependent variable: proportion freshman)



is it feasible?

- is it feasible to collect network data on hidden populations?
- (1) 2010 NSIT
- (2) cost effectiveness of gains in precision.

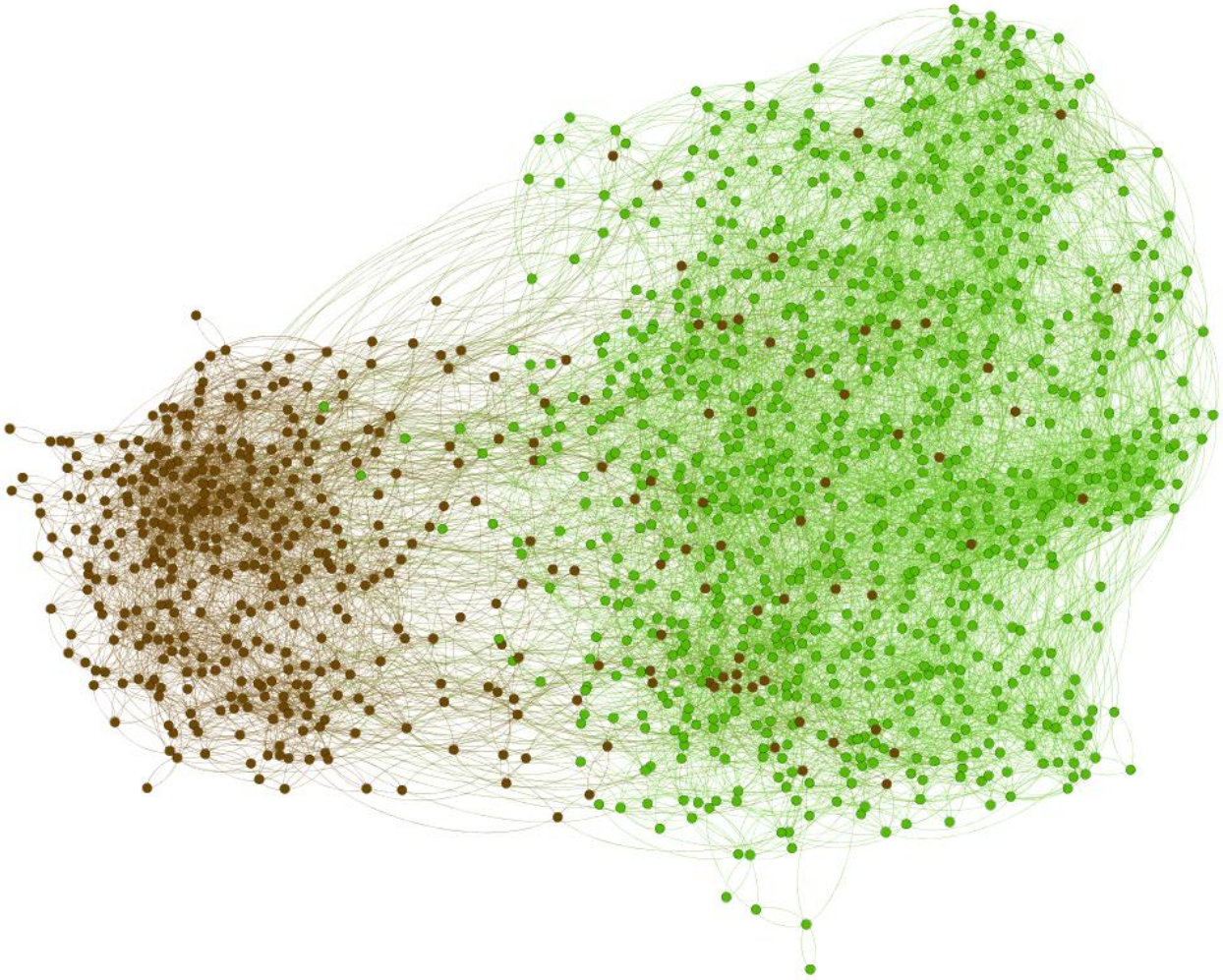
"practical" version

- 2012 paper required recontact of respondent to get contact information on alters
- "forward" sampling variant "FNSM"
- ask for contact information on a small number of alters from each interview

forward NSM

- let m=number of alters to collect
- let k=subset of m with contact information
- in the 2012 paper, m=k=20
- what about an easier approach, m=7, k=3?
- How well does FNSM work in simulated sampling? (insert results table)

Figure 1: The social network for Add Health School #150, by race –the test network for this paper
Green nodes = white
Brown nodes = non-white



Descriptive statistics for Add Health network 150

1,281 students (“nodes”),

67.3% white

10,414 ties in the data

587 cross race ties (white to non-white)

7.919% of whites’ friends are non-white (587 out of 6,838 ties)

Conclusions: there is homophily in the data, but there are no “choke points” in the network because there are lots of cross-group ties.

Test simulated sampling: collect 500 samples of 500 interviews for each sampling method.

Calculate the 95% interval of the estimated proportion white in the school and the design effect.

Figure 2

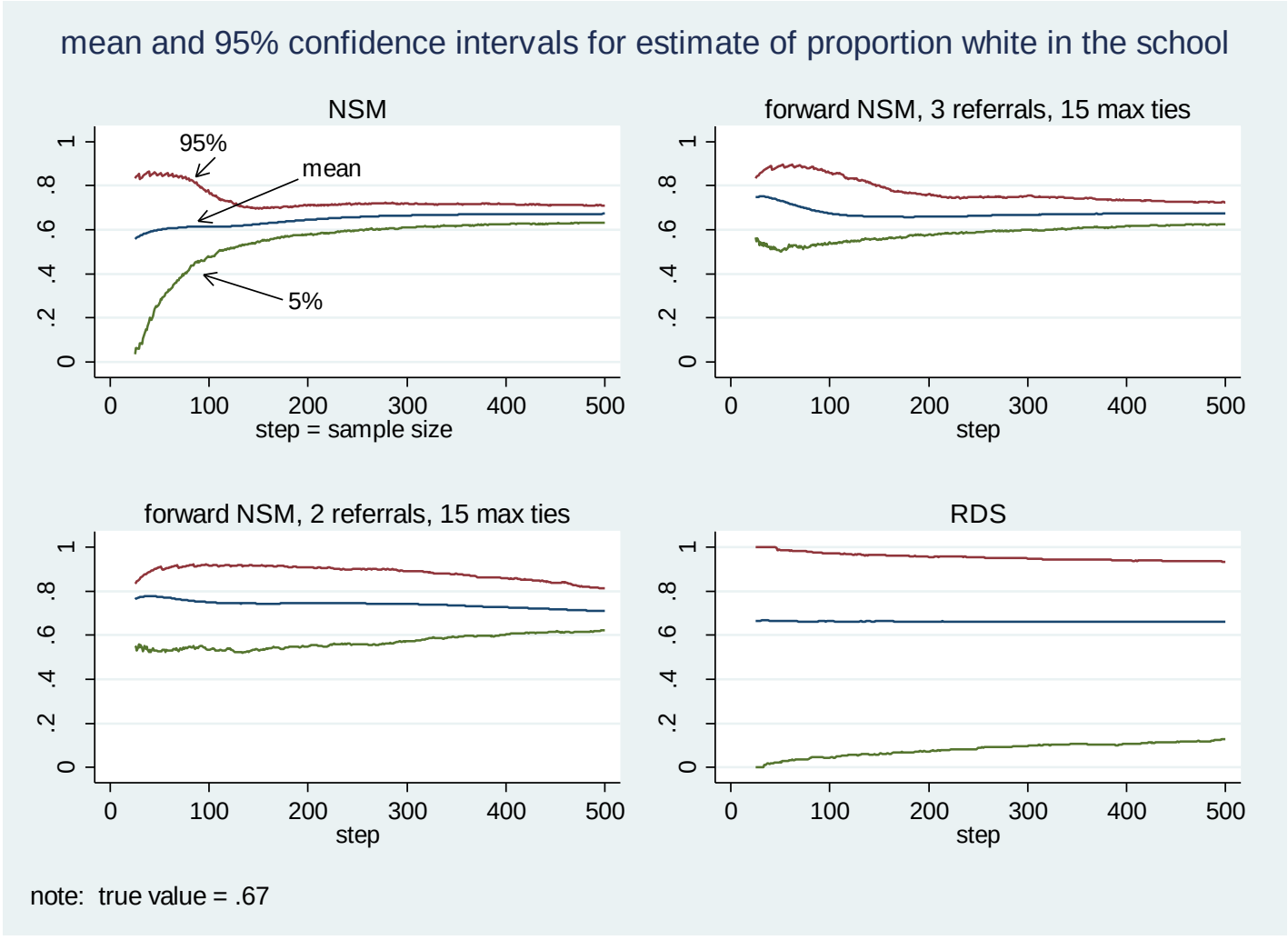
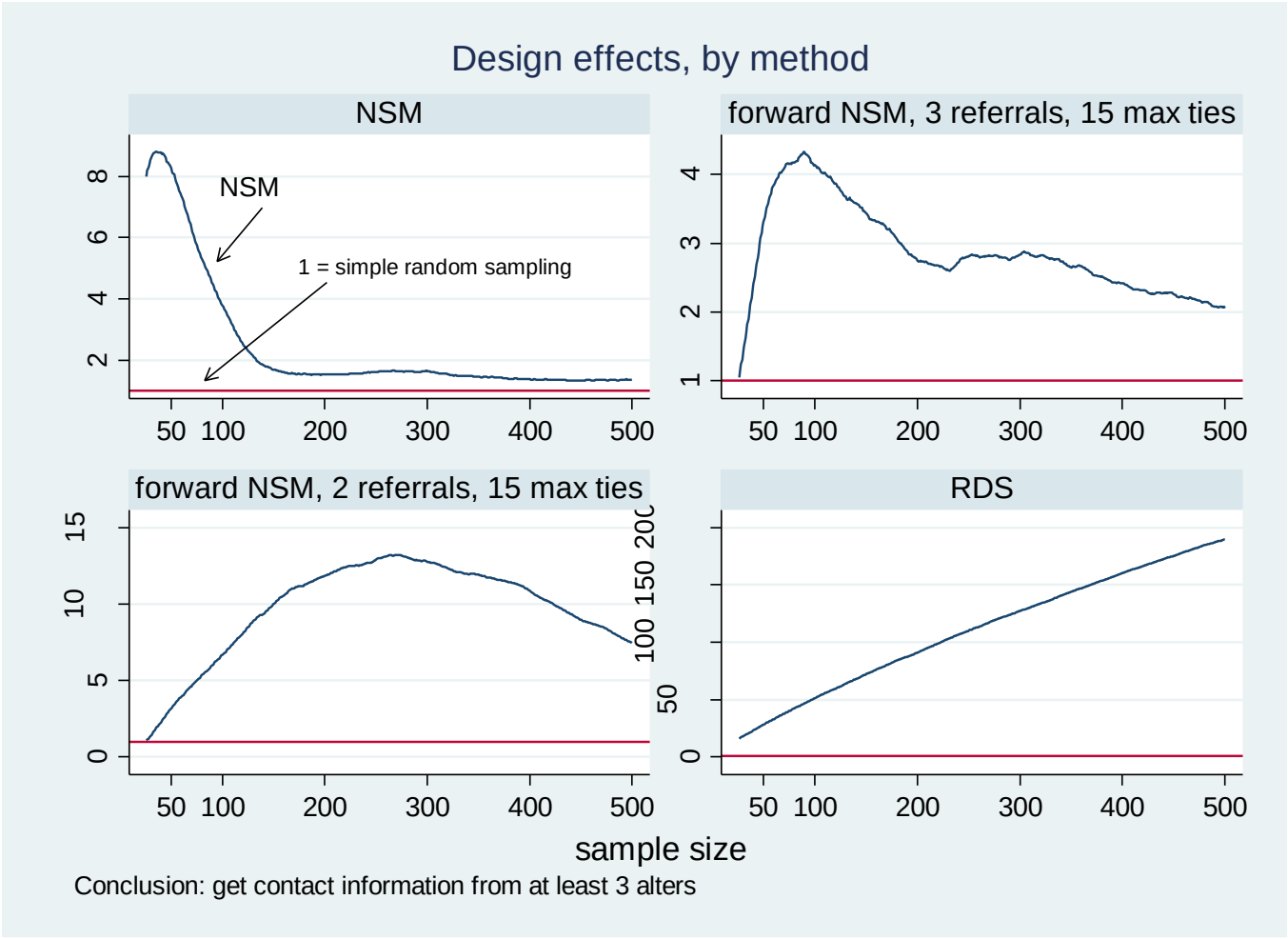

mean and 95% confidence intervals for estimate of proportion white in the school
NSM
95%
mean
5%
step = sample size
forward NSM, 3 referrals, 15 max ties
step
forward NSM, 2 referrals, 15 max ties
step
RDS
step
0
.2
.4
.6
.8
1
0
100
200
300
400
500
note: true value = .67

Figure 3


Design effects, by method
NSM
NSM
1 = simple random sampling
forward NSM, 3 referrals, 15 max ties
forward NSM, 2 referrals, 15 max ties
RDS
sample size
Conclusion: get contact information from at least 3 alters

Design Effects for 5 large university based Facebook networks (2005).
Y = first year student
#nodes: mean= 9,880 (6,788-12,660)

Method	Mean DE	Std. Dev.	Freq.
1. NSM, k=m=20	0.85	0.187	5
2. fNSM, k=m=7	2.35	0.627	5
3. fNSM, k=2, m=7	4.38	0.658	5
4. fNSM, k=3, m=7	2.18	0.563	5
5. RDS	38.19	6.24	5

boostrap standard errors

- advantage of NSM approach is that we have network data
- –> have cross-ties
- we can use this to get estimated standard errors

bootstrap

- in sampled data, we do not know the ties between nominated (but unsampled) cases
- use ERGM models to predict these ties
- add to the sampled data
- then resample the combined data
- repeat multiple times to calculate % of sampling
- use the inverse of these as weights.
- (see slides)

Test of bootstrap resampling for sampling weights.

Test for samples of 100 interviews.

--we want to test the efficacy of the weights when the sampling process has not converged to a DE of 1.

Procedure:

a) for each of the 500 replications of simulated sampling, take the first 100 interviews. These are simulated surveys of 100 cases.

b) For each of these simulated surveys, collect the network composed of the sampled and nominated ties.

These are the "sampled networks".

c) For each sampled network, run an exponential random graph model (ERGM), and use the results to predict the existence of ties between nominated but un-sampled nodes. Add these predicted ties to the actual sampled ties. This is a "simulated network" → composed of sampled ties and predicted ties.

Create 100 of these simulated networks for every sampled network. [For a total of 100 x 500 networks for each method—the number of simulated networks multiplied by the number of replications]

d) For each simulated network, resample the data using simulated sampling, keeping track of the number of times each node was sampled. Start every sample with the original seeds, so the bias of non-random seeds is minimized.

e) calculate the sampling importance weights as the inverse of the number of times a node was sampled.

Results of simulated bootstrap sampling

1. Got several emails from the computer support tech asking what I was doing running so many jobs on the Unix system.

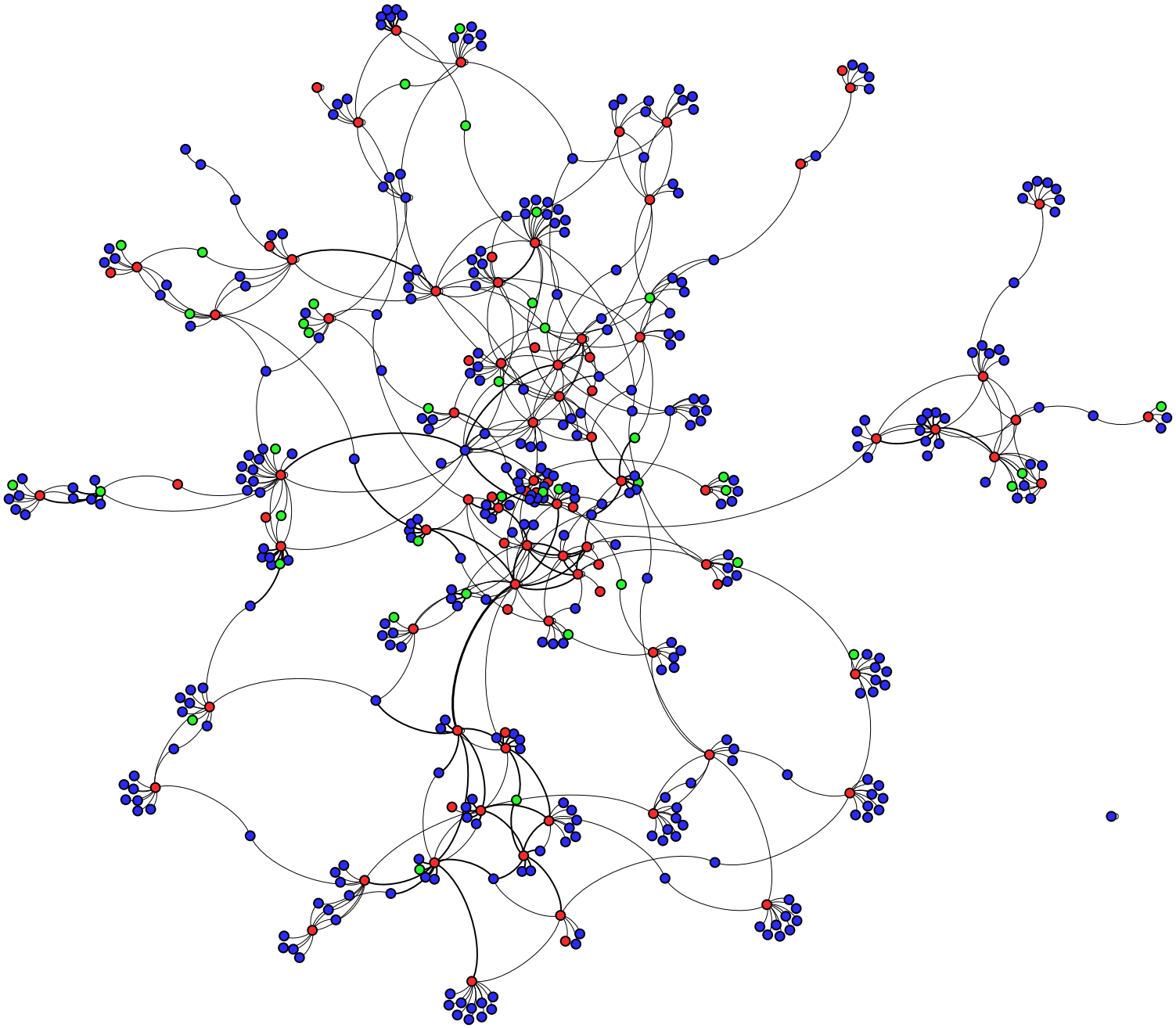
2. With a sample size of 100, NSM exhibits a small finite sample bias because it is still sampling in the "search" mode. Using the weights reduces this bias (for this network) from .03 to .01.

3. The resampling weights do not reduce the design effects for NSM or RDS, with or without the networks.

Conclusion: resampling weights show promise for improving the accuracy of NSM samples that stopped early, but more work is needed to use them to improve the DE.

pilot study

- A Pilot Study of the Health of Chinese Migrants in Tanzania
- alter information: last name & final 4-digits of phone number
- up to 10 alters (m=10)
- up to 3 with contact information (k=3)



2013 The Health of Chinese Migrants in Tanzania Survey
Red dots = interviewed nodes
Blue dots = nominated, but not interviewed
Green dots = refused

estimated population size

- use capture-recapture method
- L=number of nominated nodes
- p1 = proportion of nominated nodes that have been nominated only 1 time (and not sampled)
- G = estimated population size = L/(1-p1)
- approximate –> these are correlated samples –> but less correlated as the sample size increases
- also (obviously) depends on network structure

estimated population size

- G=L/(1-p1) works well in practice for NSM
- (example: diagnostics for Facebook network 10)
- similar in all other networks in the NSM paper

sample diagnostics

- trend in G over time (i.e., as sample size increases) can be used as a diagnostic
- if G stablizes, then the sampling process is quasi-random
- (intuition) if G is increasing, that means you are still in the "discovery" phase of the sample
- –>could be an indication that you have a "snowball"-like sample where G is underestimated early in the sample
- (need to jump across bridges to other clusters)
- (insert diagnostic slides)

Figure 8: Diagnostic Properties for Facebook network #10 by sample size (step)

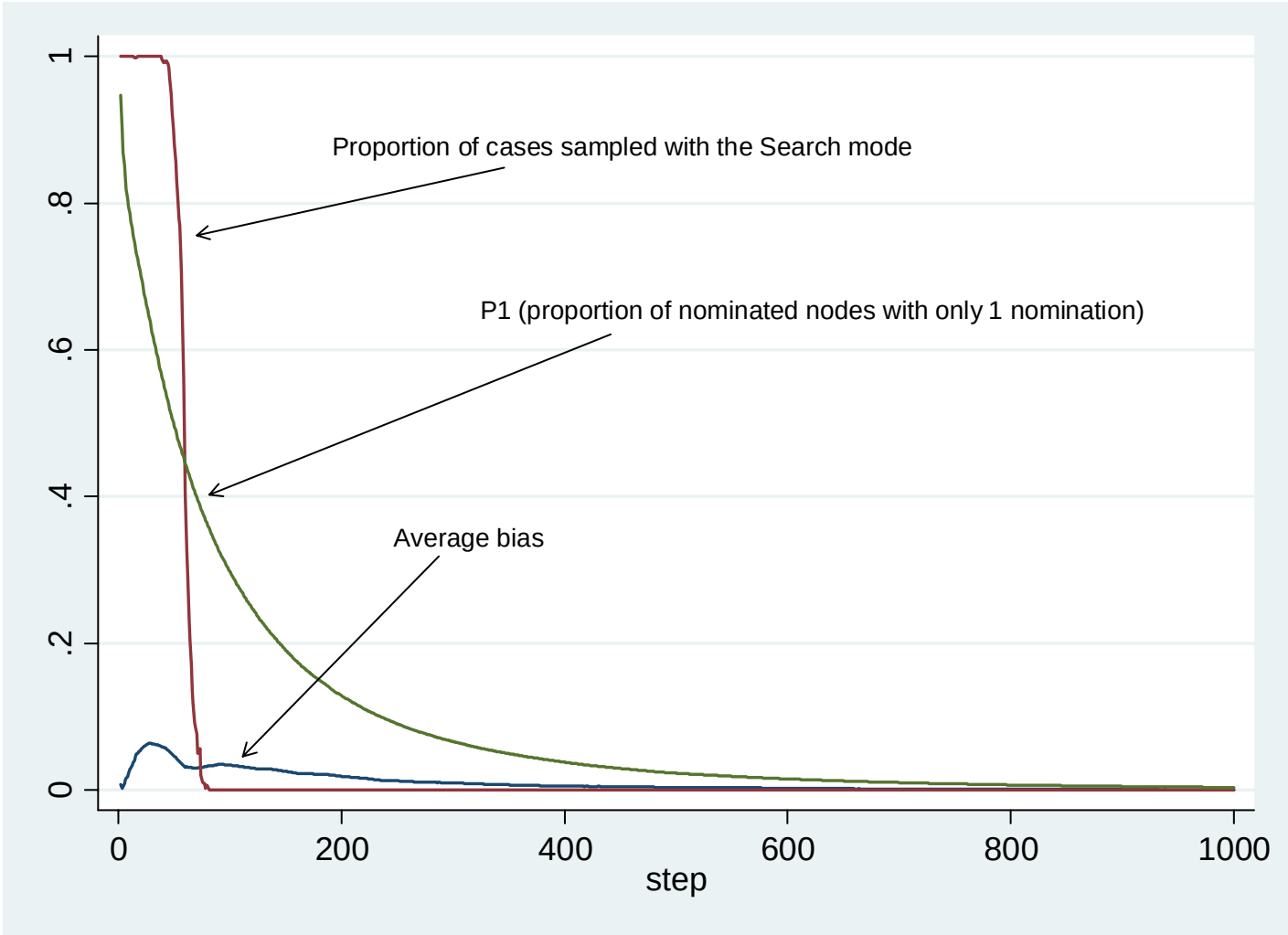
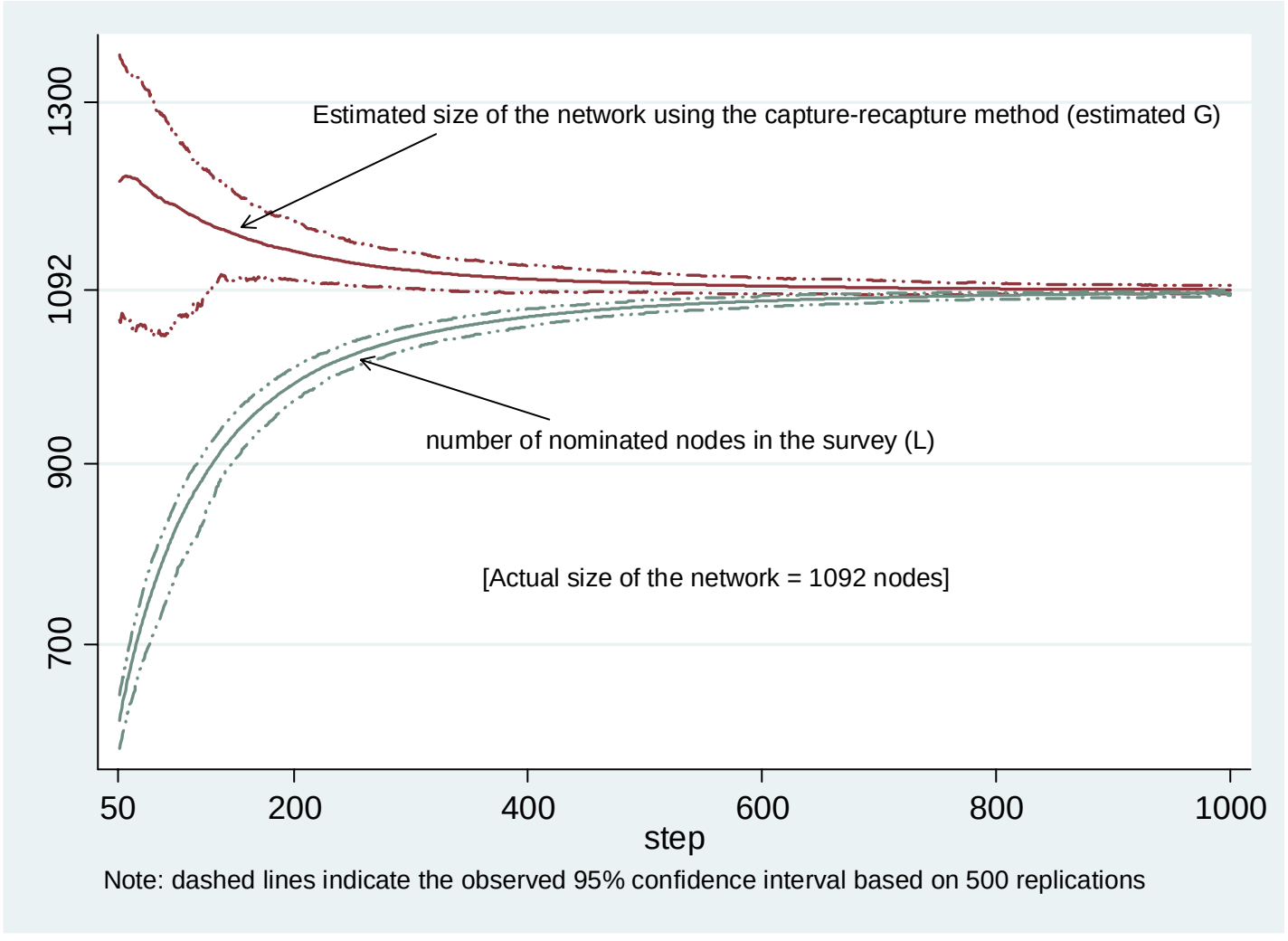
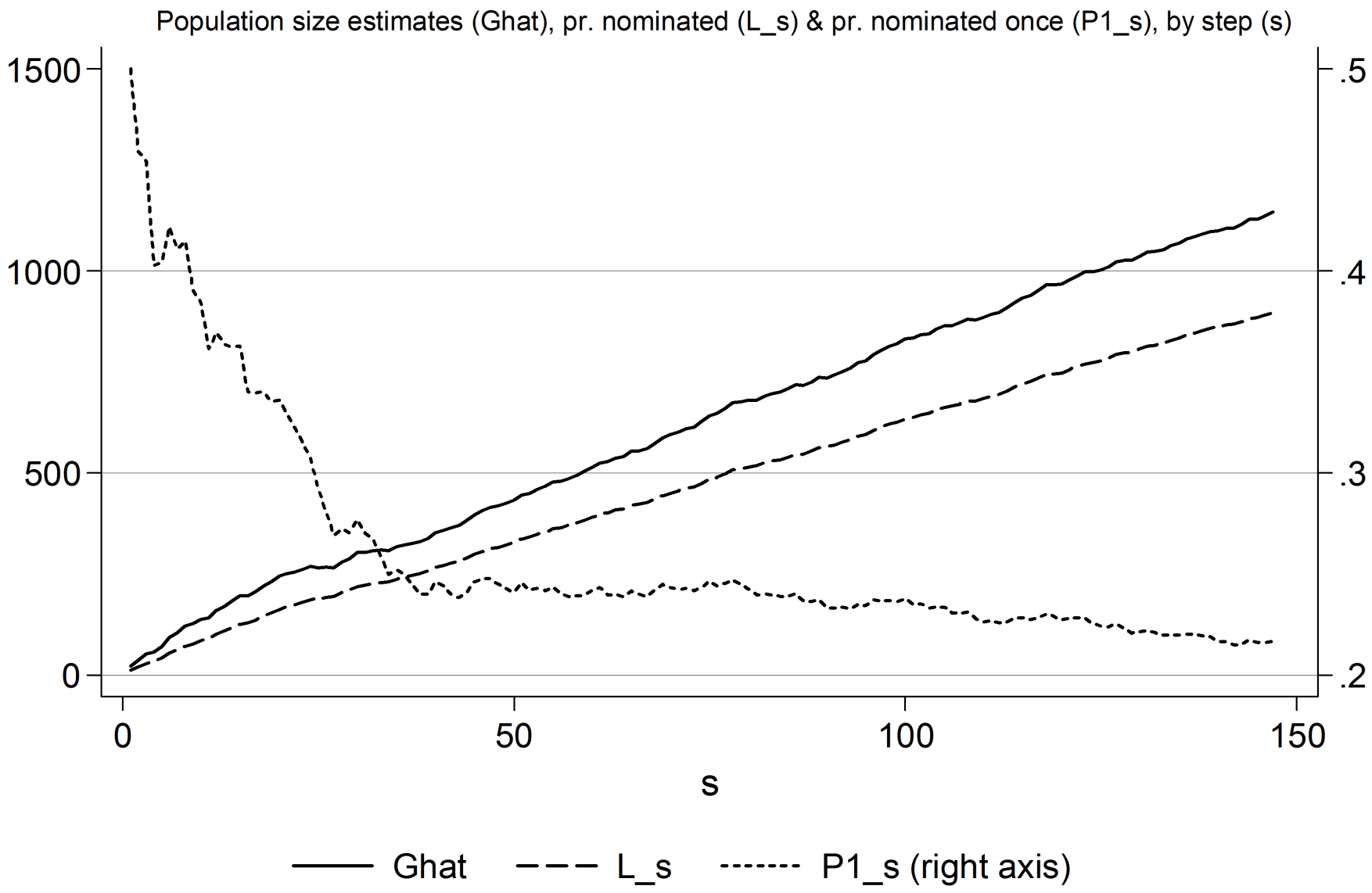


Figure 9: Number of nominated nodes and estimated network size by sample size (step), Facebook network #10




Population size estimates (Ghat), pr. nominated (L_s) & pr. nominated once (P1_s), by step (s)
1500
1000
500
0
.5
.4
.3
.2
0
50
100
150
s
Ghat
L_s
P1_s (right axis)

stopping point

- choose % of network to sample beforehand = %p
- stop when G stabilizes and you have reached %p