Day-1 Session-1: Graph Sampling Designs for Epidemic Prevalence Estimation

Li-Chun Zhang[1,2,3] **and** ***Melike Oguz-Alper***[2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyraa, Norway***
[3] ***Universitetet i Oslo***

Using individual contacts for sampling

Q. Suppose there are 1000 infected individuals in a population of size 100 000. Which is the most efficient design (of sample size n) for estimating the prevalence of the infected, where $1000 \leq n < 100000$*?*

A. If anyone infected has probability 1 to be sampled, because the Horvitz-Thompson estimator of the total no. infected is then 1000 with zero sampling variance.

Let $\pi_i = Pr(i \in s_0)$ *be the inclusion probability to an initial sample* s_0*. Repeatedly add to the sample all the contacts of anyone infected until no one can be added.*
Let s *be the final sample and* $\pi_{(i)} = Pr(i \in s)$*.*
We have $\pi_{(i)} > \pi_i$ *for any infected individual* i*.*

Population, prevalence, graph, case network

Population $U = \{1, 2, ..., N\}$. At given time, for $i \in U$:

$$y_i = 1 \text{ if } i \text{ is a } \textit{case} \quad \text{and} \quad y_i = 0 \text{ otherwise}$$

Case total: $\theta = \sum_{i \in U} y_i$ *prevalence:* $\mu = \frac{\theta}{N}$

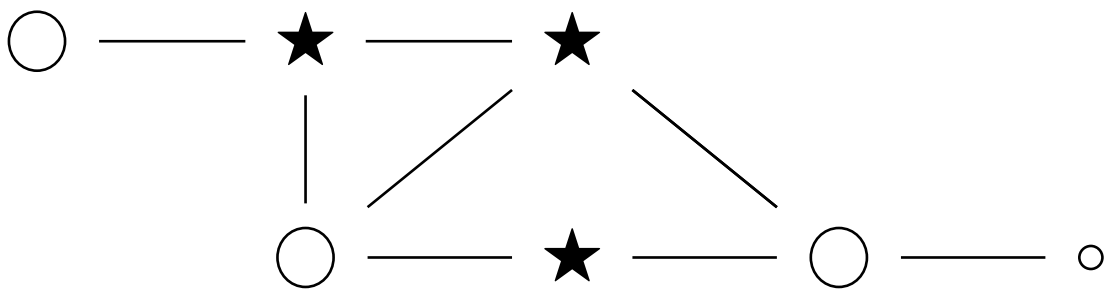
Population graph $G = (U, A) = (\text{nodes, edges})$
where $(ij) \in A$ for $i, j \in U$ if i and j are *in-contact*

- undirected: $(ij) = (ji)$ by definition
- simple: not multiple edges between any i and j

Case network in G each of connected nodes with $y_i = 1$

Setting: U is known, but not edges A or case networks

Adaptive network tracing



edge node: noncase ◯ adjacent to case ★; other noncase ∘

Adaptive tracing: sample all the nodes adjacent to i,

$$\nu_i = \{j : (ij) \in A\}$$

only if i is a case ★, not if i is a noncase ◯ or ∘... repeatedly as long as it is possible... case network κ with nodes β_κ is sampled *iff* $s_0 \cap \beta_\kappa \neq \emptyset$, hence

$$\pi_{(\kappa)} \equiv \pi_{(i)} = \Pr(s_0 \cap \beta_\kappa \neq \emptyset) \quad \text{for any } i \in \kappa$$

whereas $\pi_{(i)}$ may be unknown generally if ◯ or ∘

Adaptive cluster sampling (ACS, Thompson, 1990)

ACS: neighbourhood ν_i, binary y_i for $i \in U$

Noncase contributes $y_i/\pi_{(i)} = 0$ regardless of $\pi_{(i)}$

Each case network is a cluster of nodes

Ω = set of all case networks (or clusters) in G

Ω_s = sampled case networks

$n_\kappa = |\beta_\kappa|$ = no. cases in network κ, $n_\kappa \geq 1$ for $\kappa \in \Omega$

HT-estimator of population case total:

$$\hat{\theta}_y = \sum_{\kappa \in \Omega} \mathbb{I}(\kappa \in \Omega_s) \frac{n_\kappa}{\pi_{(\kappa)}}$$

$$V(\hat{\theta}_y) = \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} \left(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)} \pi_{(\ell)}} - 1 \right) n_\kappa n_\ell$$

Combating disease *and* estimating prevalence at once

Some simulation results (Zhang, 2020)

ACS given equal-size (c) case networks in population of size $N = 10^5$, prevalence $\mu = 0.01$ and $\theta = 10^3$. Initial sample of size m by SRS ($\eta = 1$) or size-biased sampling ($\eta = 2$), ACS with sample size $n = |s|$ by adaptive network tracing.

SRS ($\eta = 1$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.31	1631	0.24	0.58	1085	0.31	0.96	1010	0.31	0.99
1630	0.24	2423	0.15	0.40	1766	0.24	0.93	1646	0.24	0.99
2420	0.20	3306	0.10	0.23	2614	0.19	0.89	2443	0.20	0.99
5000	0.14	5944	0.02	0.03	5352	0.12	0.79	5048	0.14	0.97
10000	0.09	10900	0.00	0.00	10551	0.07	0.59	10090	0.09	0.95
Size-biased ($\eta = 2$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.22	1840	0.13	0.32	1160	0.21	0.91	1020	0.22	0.99
5000	0.09	5901	0.00	0.00	5549	0.07	0.60	5090	0.09	0.95
10000	0.06	10802	0.00	0.00	10692	0.04	0.31	10159	0.06	0.89

Size-biased sampling of s_0 if $\eta \equiv \pi(y_i = 1)/\pi(y_i = 0) \neq 1$

Population graph $G_t = (U_t, A_t)$ over time

Unless lockdown, A_t varies over time even when $U_t \equiv U$
Basic designs (Zhang, 2020), with fixed s_0 over time:

a. Panel without tracing, same $s(t) = s_0$ over t

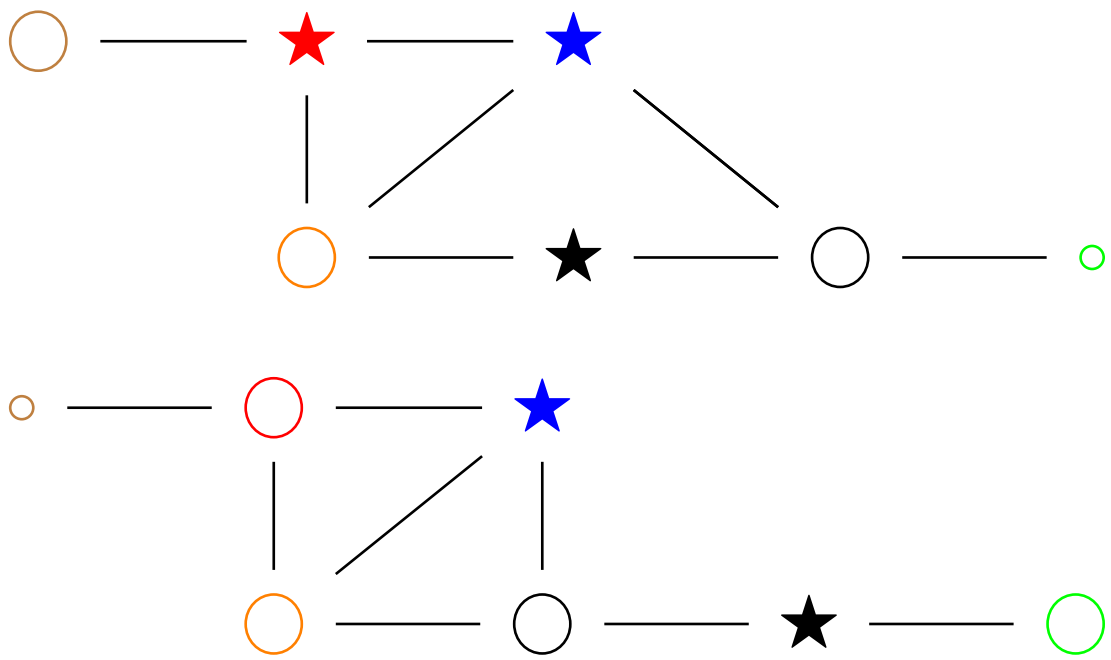
b. Panel ACS (pACS)

- ◇ sample $s(t)$ by ACS based on s_0 and A_t
- ◇ sample $s(t+1)$ by ACS based on s_0 and A_{t+1}

c. Iterative ACS (iACS)

- ◇ sample $s(t)$ by ACS based on s_0 and A_t
  known $\pi_i(t)$, $\pi_{ij}(t)$ for cases $i, j \in s(t)$, not $U_t \setminus s(t)$
- ◇ sample $s(t+1)$ by ACS based on $s(t)^*$ and A_{t+1}
  where $s(t)^* = s_0 \cup \{i \in s(t) \setminus s_0 : y_{i,t} = 1\}$
  $\pi_i(t+1)$, $\pi_{ij}(t+1)$ unknown for cases $i, j \in s(t+1)$

Population graph $G_t = (U_t, A_t)$ over time



Suppose ★ $\in s_0$ in the example here:

- Panel: ★ $\in s(1)$ and ★ $\in s(2)$
- pACS: $\{$★, ★, ○, ○, ○$\} \subset s(1)$ and ○ $\in s(2)$
- iACS: $\{$★, ★, ○, ○, ○$\} \subset s(1)$ and $\{$○, ★, ○, ○$\} \subset s(2)$

Change estimator

Change of prevalence: $\nabla_{t,t+1} = \mu_{t+1} - \mu_t$

Panel: HT-estimator

$$\hat{\nabla}_{t,t+1}^{panel} = \frac{1}{N} \sum_{i \in s_0} \frac{1}{\pi_i} (y_{i,t+1} - y_{i,t})$$

pACS: HT-estimator

$$\hat{\nabla}_{t,t+1}^{pACS} = \frac{1}{N_{t+1}} \sum_{i \in s(t+1)} \frac{y_{i,t+1}}{\pi_i(t+1)} - \frac{1}{N_t} \sum_{i \in s(t)} \frac{y_{i,t}}{\pi_i(t)}$$

iACS: Unbiased estimator

$$\hat{\nabla}_{t,t+1}^{iACS} = \frac{1}{N_{t+1}} \Big(\sum_{\substack{i \in s(t)^* \\ y_{i,t}=1}} \frac{y_{i,t+1}}{\pi_i(t)} + \sum_{\substack{i \in s(t)^* \\ y_{i,t}=0}} \frac{y_{i,t+1}}{\pi_i} \Big) - \frac{1}{N_t} \sum_{i \in s(t)} \frac{y_{i,t}}{\pi_i(t)}$$

NB. $s^{pACS}(t+1) \subseteq s^{iACS}(t+1)$, but $V(\hat{\theta}_{t+1}) > V(\hat{\theta}_{t+1})$ possible

Can trace $\hat{\theta}_{t+1}$ to Thompson (1990), Birnbaum and Sirken (1965)...

Some simulation results

Populations of constant size $N = 10^5$ and total $\theta = 10^3$ at $t = 1, 2$. With (number, size) of case networks: at $t = 1$, $(\bar{\theta}, c)$ networks; at $t = 2$, $(\bar{\theta}_+, c_+)$ or $(\bar{\theta}_-, c_-)$ existing networks of increasing or decreasing sizes, and $(\bar{\theta}', c')$ emerging networks.

		$t = 1$	$t = 2$		
	Characterisation	$(\bar{\theta}, k)$	$(\bar{\theta}_+, c_+)$	$(\bar{\theta}_-, c_-)$	$(\bar{\theta}', c')$
L1	Large, Quickly Evolving	(10, 100)	(2, 180)	(8, 80)	(0, 0)
L2	Large, Quickly Emerging	(10, 100)	(0, 0)	(10, 80)	(2, 100)
L3	Large, Slowly Emerging	(10, 100)	(0, 0)	(10, 90)	(5, 20)
M1	Medium, Quickly Evolving	(100, 10)	(10, 46)	(90, 6)	(0, 0)
M2	Medium, Quickly Emerging	(100, 10)	(0, 0)	(100, 6)	(10, 40)
M3	Medium, Slowly Emerging	(100, 10)	(0, 0)	(100, 9)	(10, 10)
S1	Small, Quickly Evolving	(500, 2)	(10, 42)	($\leq$490, $\leq$2)	(0, 0)
S2	Small, Quickly Emerging	(500, 2)	(0, 0)	($\leq$500, $\leq$2)	(10, 40)
S3	Small, Slowly Emerging	(500, 2)	(0, 0)	($\leq$500, $\leq$2)	(50, 2)

Some simulation results

	Initial SRS of Size $m = 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}_{t,t+1}^{panel}$)	0.20	0.20	0.14	0.28	0.28	0.14	0.28	0.28	0.13
RE($\hat{\nabla}_{t,t+1}^{pACS}$)	0.71	0.73	0.90	0.89	0.89	0.98	0.89	0.91	0.98
RE($\hat{\nabla}_{t,t+1}^{iACS}$)	0.57	0.60	0.52	0.69	0.70	0.52	0.84	0.85	0.75
	Initial SRS of Size $m = 5 \times 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}_{t,t+1}^{panel}$)	0.09	0.09	0.06	0.12	0.12	0.06	0.12	0.12	0.06
RE($\hat{\nabla}_{t,t+1}^{pACS}$)	0.09	0.11	0.38	0.62	0.63	0.87	0.65	0.64	0.91
RE($\hat{\nabla}_{t,t+1}^{iACS}$)	0.49	0.51	0.49	0.67	0.67	0.53	0.85	0.84	0.76
	Initial Size-biased Sampling ($\eta = 2$) of Size $m = 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}_{t,t+1}^{panel}$)	0.17	0.17	0.12	0.24	0.24	0.12	0.24	0.24	0.12
RE($\hat{\nabla}_{t,t+1}^{pACS}$)	0.31	0.33	0.41	0.51	0.51	0.50	0.52	0.53	0.51
RE($\hat{\nabla}_{t,t+1}^{iACS}$)	0.70	0.69	0.68	0.79	0.81	0.68	0.89	0.90	0.84

(pACS, iACS) complement each other, outperform traditional Panel design

Variations and challenges

Doubly adaptive: let ξ_{ij} be *strength* of $(ij) \in A$, and

$$\nu_i^* = \{j : (ij) \in A, \xi_{ij} > \xi_0\}$$

Given $y_i = 1$ with ν_i^* as neighbourhood (instead of ν_i)
Challenge: case ★ *(with* $y_i = 1$*) can be edge node... to be dealt with shortly...*

Adaptive snowball sampling, $T =$ maximum no. waves
Challenge: multiplicity/ancestry of sample cases ★

If only ★ $\in s_0$ and $T = 1$, observe ★ in s but not ★
What is $\pi_{(\bigstar)}$ under 1-wave adaptive snowball sampling?
To be dealt with under snowball sampling (SBS) later...

[1] Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

[2] Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.

[3] Zhang, L.-C. (2020). Sampling designs for epidemic prevalence estimation. `https://arxiv.org/abs/2011.08669`