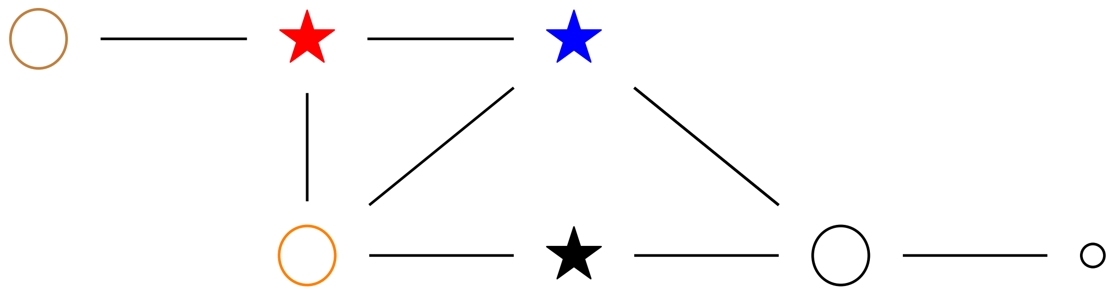
Day-1 Session-2: Bipartite Incidence Graph for Adaptive Cluster Sampling

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

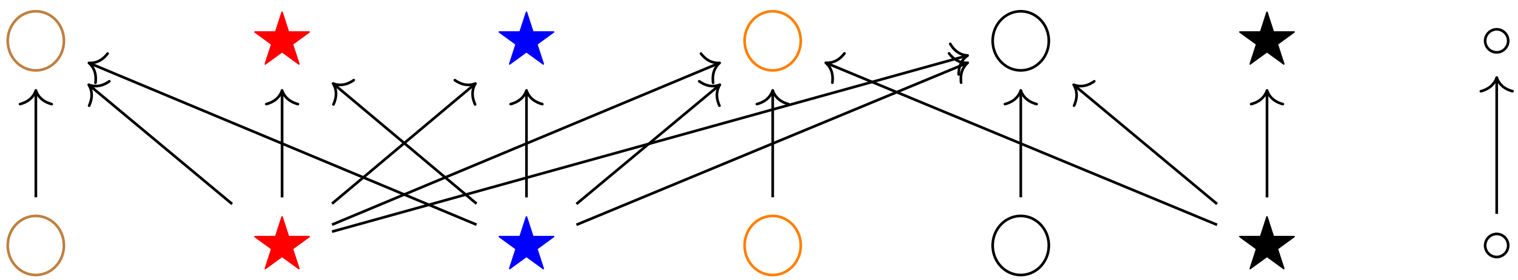
[1] *University of Southampton (L.Zhang@soton.ac.uk)*
[2] *Statistisk sentralbyraa, Norway*
[3] *Universitetet i Oslo*

Bipartite incidence graph (BIG)

ACS from $G = (U, A)$, a part of which as given below:



<u>All</u> the observational links in $\mathcal{B} = (F, \Omega; H)$ below:



$\mathcal{B}$ has simple directed edge set $H : F \to \Omega$, where F contains *samling units* and Ω the *study units*, and

$$(i\kappa) \in H \quad \text{only if} \quad \Pr(\kappa \in \Omega_s | i \in s_0) = 1$$

Multiplicity/ancestry under BIG sampling (BIGS)

The sampling units that can lead to a given $\kappa \in \Omega$ are

$$\beta_\kappa = \{i \in F : (i\kappa) \in H\}$$

the *ancestors* of κ under BIGS (Zhang, 2021; Zhang and Patone, 2017), similar to *multiplicity* (Birnbaum and Sirken, 1965) under indirect sampling.

Problem: ancestry knowledge β_κ for BIGS from $\mathcal{B}$ not always guaranteed under original sampling from G

- For any case node $\kappa \in \Omega$, ACS from G yields all its case network nodes β_κ which are also its ancestors under BIGS.
- For any noncase non-edge node κ, we have $\beta_\kappa = \{i_\kappa\}$ under ACS from G, which is always observed if κ is observed.
- For any edge node κ, like $\bigcirc$, ACS from G does not always yield β_κ which includes *all* its adjacent networks.

Thompson (1990): initial SRS and $|s_0| = 2$

A spatial graph G for ACS, with threshold $y_i \geq 5$:

$$1 \text{ --- } 0 \text{ --- } 2 \text{ --- } 10 \text{ --- } 1000$$

$\mathcal{B}$ including all the observation links to edge node 2:

$$\begin{tikzcd}
1 & 0 & 2 & 10 & 1000 \\
1 \arrow[u] & 0 \arrow[u] & 2 \arrow[u] & 10 \arrow[u] \arrow[ul] \arrow[ur] & 1000 \arrow[u] \arrow[ul] \arrow[ull]
\end{tikzcd}$$

<u>Modified HTE</u>: $\quad \hat{\theta}^*_{HT} = \sum_{\kappa \in \Omega_s} W_\kappa y_\kappa$

with $W_\kappa = \pi^{-1}_{(\kappa)}$ except for any *terminal* node where

$$W_\kappa = \begin{cases} \Pr(i_\kappa \in s_0)^{-1} & \text{if } i_\kappa \in s_0 \\ 0 & \text{otherwise} \end{cases}$$

such that $E\Big(\mathbb{I}(\kappa \in \Omega_s) W_\kappa\Big) \equiv 1$ for any $\kappa \in \Omega = U$

Modified BIG (Zhang and Oguz-Alper, 2020)

Remove corresponding links to 2 in $\mathcal{B}$ $\Rightarrow$ modified $\mathcal{B}^*$:

$$\begin{tikzcd}
1 & 0 & 2 & 10 & 1000 \\
1 \arrow[u] & 0 \arrow[u] & 2 \arrow[u] & 10 \arrow[u] \arrow[ur] & 1000 \arrow[u] \arrow[ul]
\end{tikzcd}$$

Under BIGS from $\mathcal{B}^*$, edge node 2 observed iff $2 \in s_0$

$$\text{HTE:} \qquad \hat{\theta}_{HT} = \sum_{\kappa \in \Omega_s} y_\kappa / \pi_{(\kappa)}$$

where $\pi_{(\kappa)}$ refers to $\Pr(\kappa \in \Omega_s)$ under BIGS from $\mathcal{B}^*$

<u>Strategies</u> ($\mathcal{B}$, MHT) and ($\mathcal{B}^*$, HT) yield always the same estimate but differ w.r.t. Rao-Blackwell (RB) method
NB. sampling strategy = (sampling method, estimator)

Numerical results for $\mu = \theta/N$

1 —— 0 —— 2 —— 10 —— 1000

	($\mathcal{B}$, MHT)		($\mathcal{B}^*$, HT)	
s_0	$\Omega_s = s$	$\hat{\mu}^*_{HT}$	Ω_s	$\hat{\mu}_{HT}$
1,0	1,0	0.500	1,0	0.500
1,2	1,2	1.500	1,2	1.500
0,2	0,2	1.000	0,2	1.000
1,10	1,10,*2*,1000	289.071	1,10,1000	289.071
1,1000	1,1000,*2*,10	289.071	1,1000,10	289.071
0,10	0,10,*2*,1000	288.571	0,10,1000	288.571
0,1000	0,1000,*2*,10	288.571	0,1000,10	288.571
2,10	2,10,1000	289.571	2,10,1000	289.571
2,1000	2,1000,10	289.571	2,1000,10	289.571
10,1000	10,1000,*2*	288.571	10,1000	288.571
Variance		17418.4		17418.4

Strategy ($\mathcal{B}$, MHT): the last three samples are all $\Omega_s = \{2, 10, 1000\}$, but $\hat{\mu}^*_{HT}$ differs because *2* is unused when $s_0 = \{10, 1000\}$. The RB method yields $E\big[\hat{\mu}^*_{HT} | \Omega_s = \{2, 10, 1000\}\big] = 289.238$.

Modified $\mathcal{B}^{\dagger}$, where 2 is only observed from $\{10, 1000\}$:

$$
\begin{tikzcd}
1 & 0 & 2 & 10 & 1000 \\
1 \arrow[u] & 0 \arrow[u] & 2 & 10 \arrow[ul] \arrow[u] \arrow[ur] & 1000 \arrow[ull] \arrow[ul] \arrow[u]
\end{tikzcd}
$$

$\mathcal{B}^{\dagger}$ identified with $\Pr(s_0 \cap \{10, 1000\} \neq \emptyset) = 0.7$, e.g. $s_0 = \{1, 10\}$

$$
\begin{tikzcd}
1 & 2 & 10 & 1000 & \\
1 \arrow[u] & 2 & 10 \arrow[ul] \arrow[u] \arrow[ur] & 1000 \arrow[ull] \arrow[ul] \arrow[u] &
\end{tikzcd}
\qquad \mathcal{B}_s^{\dagger}
$$

$\mathcal{B}^{*}$ needed with $\Pr(s_0 \cap \{2, 10, 1000\} = \{2\}) = 0.2$, e.g. $s_0 = \{0, 2\}$

$$
\begin{tikzcd}
0 & 2 \\
0 \arrow[u] & 2 \arrow[u]
\end{tikzcd}
\qquad \mathcal{B}_s^{*}
$$

Sample-dependent BIGS

1 —— 0 —— 2 —— 10 —— 1000

	$(\mathcal{B}^*$, HT)		$(\mathcal{B}^\dagger$, HT)	
s_0	Ω_s	$\hat{\mu}_{HT}$	Ω_s	$\hat{\mu}_{HT}$
1,0	1,0	0.500	1,0	0.500
1,2	1,2	1.500	1	0.500
0,2	0,2	1.000	0	0.000
1,10	1,10,1000	289.071	1,10,2,1000	289.643
1,1000	1,1000,10	289.071	1,1000,2,10	289.643
0,10	0,10,1000	288.571	0,10,2,1000	289.143
0,1000	0,1000,10	288.571	0,1000,2,10	289.143
2,10	2,10,1000	289.571	2,10,1000	289.143
2,1000	2,1000,10	289.571	2,1000,10	289.143
10,1000	10,1000	288.571	10,1000,2	289.143
Variance		17418.4		17533.7

Repeated ACS from G: indifferent choice if $s_0 = \{1, 0\}$, more repetitions 'needed' to decide; adopt $\mathcal{B}^*$ if 2 first sampled from $s_0 = \{1, 2\}$ or $\{0, 2\}$, or $\mathcal{B}^\dagger$ otherwise

Sample-dependent BIGS

Odds of adopting $\mathcal{B}^\dagger$ or $\mathcal{B}^*$ for edge node 2 is $7:2$

Unconditional inference impossible: 'first sample' decides

$\Downarrow$

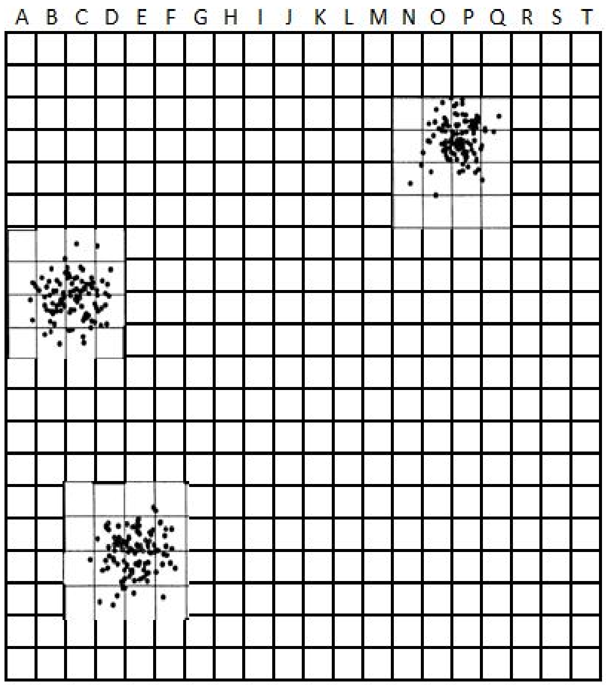
Conditional inference w.r.t. the adopted strategy

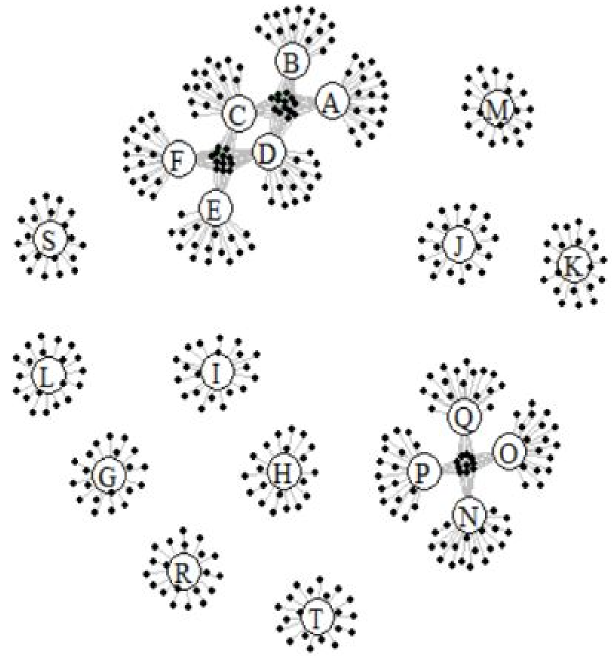
$$\begin{array}{ccccccccc} & & & & 20 & & & & \\ & & & & | & & & & \\ 1 & \text{———} & 0 & \text{———} & 2 & \text{———} & 10 & \text{———} & 1000 \end{array}$$

SRS, $|s_0| = 2$: probability $\frac{2}{15}$ for observing both networks $\{20\}$ and $\{10, 1000\}$, in which case one can e.g. let $\beta'_2 = \{20\}$ or $\beta'_2 = \{10, 1000\}$ or $\beta'_2 = \{20, 10, 1000\}$

Under ACS: $\{20\}$ and $\{10, 1000\}$ cannot be observed *from* each other

An example of two-stage ACS (Thompson, 1991)





1st-stage: 20 strips; 2nd-stage: neighbouring grids to any nonempty one, and so on
Thus, ACS applied at the second stage and terminated if no more non-empty grids
An edge grid is an empty grid that is contiguous to one or more non-empty grids
Modified $\mathcal{B}^*$ (Zhang and Oguz-Alper, 2020): F = strips, Ω = grids
10 star-like subgraphs, where an empty strip is adjacent to its 20 empty grids;
three clusters of non-empty grids; rest empty grids to 10 non-empty strips
An edge grid non-adjacent to neighbour strip, removing link under two-stage ACS

[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] Thompson, S. K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, 47:1103–1115.

[4] Zhang, L.-C. (2021). Graph sampling: An introduction. *The Survey Statistician*, 83:27-37.

[5] Zhang, L.-C. and Oguz-Alper, M. (2020). Bipartite incidence graph sampling. `https://arxiv.org/abs/2003.09467`

[6] Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, **75**, 277-299. `DOI:10.1007/s40300-017-0126-y`