Network Sampling

- Network data are themselves the primary object of interest
- Or is sampling over the network important?
- $G^* = (V^*, E^*)$ subgraph of G, $\eta(G)$ a summary statistic
- Goal: $\eta(G^*) \equiv \hat{\eta}$ how good? (e.g. # nodes, # edges, degree, clustering coeff)

Q: Plug-in good? A: No.

Ex: Induced subgraph sampling introduces bias in estimating average degree

↳ Correction $d_i^* = n d_i / N_V$

- Particularly interested in "design-based" inference.

Finite population

$\mathcal{U} = \{1, \dots, N_u\}$ associated with obs. y_i

Goal: $\mu = \frac{1}{N_u} \sum_{i=1}^{N_u} y_i$

Sample n units and construct $\mathbb{E}[\hat{\mu}] = \mu$

Idea: $\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{N_u} y_i \, \mathbb{E}[I(y_i \in S)] = \frac{1}{n} \sum_{i=1}^{N_u} y_i \pi_i$

$\Rightarrow$ unbiased iff $\pi_i = \frac{n}{N_u}$ $\Rightarrow$ In general top true

Horvitz-Thompson

An unbiased estimator $\hat{\mu}_\pi = \frac{1}{N_u}\hat{\tau}_\pi$, $\hat{\tau}_\pi = \sum_{i=1}^{N_v} \frac{y_i s_i}{\pi_i}$

Variance is included on the slides.

Estimation of Network Totals

Sampling Designs

- Induced subgraph (a) SRS of vertices (b) observe edges on induced subgraphs

$\pi_i = \frac{n}{N_v}$ $\pi_{\{i,j\}} = \frac{n(n-1)}{N_v(N_v-1)}$ (need to know # of vertices)

- Incident Subgraph (a) Edges SRS (b) observe vertices in induced subgraphs

$\pi_{(i,j)} = \frac{n}{N_E}$ $\pi_i = 1 - \mathbb{P}(\text{no edge incident to } i \text{ is sampled})$

$$= \begin{cases} 1 - \binom{N_e - d_i}{n} & n \le N_e - d_i \\ 1 & n > N_e - d_i \end{cases}$$

· Requires number of edges

- Snowball Sampling (a) SRS vertices (b) Sample subset of neighbours
  - Several different types
- Respondent Driven Sampling (coupons)

- Path sampling and traceroute