Sampling and Estimation in Network Graphs

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

March 27, 2020

Network sampling

Network sampling and challenges

Background on statistical sampling theory

Network graph sampling designs

Estimation of network totals and group size

Estimation of degree distributions

Sampling network graphs

- Measurements often gathered only from a portion of a complex system
  - Ex: social study of high-school class vs. large corporation, Internet
  - Network graph $\rightarrow$ sample from a larger underlying network
- Goal: use sampled network data to infer properties of the whole system
  - Approach using principles of statistical sampling theory
- Sampling in network contexts introduces various potential challenges

System under study		Available measurements
$G(V, E)$	$\xrightarrow{\text{Random Procedure}}$	$G^*(V^*, E^*)$
Population graph		Sampled graph

- G^* often a subgraph of G (i.e., $V^* \subseteq V$, $E^* \subseteq E$), but may not be

The fundamental problem

- Suppose a given graph characteristic or summary $\eta(G)$ is of interest
  - Ex: order N_v, size N_e, degree d_v, clustering coefficient $\mathrm{cl}(G)$, ...
- Typically impossible to recover $\eta(G)$ exactly from G^*

  $\Rightarrow$ Q: Can we still form a useful estimate $\hat{\eta} = \hat{\eta}(G^*)$ of $\eta(G)$?
- Plug-in estimator $\hat{\eta} := \eta(G^*)$
  - Boils down to computing the characteristic of interest in G^*
  - Many familiar estimators in statistical practice are of this type

    Ex: sample means, standard deviations, covariances, quantiles...
- Oftentimes $\eta(G^*)$ is a poor representation of $\eta(G)$

Example: Estimating average degreee

- Let $G(V, E)$ be a network of protein interactions in yeast

  $\Rightarrow$ Characteristic of interest is average degree

$$\eta(G) = \frac{1}{N_v} \sum_{i \in V} d_i$$

- Here $N_v = 5,151$, $N_e = 31,201 \Rightarrow \eta(G) = 12.115$

- Consider two sampling designs to obtain G^*
  - First sample n vertices $V^* = \{i_1, \ldots, i_n\}$ **without replacement**
  - Design 1: For each $i \in V^*$, observe incident edges $(i, j) \in E$
  - Design 2: Observe edge (i, j) only when both $i, j \in V^*$

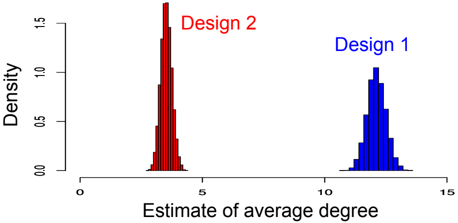
- Estimate $\eta(G)$ by averaging the observed degree sequence $\{d_i^*\}_{i \in V^*}$

$$\eta(G^*) = \frac{1}{n} \sum_{i \in V^*} d_i^*$$

Example: Estimating average degreee (cont.)

- Random sample of $n = 1,500$ vertices, Designs 1 and 2 for edges

  $\Rightarrow$ Process repeated for 10,000 trials $\Rightarrow$ histogram of $\eta(G^*)$



- Under-estimate $\eta(G)$ for Design 2, but Design 1 on target. Why?
  - Design 1: sample vertex degree explicitly, i.e., $d_i^* = d_i$
  - Design 2: (implicitly) sample vertex degree with bias, i.e., $d_i^* \approx \frac{n}{N_v} d_i$

Improving estimation accuracy

- In order to do better we need to incorporate the effects of
  - $\Rightarrow$ Random sampling; and/or
  - $\Rightarrow$ Measurement error
- Sampling design, topology of G, nature of $\eta(\cdot)$ all critical
- Model-based inference $\rightarrow$ Likelihood-based and Bayesian paradigms
- Design-based methods $\rightarrow$ Statistical sampling theory
  - Assume observations made without measurement error
  - Only source of randomness $\rightarrow$ sampling procedure
- Ex: Estimating average degree
  - Under Design 2 the estimate is biased, with mean of only 3.528
  - Adjusting $\eta(G^*)$ upward by a factor $\frac{N_v}{n} = 3.434$ yields 12,115
- Will see how statistical sampling theory justifies this correction

Background

Network sampling and challenges

Background on statistical sampling theory

Network graph sampling designs

Estimation of network totals and group size

Estimation of degree distributions

Statistical sampling theory

- Suppose we have a population $\mathcal{U} = \{1, \ldots, N_u\}$ of N_u units
  - Ex: People, animals, objects, vertices, ...
- A value y_i is associated with each unit $i \in \mathcal{U}$
  - Ex: Height, age, gender, infected, membership, ...
- Typical interest in the population **totals** τ and **averages** μ

$$\tau := \sum_{i \in \mathcal{U}} y_i \quad \text{and} \quad \mu := \frac{1}{N_u} \sum_{i \in \mathcal{U}} y_i = \frac{1}{N_u} \tau$$

- Basic sampling theory paradigm oriented around these steps:

S1: Randomly sample n units $\mathcal{S} = \{i_1, \ldots, i_n\}$ from $\mathcal{U}$

S2: Observe the value y_{i_k} for $k = 1, \ldots, n$

S3: Form an unbiased estimator $\hat{\mu}$ of μ, i.e., $\mathbb{E}[\hat{\mu}] = \mu$

S4: Evaluate or estimate the variance $\text{var}[\hat{\mu}]$

Inclusion probabilities

- **Def:** For given sampling design, the inclusion probability π_i of unit i is

$$\pi_i := \mathrm{P}\left(\text{unit } i \text{ belongs in the sample } \mathcal{S}\right)$$

- Simple random sampling (SRS): n units sampled uniformly form $\mathcal{U}$

  Without replacement: i_1 chosen from $\mathcal{U}$, i_2 from $\mathcal{U} \setminus \{i_1\}$, and so on

  $\Rightarrow$ There are $\binom{N_u}{n}$ such possible samples of size n

  $\Rightarrow$ There are $\binom{N_u-1}{n-1}$ samples which include a given unit i

- The inclusion probability is

$$\pi_i = \frac{\binom{N_u-1}{n-1}}{\binom{N_u}{n}} = \frac{n}{N_u}$$

Sample mean estimator

- Definition of sample mean estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i$$

- Using indicator RVs $\mathbb{I}\{i \in \mathcal{S}\}$ for $i \in \mathcal{U}$, where $\mathbb{E}\left[\mathbb{I}\{i \in \mathcal{S}\}\right] = \pi_i$

$$\Rightarrow \mathbb{E}\left[\hat{\mu}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}} y_i\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N_u} y_i \mathbb{I}\{i \in \mathcal{S}\}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{N_u} y_i \mathbb{E}\left[\mathbb{I}\{i \in \mathcal{S}\}\right] = \frac{1}{n}\sum_{i=1}^{N_u} y_i \pi_i$$

- SRS without replacement $\rightarrow$ unbiased because $\pi_i = \frac{n}{N_u}$
- Unequal probability sampling
  - More common than SRS, especially with networks. (More soon)
  - Sample mean can be a poor (i.e., biased) estimator for μ

Horvitz-Thompson estimation for totals

- Idea: weighted average using inclusion probabilities as weights

Horvitz-Thompson (HT) estimator

$$\hat{\mu}_\pi = \frac{1}{N_u} \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} \quad \text{and} \quad \hat{\tau}_\pi = N_u \hat{\mu}_\pi$$

- Remedies the bias problem

$$\mathbb{E}\left[\hat{\mu}_\pi\right] = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{y_i}{\pi_i} \mathbb{E}\left[\mathbb{I}\left\{i \in \mathcal{S}\right\}\right] = \frac{1}{N_u} \sum_{i=1}^{N_u} y_i = \mu$$

$\Rightarrow$ Size of the population N_u assumed known

$\Rightarrow$ Broad applicability, but π_i may be difficult to compute

Horvitz-Thompson estimator variance

- **Def:** Joint inclusion probability π_{ij} of units i and j is

$$\pi_{ij} := \mathrm{P}\left(\text{units } i \text{ and } j \text{ belong in the sample } \mathcal{S}\right)$$

- If inclusion of units i and j are independent events $\Rightarrow \pi_{ij} = \pi_i \pi_j$
- Ex: Simple random sampling without replacement yields

$$\pi_{ij} = \frac{n(n-1)}{N_u(N_u - 1)}$$

- Variance of the HT estimator:

$$\mathrm{var}\left[\hat{\tau}_\pi\right] = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right), \quad \mathrm{var}\left[\hat{\mu}_\pi\right] = \frac{\mathrm{var}\left[\hat{\tau}_\pi\right]}{N_u^2}$$

$\Rightarrow$ Typically estimated in an unbiased fashion from the sample $\mathcal{S}$

Probability proportional to size sampling

- Unequal probability sampling

  $\Rightarrow$ n units selected w.r.t. a distribution $\{p_1, \ldots, p_{N_u}\}$ on $\mathcal{U}$

  $\Rightarrow$ Uniform sampling: special case with $p_i = \frac{1}{N_u}$ for all $i \in \mathcal{U}$

- Probability proportional to size (PPS) sampling

  $\Rightarrow$ Probabilities p_i proportional to a characteristic c_i

  Ex: households chosen by drawing names from a database

- If sampling with replacement, PPS inclusion probabilities are

$$\pi_i = 1 - (1 - p_i)^n, \text{ where } p_i = \frac{c_i}{\sum_k c_k}$$

- Joint inclusion probabilities for variance calculations

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$$

Estimation of group size

- So far implicitly assumed N_u known $\rightarrow$ Often not the case!

  Ex: endangered animal species, people at risk of rare disease

- Special population total often of interest is the group size

$$N_u = \sum_{i \in \mathcal{U}} 1$$

- Suggests the following HT estimator of N_u

$$\hat{N}_u = \sum_{i \in \mathcal{S}} \pi_i^{-1}$$

$\Rightarrow$ Infeasible, since knowledge of N_u needed to compute π_i

Capture-recapture estimator

- Capture-recapture estimators overcome HT limitations in this setting
- Two rounds of SRS without replacement $\Rightarrow$ Two samples $\mathcal{S}_1$, $\mathcal{S}_2$

Round 1: Mark all units in sample $\mathcal{S}_1$ of size n_1 from $\mathcal{U}$

- Ex: tagging a fish, noting the ID number...
- All units in $\mathcal{S}_1$ are returned to the population

Round 2: Obtain a sample $\mathcal{S}_2$ of size n_2 from $\mathcal{U}$

Capture-recapture estimator of N_u

$$\hat{N}_u := \frac{n_2}{m} n_1, \quad \text{where } m := |\mathcal{S}_1 \cap \mathcal{S}_2|$$

- Factor m/n_2 indicative of marked fraction of the overall population

$\Rightarrow$ Can derive using model-based arguments as an ML estimator

Common network graph sampling designs

Network sampling and challenges

Background on statistical sampling theory

Network graph sampling designs

Estimation of network totals and group size

Estimation of degree distributions

Graph sampling designs

- Q: What are common designs for sampling a network graph G?
- A: Will see a few examples, along with their inclusion probabilities π_i
- Graph-based sampling designs

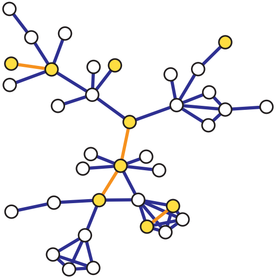
  $\Rightarrow$ Two inter-related classes of units, vertices i and edges (i, j)
- Often two stages
  - Selection among one class of units (e.g., vertices)
  - Observation of units from the other class (e.g., edges)
- Inclusion probabilities offer insight into the nature of the designs

  $\Rightarrow$ Central to HT estimators of network graph characteristics $\eta(G)$

Induced subgraph sampling

S) Sample n vertices $V^* = \{i_1, \ldots, i_n\}$ without replacement (SRS)

O) Observe edges $(i,j) \in E^*$ only when both $i,j \in V^*$ (induced by V^*)



- Ex: construction of contact networks in social network research
- Vertex and edge inclusion probabilities are uniformly equal to

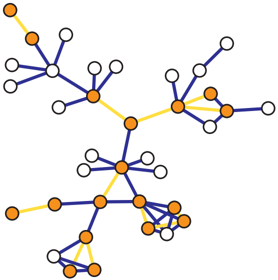
$$\pi_i = \frac{n}{N_v} \quad \text{and} \quad \pi_{\{i,j\}} = \frac{n(n-1)}{N_v(N_v-1)}$$

Incident subgraph sampling

- Consider a complementary design to induced subgraph sampling

S) Sample n edges E^* without replacement (SRS)

O) Observe vertices $i \in V^*$ incident to those selected edges in E^*



- Ex: construction of sampled telephone call graphs

Inclusion probabilities

- For incident subgraph sampling, edge inclusion probabilities are

$$\pi_{\{i,j\}} = \frac{n}{N_e}$$

- Vertex in V^* if any one or more of its incident edges are sampled

$$\begin{aligned} \pi_i &= \mathrm{P}\left(\text{vertex } i \text{ is sampled}\right) \\ &= 1 - \mathrm{P}\left(\text{no edge incident to } i \text{ is sampled}\right) \\ &= \begin{cases} 1 - \frac{\binom{N_e - d_i}{n}}{\binom{N_e}{n}}, & \text{if } n \leq N_e - d_i \\ 1, & \text{if } n > N_e - d_i \end{cases} \end{aligned}$$

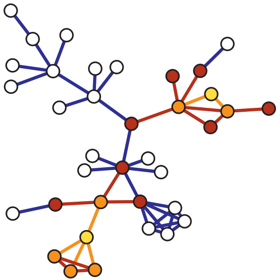
- Vertices included with unequal probs. that depend on their degrees
  - $\Rightarrow$ Probability proportional to size (degree) sampling of vertices
  - $\Rightarrow$ Requires knowledge of N_e and degree sequence $\{d_i\}_{i \in V^*}$

Snowball sampling

S) Sample n vertices $V_0^* = \{i_1, \ldots, i_n\}$ without replacement (SRS)

O1) Observe edges E_0^* incident to each $i \in V_0^*$, forming the initial wave

O2) Observe neighbors $\mathcal{N}(V_0^*)$ of $i \in V_0^*$, i.e., $V_1^* = \mathcal{N}(V_0^*) \cap (V_0^*)^c$



- Iterate to a desired number of e.g., k waves, or until V_k^* empty

  $\Rightarrow$ G^* has $V^* = V_0^* \cup V_1^* \cup \ldots \cup V_k^*$, and their incident edges

- Ex: ‘spiders’ or ‘crawlers’ to discover the WWW’s structure

Star sampling

- Difficult to compute inclusion probabilities beyond a single wave
  - $\Rightarrow$ Single-wave snowball sampling reduces to star sampling
- Unlabeled: $V^* = V_0^*$ and $E^* = E_0^*$ their incident edges
  - Ex: Count all co-authors of n sampled authors
  - Vertex inclusion probabilities are simply $\pi_i = n/N_v$
- Labeled: $V^* = V_0^* \cup (\mathcal{N}(V_0^*) \cap (V_0^*)^c)$ and $E^* = E_0^*$
  - Ex: Count and identify all co-authors of n sampled authors
  - Vertex inclusion probabilities can be shown to look like

$$\pi_i = \sum_{L \subseteq \mathcal{N}_i} (-1)^{|L|+1} \mathrm{P}(L), \quad \text{where } \mathrm{P}(L) = \frac{\binom{N_v - |L|}{n - |L|}}{\binom{N_v}{n}}$$

  - Denoted by $\mathcal{N}_i$ the neighborhood of vertex i (including i itself)

Link tracing

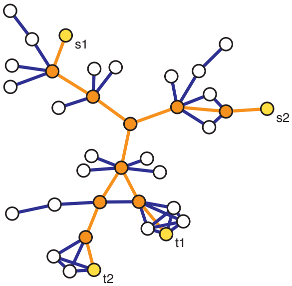
- Link-tracing designs
  - $\Rightarrow$ Select an initial sample of vertices V_S^*
  - $\Rightarrow$ Trace edges (links) from V_s^* to another set of vertices V_T^*
- Snowball sampling: special case where all incident edges are traced
- May be infeasible to follow all incident edges to a given vertex

  Ex: lack of recollection/deception in social contact networks
- Path sampling designs
  - $\Rightarrow$ Source nodes $V_S^* = \{s_1, \ldots, s_{n_S}\} \subset V$
  - $\Rightarrow$ Target nodes $V_T^* = \{t_1, \ldots, t_{n_T}\} \subset V \setminus V_S^*$
  - $\Rightarrow$ Traverse and measure the path between each pair (s_i, t_j)

  Ex: `Traceroute` Internet studies, Milgram's "Six Degrees" experiment

Traceroute sampling

- Trace shortest paths from each source to all targets



- Vertex and edge inclusion probabilities roughly [Dall'Asta et al '06]:

$$\pi_i \approx 1 - (1 - \rho_S - \rho_T) e^{-\rho_S \rho_T c_{Be}(i)} \text{ and } \pi_{\{i,j\}} \approx 1 - e^{-\rho_S \rho_T c_{Be}(\{i,j\})}$$

- Source and target sampling fractions $\rho_S := n_S / N_v$ and $\rho_T := n_T / N_v$
  $\Rightarrow$ Induces PPS sampling, size given by betweenness centralities

Estimation of totals in network graphs

Network sampling and challenges

Background on statistical sampling theory

Network graph sampling designs

Estimation of network totals and group size

Estimation of degree distributions

Network summaries as totals

- Various graph summaries $\eta(G)$ are expressible in terms of totals τ

  Average degree: Let $\mathcal{U} = V$ and $y_i = d_i$, then $\eta(G) = \bar{d} \propto \sum_{i \in V} d_i$

  Graph size: Let $\mathcal{U} = E$ and $y_{ij} = 1$, then $\eta(G) = N_e = \sum_{(i,j) \in E} 1$

  Betweenness centrality: Let $\mathcal{U} = V^{(2)}$ (unordered vertex pairs) and $y_{ij} = \mathbb{I}\left\{k \in \mathcal{P}_{(i,j)}\right\}$. For unique shortest $i - j$ paths $\mathcal{P}_{(i,j)}$, then

$$\eta(G) = c_{Be}(k) = \sum_{(i,j) \in V^{(2)}} \mathbb{I}\left\{k \in \mathcal{P}_{(i,j)}\right\}$$

  Clustering coefficient: Let $\mathcal{U} = V^{(3)}$ (unordered vertex triples), then

$$\eta(G) = \mathrm{cl}(G) = 3 \times \frac{\text{total number of triangles}}{\text{total number of connected triples}}$$

- Often such totals can be obtained from sampled G^* via HT estimation

Vertex totals

- Vertex totals are of the form $\tau = \sum_{i \in V} y_i$, averages are τ / N_v
  - Ex: average degree where $y_i = d_i$
  - Ex: nodes with characteristic $\mathcal{C}$, where $y_i = \mathbb{I}\{i \in \mathcal{C}\}$
- Given a sample $V^* \subseteq V$, the HT estimator for vertex totals is

$$\hat{\tau}_\pi = \sum_{i \in V^*} \frac{y_i}{\pi_i}$$

$\Rightarrow$ Variance expressions carry over, let $\mathcal{U} = V$ and V^* for estimates

- Inclusion probabilities π_i depend on network sampling design

$\Rightarrow$ Sampling also influences whether y_i is observable, e.g., $y_i = d_i$

Totals on vertex pairs

- Quantity y_{ij} corresponding to vertex pairs $(i,j) \in V^{(2)}$ of interest

  $\Rightarrow$ Totals $\tau = \sum_{(i,j)\in V^{(2)}} y_{ij}$ become relevant

  - Ex: graph size N_e and betweenness $c_{Be}(k)$ where $y_{ij} = \mathbb{I}\left\{k \in \mathcal{P}_{(i,j)}\right\}$
  - Ex: shared gender in friendship network, average dissimilarity

- The HT estimator in this context is

$$\hat{\tau}_\pi = \sum_{(i,j)\in V^{(2)*}} \frac{y_{ij}}{\pi_{ij}}$$

  $\Rightarrow$ Edge totals a special case, when $y_{ij} \neq 0$ only for $(i,j) \in E$

- Variance expression increasingly complicated, namely

$$\text{var}\left[\hat{\tau}_\pi\right] = \sum_{(i,j)\in V^{(2)}} \sum_{(k,l)\in V^{(2)}} y_{ik} y_{kl} \left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - 1\right)$$

  $\Rightarrow$ Depends on inclusion probabilities π_{ijkl} of vertex quadruples

Example: Estimating network size

- Consider estimating N_e as an edge total, i.e.,

$$N_e = \sum_{(i,j)\in E} 1 = \sum_{(i,j)\in V^{(2)}} A_{ij}$$

- Bernoulli sampling (BS): $\mathbb{I}\{i \in V^*\} \sim \text{Ber}(p)$ i.i.d. for all $i \in V$
  $\Rightarrow$ Edges E^* obtained via induced subgraph sampling $\Rightarrow \pi_{ij} = p^2$
- The HT estimator of N_e is

$$\hat{N}_e = \sum_{(i,j)\in V^{(2)*}} \frac{A_{ij}}{\pi_{ij}} = p^{-2} N_e^*$$

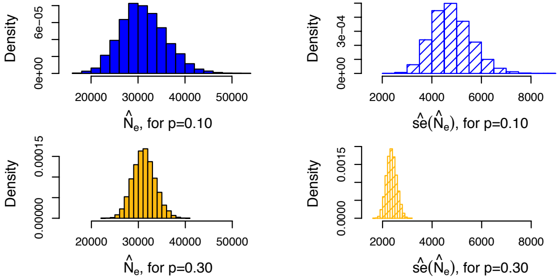
  $\Rightarrow$ Scales up the empirically observed edge total N_e^* by $p^{-2} > 1$

- Variance can be shown to take the form [Frank '77]

$$\text{var}\left[\hat{N}_e\right] = (p^{-1} - 1)\sum_{i\in V} d_i^2 + (p^{-2} - 2p^{-1} + 1)N_e$$

Example: Estimating network size (cont.)

- Protein network: $N_v = 5,151$, $N_e = 31,201$
  - $\Rightarrow$ BS of vertices with $p = 0.1$ and $p = 0.3$
  - $\Rightarrow$ Process repeated for 10,000 trials $\Rightarrow$ histogram of $\hat{N}_e$



- Average of $\hat{N}_e$ was $31,116$ and $31,203$ $\Rightarrow$ Unbiasedness supported
  - $\Rightarrow$ Mean and variability of $\hat{se}$ shrinks with p (larger sample)

Example: Estimating clustering coefficient

- Average clustering coefficient $\text{cl}(G)$ can be expressed as

$$\text{cl}(G) = 3 \times \frac{\tau_{\triangle}(G)}{\tau_3(G)}$$

- Involves the quotient of two totals on vertex triples

$$\tau = \sum_{(i,j,k) \in V^{(3)}} y_{ijk} \Rightarrow \hat{\tau}_\pi = \sum_{(i,j,k) \in V^{(3)*}} \frac{y_{ijk}}{\pi_{ijk}}$$

- Total number of triangles $\tau_{\triangle}(G)$, where

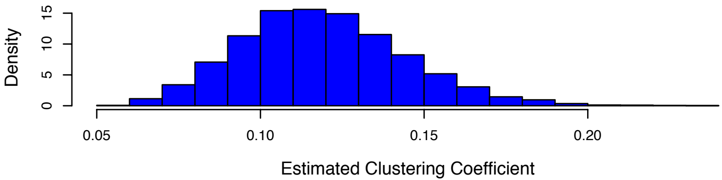
$$y_{ijk} = A_{ij} A_{jk} A_{ki}$$

- Total number of connected triples $\tau_3(G)$, where

$$y_{ijk} = A_{ij} A_{jk}(1 - A_{ki}) + A_{ij}(1 - A_{jk}) A_{ki} + (1 - A_{ij}) A_{jk} A_{ki}$$

Example: Estimating clustering coefficient (cont.)

- Protein network: $\tau_\triangle(G) = 44,858$, $\tau_3(G) \approx 1M$, and $\text{cl}(G) = 0.1179$
  - $\Rightarrow$ BS of vertices with $p = 0.2$
  - $\Rightarrow$ Induced subgraph sampling of edges
  - $\Rightarrow$ Process repeated for 10,000 trials $\Rightarrow$ histogram of $\hat{\text{cl}}(G)$



- Unbiased HT estimators $\hat{\tau}_\triangle = p^{-3}\tau_\triangle(G^*)$ and $\hat{\tau}_3 = p^{-3}\tau_3(G^*)$
  - $\Rightarrow$ Plug-in estimator $\hat{\text{cl}}(G) = 3\hat{\tau}_\triangle/\hat{\tau}_3$ results in $\hat{\text{cl}}(G) = \text{cl}(G^*)$
  - $\Rightarrow$ Quite accurate with mean 0.1191 and $\hat{\text{se}}$ of 0.0251

Caveat emptor

- Horvitz-Thompson framework fairly straightforward in its essence
- Success in network sampling and estimation rests on interaction among
  a) Sampling design;
  b) Measurements taken; and
  c) Total to be estimated
- Three basic elements must be present in the problem
  1) Network summary statistic $\eta(G)$ expressible as total;
  2) Values y either observed, or obtainable from measurements; and
  3) Inclusion probabilities π computable for the sampling design
- Unfortunately, often not all three are present at the same time ...

Example: Estimating average degreee

- Recall our first example on estimation of average degree $\frac{1}{N_v}\sum_{i \in V} d_i$
  - Design 1: Unlabeled star sampling, observes degrees d_i, $i \in V^*$
  - Design 2: Induced subgraph sampling, does not observe degrees
- Average degree is a scaling of a vertex total (N_v known)

  $\Rightarrow$ HT estimation applicable so long as $y_i = d_i$ observed
- True for unlabeled star sampling, and since $\pi_i = n/N_v$ we have

$$\hat{\mu}_{St} = \frac{\hat{\tau}_{St}}{N_v}, \quad \text{where } \hat{\tau}_{St} = \sum_{i \in V^*_{St}} \frac{d_i}{n/N_v}$$

- We do not observe d_i under induced subgraph sampling

  $\Rightarrow$ Not amenable to HT estimation as vertex total for this design

Example: Estimating average degreee (cont.)

- Identity $\mu = \frac{2N_e}{N_v} \Rightarrow$ Tackle instead as estimation of network size N_e
- For induced subgraph sampling $\pi_{ij} = \frac{n(n-1)}{N_v(N_v-1)}$, so HT estimator is

$$\hat{N}_{e,IS} = \sum_{(i,j)\in V^{(2)*}} \frac{A_{ij}}{n(n-1)/[N_v(N_v-1)]} = \frac{N_v(N_v-1)}{n(n-1)} N^*_{e,IS}$$

$\Rightarrow$ Desired unbiased estimator for the average degree is

$$\hat{\mu}_{IS} = \frac{2\hat{N}_{e,IS}}{N_v}$$

- Estimators under both designs can be compared by writing them as

$$\hat{\mu}_{St} = \frac{2N^*_{e,St}}{n} \text{ and } \hat{\mu}_{IS} = \frac{2N^*_{e,IS}}{n} . \frac{N_v - 1}{n-1}$$

$\Rightarrow$ Design 1: uses the identity $\mu = \frac{2N_e}{N_v}$ on G^*_{St}

$\Rightarrow$ Design 2: same but inflated by $\frac{N_v-1}{n-1}$, compensates $d^*_{i,IS} < d_i$

Estimation of network group size

- Assuming that N_v is known may not be on safe grounds
  - $\Rightarrow$ Human or animal groups too mobile or elusive to count accurately
  - $\Rightarrow$ All Web pages or Internet routers are too massive and dispersed
- Often estimating N_v may well be the prime objective
- If vertex SRS or BS feasible, could sample twice 'marking' in between
  - $\Rightarrow$ Facilitates usage of capture-recapture estimators 'off-the-shelf'
- If sampling infeasible, or capture-recapture performs poorly
  - $\Rightarrow$ Develop estimators of N_v tailored to the graph sampling at hand

Estimating the size of a "hidden population"

- Hidden population: individuals do not wish to expose themselves
  - Ex: humans of socially sensitive status, such as homeless
  - Ex: involved in socially sensitive activities, e.g., drugs, prostitution
- Such groups are often small $\Rightarrow$ Estimating their size is challenging
- Snowball sampling used to estimate the size of hidden populations
- O. Frank and T. Snijders, "Estimating the size of hidden populations using snowball sampling," *J. Official Stats.*, vol. 10, pp. 53-67, 1994

Sampling a hidden population

- Directed graph $G(V, E)$, V the members of the hidden population
  - $\Rightarrow$ Graph describing willingness to identify other members
  - $\Rightarrow$ Arc (i, j) when ask individual i, mentions j as a member
- Graph G^* obtained via one-wave snowball sampling, i.e., $V^* = V_0^* \cup V_1^*$
  - $\Rightarrow$ Initial sample V_0^* obtained via BS from V with probability p_0
- Consider the following random variables (RVs) of interest
  - $N = |V_0^*|$: size of the initial sample
  - M_1: number of arcs among individuals in V_0^*
  - M_2: number of arcs from individuals in V_0^* to individuals in V_1^*
- Snowball sampling yields measurements n, m_1, and m_2 of these RVs

Method of moments estimator

- Method of moments: equate moments to sample counterparts

$$\mathbb{E}\left[N\right] = \mathbb{E}\left[\sum_i \mathbb{I}\left\{i \in V_0^*\right\}\right] = N_v p_0 = n$$

$$\mathbb{E}\left[M_1\right] = \mathbb{E}\left[\sum_j \sum_{i \neq j} \mathbb{I}\left\{i \in V_0^*\right\} \mathbb{I}\left\{j \in V_0^*\right\} A_{ij}\right] = N_e p_0^2 = m_1$$

$$\mathbb{E}\left[M_2\right] = \mathbb{E}\left[\sum_j \sum_{i \neq j} \mathbb{I}\left\{i \in V_0^*\right\} \mathbb{I}\left\{j \notin V_0^*\right\} A_{ij}\right] = N_e p_0 (1 - p_0) = m_2$$

- Expectation w.r.t. randomness in selecting the sample V_0^*. Solution:

$$\hat{N}_v = n \left(\frac{m_1 + m_2}{m_1}\right)$$

$\Rightarrow$ Size of initial sample inflated by estimate of the sampling rate

Estimation of degree distributions

Network sampling and challenges

Background on statistical sampling theory

Network graph sampling designs

Estimation of network totals and group size

Estimation of degree distributions

Estimation of other network characteristics

- Classical sampling theory rests heavily on Horvitz-Thompson framework

  $\Rightarrow$ Scope limited to network totals

  $\Rightarrow$ Q: Other network summaries, e.g., degree distributions?

- Findings on the effect of sampling on observed degree distributions:
  - Highly unrepresentative of actual degree distributions; and
  - Unhelpful to characterizing heterogeneous distributions

- Ex: Internet `traceroute` sampling [Lakhina et al' 03]

  $\Rightarrow$ Broad degree distribution in G^*, while concentrated in G

- Ex: Sampling protein-protein interaction networks [Han et al' 05]

  $\Rightarrow$ Power-law exponent estimate from G^* underestimates α in G

Impact of sampling on degree distribution

- Let $N(d)$ denote the number of vertices with degree d in G
  - $\Rightarrow$ Let $N^*(d)$ be the counterpart in a sampled graph G^*
  - $\Rightarrow$ Introduce vectors $\mathbf{n} = [N(0), \ldots, N(d_{\text{max}})]^\top$ and likewise $\mathbf{n}^*$
- Under a variety of sampling designs, it holds that

$$\mathbb{E}\left[\mathbf{n}^*\right] = \mathbf{P}\mathbf{n}$$

  - $\Rightarrow$ Matrix $\mathbf{P}$ depends fully on the sampling, not G itself
  - $\Rightarrow$ Expectation w.r.t. randomness in selecting the sample G^*
- O. Frank, "Estimation of the number of vertices of different degrees in a graph," *J. Stat. Planning and Inference,* vol. 4, pp. 45-50, 1980

An inverse problem

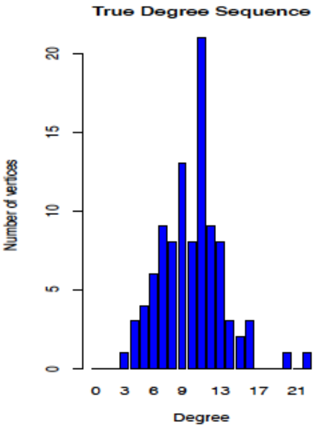
- Recall the identity $\mathbb{E}[\mathbf{n}^*] = \mathbf{P}\mathbf{n}$ $\Rightarrow$ Face a linear inverse problem
- Unbiased estimator of the degree distribution $\mathbf{n}$

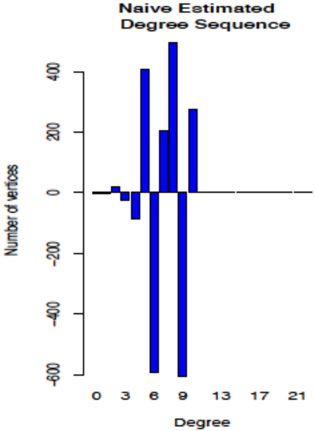
$$\hat{\mathbf{n}}_{\text{naive}} = \mathbf{P}^{-1}\mathbf{n}^*$$

- While natural, two problems with this simple solution
  - $\Rightarrow$ Matrix $\mathbf{P}$ typically not invertible in practice; and
  - $\Rightarrow$ Non-negativity of the solution is not guaranteed
- We actually have an ill-posed linear inverse problem

Performance of naive estimator

- Erdös-Renyi graph with $N_v = 100$ and $N_e = 500$
  - $\Rightarrow$ BS of vertices with $p = 0.6$
  - $\Rightarrow$ Induced subgraph sampling of edges





Penalized least-squares formulation

- Constrained, penalized, weighted least-squares [Zhang et al '14]

$$\min_{\mathbf{n}} \; (\mathbf{P}\mathbf{n} - \mathbf{n}^*)^\top \mathbf{C}^{-1} (\mathbf{P}\mathbf{n} - \mathbf{n}^*) + \lambda \text{pen}(\mathbf{n})$$
$$\text{s. to} \;\; N(d) \geq 0, \; d = 0, 1, \ldots, d_{\max},$$
$$\sum_{d=1}^{d_{\max}} N(d) = N_v$$

$\Rightarrow$ Matrix $\mathbf{C}$ denotes the covariance of $\mathbf{n}^*$

$\Rightarrow$ Functional pen($\mathbf{n}$) penalizes complexity in $\mathbf{n}$, tuned by λ

- Constraints

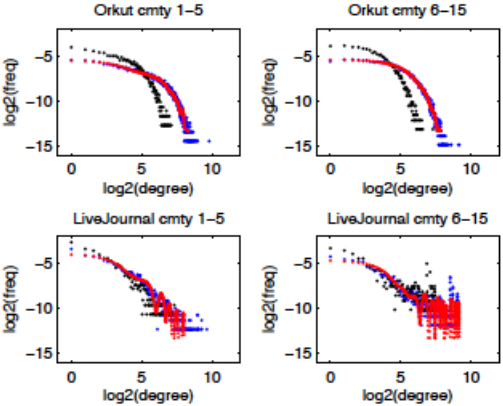
$\Rightarrow$ Non-negativity of degree counts

$\Rightarrow$ Total degree counts equal the number of vertices

$\Rightarrow$ Smoothness: pen($\mathbf{n}$) $= \|\mathbf{D}\mathbf{n}\|^2$, $\mathbf{D}$ differentiating operator

Application to online social networks

- Communities from online social networks Orkut and LiveJournal
  - $\Rightarrow$ BS of vertices with $p = 0.3$
  - $\Rightarrow$ Induced subgraph sampling of edges



- True, sampled, and estimated degree distribution

Glossary

- Enumeration and samping
- Population graph
- Sampled graph
- Plug-in estimator
- Sampling design
- Sample with(out) replacement
- Design-based methods
- Averages and totals
- Inclusion probability
- Simple random sampling
- Bernoulli sampling
- Unequal probability sampling
- Horvitz-Thompson estimator
- Probability proportional to size sampling
- Capture-recapture estimator
- Induced subgraph sampling
- Incident subgraph sampling
- Snowball and star sampling
- Traceroute sampling
- Hidden population
- Ill-posed inverse problem
- Penalized least squares