

# *Statistical Analysis of Network Data: II*

Eric D. Kolaczyk

`kolaczyk@math.bu.edu`

Dept of Mathematics and Statistics, Boston University



## Point of Departure . . .

---

Common *modus operandi* in network analysis:

- System of elements and their interactions is of interest.
- Collect elements and relations among elements.
- Represent the collected data via a network.
- Characterize properties of the network.

## Point of Departure . . .

---

Common *modus operandi* in network analysis:

- System of elements and their interactions is of interest.
- Collect elements and relations among elements.
- Represent the collected data via a network.
- Characterize properties of the network.

Sounds good . . . so what?

## Interpretation: Two Scenarios

---

With respect to what frame of reference are the network characteristics interpreted?

## Interpretation: Two Scenarios

---

With respect to what frame of reference are the network characteristics interpreted?

1. The collected network data are themselves the primary object of interest.

## Interpretation: Two Scenarios

---

With respect to what frame of reference are the network characteristics interpreted?

1. The collected network data are themselves the primary object of interest.
2. The collected network data are interesting primarily as representative of an underlying 'true' network.

## Interpretation: Two Scenarios

---

With respect to what frame of reference are the network characteristics interpreted?

1. The collected network data are themselves the primary object of interest.
2. The collected network data are interesting primarily as representative of an underlying 'true' network.

### The Point of Today's Talk:

The distinction is important!

Under Scenario 2, statistical sampling theory becomes relevant . . . but is not trivial.

# Agenda for the Talk

---

**Goal:** Examine the implications of sampling on the inference of network graph characteristics, and describe existing work addressing these implications.

1. Establish context and notation.
2. Network Sampling and Estimation
  - (a) Horvitz-Thompson Estimation for Totals
  - (b) Network Sampling Designs and Estimates of Totals
  - (c) Beyond Total
3. Estimation of Network Size
  - (a) Estimation of Group Size / Species Problems
  - (b) Extended Example: Estimating the Size of the Internet

## Some Notation

---

Let

- $G = (V, E)$  be a network graph
- $G^* = (V^*, E^*)$  be a sampled subgraph of  $G$
- $\eta(G)$  be a summary characteristic of  $G$

## Some Notation

---

Let

- $G = (V, E)$  be a network graph
- $G^* = (V^*, E^*)$  be a sampled subgraph of  $G$
- $\eta(G)$  be a summary characteristic of  $G$

**Goal:** Accurate estimation of  $\eta = \eta(G)$   
by some  $\hat{\eta} = \hat{\eta}(G^*)$ .

# Examples of Network Summaries

---

Examples of  $\eta(G)$  include

- The number of nodes  $N_v = |V|$
- The number of links  $N_e = |E|$
- The degree  $d_i$  of a node  $i \in V$
- The fraction  $f_d$  of nodes  $i \in V$  with degree  $d_i = d$
- The clustering coefficient  $cl(G)$
- Etc.

## A Natural Starting Point

---

**Question:** How representative of  $\eta(G)$   
is the plug-in estimate  $\eta(G^*)$  ?

## A Natural Starting Point

---

**Question:** How representative of  $\eta(G)$   
is the plug-in estimate  $\eta(G^*)$  ?

**Answer:** Often  $\eta(G^*)$  is a poor representation of  $\eta(G)$  !

## A Natural Starting Point

---

**Question:** How representative of  $\eta(G)$   
is the plug-in estimate  $\eta(G^*)$  ?

**Answer:** Often  $\eta(G^*)$  is a poor representation of  $\eta(G)$  !

**Followup Question:** In that case, can we construct a better estimator from the information in  $G^*$  ?

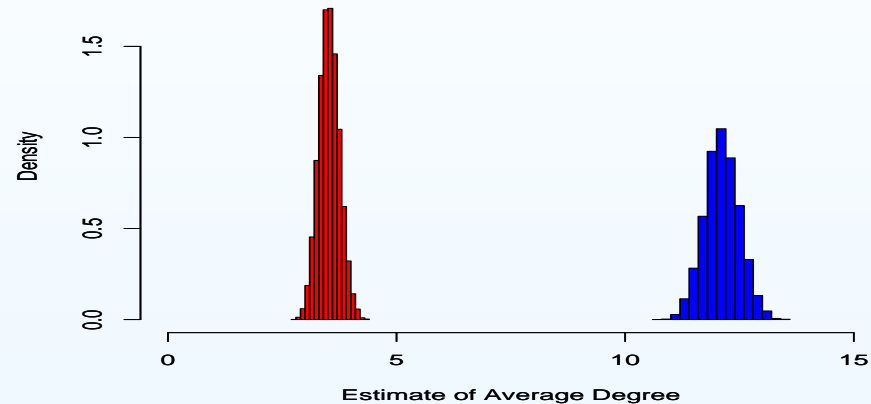
## Example: Estimating Average Degree

We conduct a small sampling experiment for illustration.

- Let  $G$  be the BioGRID network of protein interactions in yeast (i.e., *S. Cerevisiae*).
- Take  $\eta(G) = \text{Average Degree}$  (i.e., equals 12.115 in our network).
- Sample  $n$  vertices  $V^* \subseteq V$  using SRS, and then either
  1. sample all edges incident to each  $i \in V^*$ , or
  2. sample all edges  $\{i, j\}$  such that  $i, j \in V^*$ .
- Under each sampling design, estimate

$$\eta(G) = (1/N_v) \sum_{i \in V} d_i \quad \text{by} \quad \eta(G^*) = (1/n) \sum_{i \in V^*} d_i^* .$$

## Example (cont.)



Significant under-estimation in Design 2 (red) ...  
... but not in Design 1 (blue). Why?

- In Design 1, we sample vertex degree *explicitly*  
i.e.,  $d_i^* = d_i$ .
- In Design 2, we (*implicitly*) sample vertex degree with bias  
i.e.,  $d_i^* \approx nd_i/N_v$

## Additional Results

---

Lee, Kim, & Jeong (2006) provide a more comprehensive study of the naive estimator  $\hat{\eta}(G) = \eta(G^*)$ .

Study design varied network, sampling, and summary metric:

- **Networks:** BA, PPI (yeast), Internet (AS level), co-authorship (arXiv.org). Each with  $N = 30000$  nodes.
- **Sampling:** Vertex, edge, and snowball.
- **Summaries:** Degree distribution exponent, average path length, betweenness distribution exponent, assortativity, and clustering coefficient.

## Numerical Results from Lee *et al.*

	BA	PPI	AS	arXiv
Degree Exponent	↑ ↑ ↓	↑ ↑ =	= = ↓	↑ ↑ ↓
Average Path Length	↑ ↑ =	↑ ↑ ↓	↑ ↑ ↓	↑ ↑ ↓
Betweenness	↑ ↑ ↓	↑ ↑ ↓	↑ ↑ ↓	= = =
Assortativity	= = ↓	= = ↓	= = ↓	= = ↓
Clustering Coefficient	= = ↑	↑ ↓ ↑	↓ ↓ ↑	↓ ↓ ↓

Entries indicate direction of bias for vertex (**red**), edge (**green**), and snowball (**blue**) sampling.

# Improving the Accuracy of Estimation

---

In order to do better, we need to incorporate the effects of

- random sampling, and/or
- measurement error.

Focus today primarily on effects of random sampling.

Perspective one of 'design-based' inference<sup>a</sup>.

---

<sup>a</sup>As opposed to 'model-based' inference.

# Agenda for the Talk

---

**Goal:** Examine the implications of sampling on the inference of network graph characteristics, and describe existing work addressing these implications.

1. Establish context and notation.
2. **Network Sampling and Estimation**
  - (a) Horvitz-Thompson Estimation for Totals
  - (b) Network Sampling Designs and Estimates of Totals
  - (c) Beyond Totals
3. Estimation of Network Size
  - (a) Estimation of Group Size / Species Problems
  - (b) Extended Example: Estimating the Size of the Internet

## Background on Classical Sampling

- Finite population  $\mathcal{U}$  of units  $\{1, \dots, N_{\mathcal{U}}\}$ .

E.g., People, animals, objects, etc.

- A value(s)  $y_i$  associated with each  $i \in \mathcal{U}$ .

E.g., Height, weight, member/non-member, etc.

- Typical interest in averages and totals i.e.,

$$\mu \equiv (1/N_{\mathcal{U}}) \sum_{i \in \mathcal{U}} y_i \quad \text{and} \quad \tau = N_{\mathcal{U}} \mu .$$

NB: Special case of a total is  $\tau = N_{\mathcal{U}}$  .

## Sampling Background (cont.)

Basic paradigm in sampling is oriented around the following steps:

- Sample  $n$  units  $\{i_1, \dots, i_n\}$  from  $\mathcal{U}$
- Observe the value  $y_{i_k}$  for  $k = 1, \dots, n$
- Form an estimator  $\hat{\mu}$  of  $\mu$  that is unbiased i.e.,

$$\mathbb{E}[\hat{\mu}] = \mu ,$$

where the ‘ $\mathbb{E}$ ’ is expectation wrt the random sampling.

- Evaluate or estimate the variance  $\mathbb{V}(\hat{\mu})$ .

## Estimation: A First Attempt

---

**Idea:** How about using  $\hat{\mu}_n = (1/n) \sum_{k=1}^n y_{i_k}$   
i.e., the sample mean?

## Estimation: A First Attempt

**Idea:** How about using  $\hat{\mu}_n = (1/n) \sum_{k=1}^n y_{i_k}$   
i.e., the sample mean?

Let  $\pi_i = \Pr\{ \text{Unit } i \text{ is in the sample} \}$ .

Then

$$\mathbb{E}[\hat{\mu}_n] = (1/n) \sum_{i=1}^{N_{\mathcal{U}}} y_i \pi_i .$$

$\Rightarrow \bar{y}_n$  is unbiased iff  $\pi_i = n/N_{\mathcal{U}}$

## Estimation: A First Attempt

**Idea:** How about using  $\hat{\mu}_n = (1/n) \sum_{k=1}^n y_{i_k}$   
i.e., the sample mean?

Let  $\pi_i = \Pr\{ \text{Unit } i \text{ is in the sample} \}$ .

Then

$$\mathbb{E}[\hat{\mu}_n] = (1/n) \sum_{i=1}^{N_{\mathcal{U}}} y_i \pi_i .$$

$\Rightarrow \bar{y}_n$  is unbiased iff  $\pi_i = n/N_{\mathcal{U}}$

**Key Point:** The  $\pi$ 's are  $n/N_{\mathcal{U}}$  for random sampling w/out replacement; not the case more generally.

## Estimation: Horvitz-Thompson

---

**Solution:** Unequal probability sampling necessitates unequal weights when averaging.

## Estimation: Horvitz-Thompson

**Solution:** Unequal probability sampling necessitates unequal weights when averaging.

An unbiased estimator of  $\mu$  is  $\hat{\mu}_\pi = (1/N_{\mathcal{U}})\hat{\tau}_\pi$ , where

$$\hat{\tau}_\pi = \sum_{i=1}^{N_{\mathcal{U}}} \frac{y_i S_i}{\pi_i} ,$$

for

$$S_i = \begin{cases} 1 & \text{if node } i \text{ is in the sample} \\ 0 & \text{otherwise .} \end{cases}$$

## Estimation: Horvitz-Thompson

**Solution:** Unequal probability sampling necessitates unequal weights when averaging.

An unbiased estimator of  $\mu$  is  $\hat{\mu}_\pi = (1/N_{\mathcal{U}})\hat{\tau}_\pi$ , where

$$\hat{\tau}_\pi = \sum_{i=1}^{N_{\mathcal{U}}} \frac{y_i S_i}{\pi_i} ,$$

for

$$S_i = \begin{cases} 1 & \text{if node } i \text{ is in the sample} \\ 0 & \text{otherwise .} \end{cases}$$

**Caveat Emptor:**  $\pi_i$ 's can be nontrivial to compute.

## Horvitz-Thompson (cont.)

The variance of  $\hat{\tau}_\pi$  has the form

$$\mathbb{V}(\hat{\tau}_\pi) = \sum_{i=1}^{N_{\mathcal{U}}} \left( \frac{1 - \pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^{N_{\mathcal{U}}} \sum_{j \neq i} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j .$$

where  $\pi_{ij} = \Pr\{ \text{Units } i \text{ and } j \text{ are in the sample} \}$ .

This typically can be estimated from the sample.

Note: Variance of  $\hat{\tau}_\pi$  low when  $\pi_i \propto y_i$ .

## Network Totals

Many of the network summary measures of standard interest can be expressed in terms of totals.

- Let  $\mathcal{U} = V$  and  $y_i = d_i$ . Then

$$\begin{aligned}\eta(G) &= \text{Average Degree in } G \\ &\propto \sum_{i \in V} d_i .\end{aligned}$$

- Let  $\mathcal{U} = E$  and  $y_{\{i,j\}} = 1$ . Then

$$\begin{aligned}\eta(G) &= N_e \\ &= \sum_{\{i,j\} \in E} 1 .\end{aligned}$$

## Network Totals (cont.)

- Let  $\mathcal{U} = V^{(2)}$  and  $y_{(i,j)} = I_{k \in \mathcal{P}(i,j)}$ . Then for unique shortest paths  $\mathcal{P}(i,j)$ ,

$$\begin{aligned}\eta(G) &= c_B(k) \\ &= \sum_{(i,j) \in V^{(2)}} I_{k \in \mathcal{P}(i,j)} \cdot\end{aligned}$$

- Let  $\mathcal{U} = V^{(3)}$  i.e., the set of all triples of distinct vertices  $(i,j,k)$ . Then

$$\begin{aligned}\eta(G) &= \text{cl}_T(G) \\ &= \frac{\text{Total \# of Triangles}}{\text{Total \# of Connected Triples}} \cdot\end{aligned}$$

## Estimation of Network Totals

---

As a result, we can in principle bring H-T theory to bear on numerous network sampling/estimation problems.  
(Frank, 1970s)

## Estimation of Network Totals

---

As a result, we can in principle bring H-T theory to bear on numerous network sampling/estimation problems.

(Frank, 1970s)

**True . . . but *caveat emptor***

- Inclusion probabilities  $\pi$ , necessary for H-T estimators, may be for nodes or edges or . . . !

Potentially non-trivial to compute.

## Estimation of Network Totals

---

As a result, we can in principle bring H-T theory to bear on numerous network sampling/estimation problems.  
(Frank, 1970s)

**True . . . but *caveat emptor***

- Inclusion probabilities  $\pi$ , necessary for H-T estimators, may be for nodes or edges or . . . !

Potentially non-trivial to compute.

- Whether a given variable  $y$  is observable may vary with the sampling design

E.g.,  $y_i = d_i$

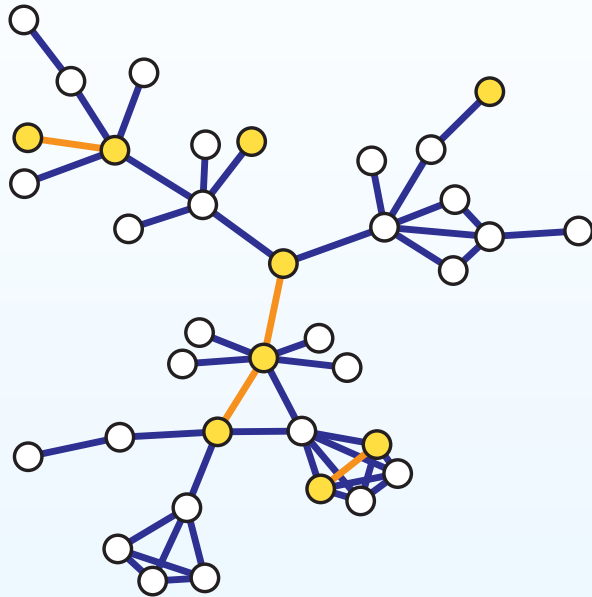
# Four Common Network Sampling Designs

---

We'll look at four common network sampling designs and their inclusion probabilities  $\pi_i$ .

- Induced Subgraph Sampling
- Incident Subgraph Sampling
- Snowball Sampling
- Link Tracing

# Induced Subgraph Sampling



Take a SRS of  $n$  vertices (yellow).

Observe all edges (orange) in the subgraph induced by  $V^*$ .

Example: Friendship networks.

## Induced Subgraph Sampling (cont.)

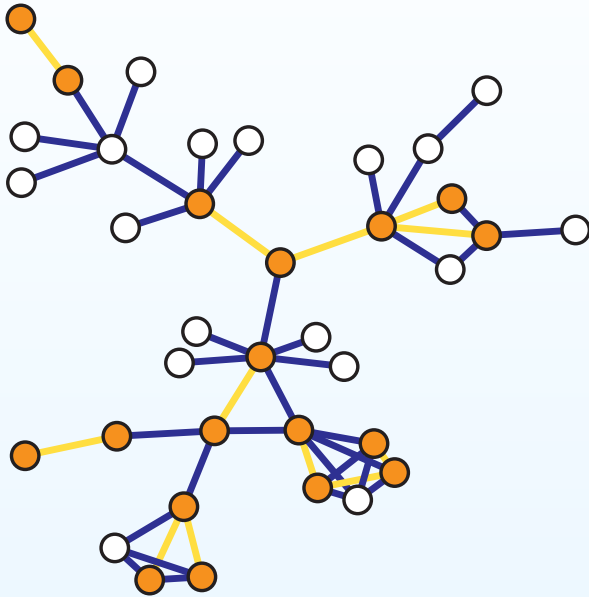
---

Vertex and edge inclusion probabilities uniformly equal to

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

Note: Calculation of these probabilities requires knowledge of  $N_v$ . If unavailable, must estimate. Will discuss this problem later.

## Incident Subgraph Sampling



Take a SRS of  $n$  edges (yellow).

Observe all vertices (orange) incident to edges in  $E^*$ .

Example: Telephone call graphs.

## Incident Subgraph Sampling (cont.)

Edge inclusion probabilities are simply

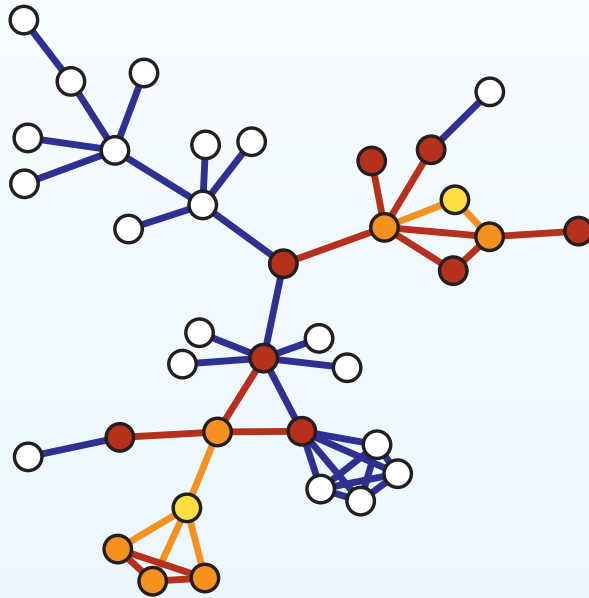
$$\pi_{\{i,j\}} = n/N_e .$$

Vertex inclusion probabilities take the form

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

Note: Calculation of these probabilities requires knowledge of  $N_e$  and the  $d_i$ 's.

# Snowball Sampling



(Two-stage)

Begin with an initial vertex sample  $V_0^*$ .

Observe

1. all incident edges, and
2. those vertices sharing these edges.

Iterate to the desired number of waves.

Examples: 'Spiders' on the WWW; sexual contact networks.

## Snowball Sampling (cont.)

---

In general, calculation of inclusion probabilities becomes increasingly intractable after one-stage snowball sampling.

With only one stage, this reduces to *star sampling*:

- *unlabeled*  
E.g., Count all co-authors for each of  $n$  authors.
- *labeled*  
E.g., Record all co-authors for each of  $n$  authors.

## Link Tracing

After selection of an initial set of vertices  $V_0$ , some subset of the edges (i.e., 'links') are traced to additional vertices.

Snowball sampling is a special case.

In general, it may not be feasible that all edges incident to a given vertex be followed (i.e., as in snowball sampling).

E.g., Lack of recollection, or simply deception, in social contact networks.

# Path Sampling

---

## Design:

- Randomly select
  - a set of source nodes  $S = \{s_1, \dots, s_{n_S}\}$
  - a set of target nodes  $T = \{t_1, \dots, t_{n_T}\}$
- Traverse the path between each pair  $(s_i, t_j)$ , taking measurements enroute.

# Path Sampling

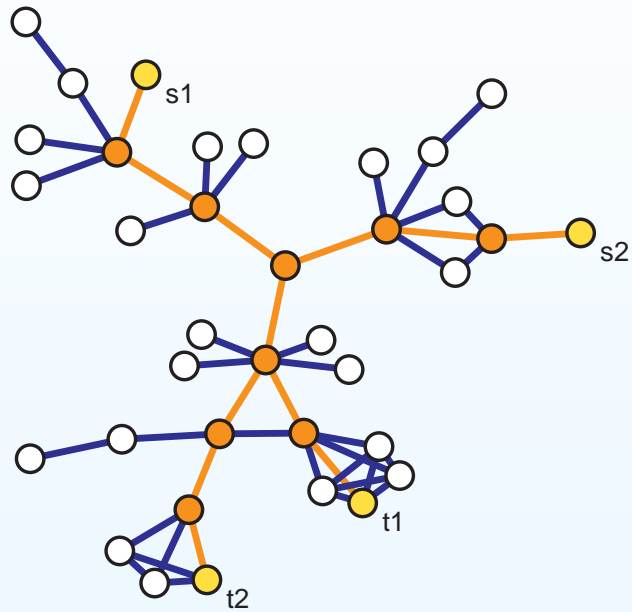
## Design:

- Randomly select
  - a set of source nodes  $S = \{s_1, \dots, s_{n_S}\}$
  - a set of target nodes  $T = \{t_1, \dots, t_{n_T}\}$
- Traverse the path between each pair  $(s_i, t_j)$ , taking measurements enroute.

## Examples:

- Traceroute studies in the Internet
- Milgram's 'Six Degrees' Study

# Traceroute Sampling



$n_S$  source nodes

$n_T$  target nodes

‘Trace’ shortest paths from each source to all targets.

## Traceroute Sampling (cont.)

Dall'Asta *et al.* (2006) show that, roughly, the inclusion probabilities behave like

$$\pi_i \approx 1 - (1 - \rho_S - \rho_T) \exp(-\rho_S \rho_T b_i)$$

and

$$\pi_{\{i,j\}} \approx 1 - \exp(-\rho_S \rho_T b_{i,j}) ,$$

for vertices and edges, respectively, where

- $b_i$  = betweenness centrality of vertex  $i$
- $b_{i,j}$  = betweenness centrality of edge  $\{i, j\}$
- $\rho_S = n_S/N$ ;  $\rho_T = n_T/N$

## Estimation of Other Network Characteristics

---

Classical sampling theory rests heavily on Horvitz-Thompson framework.

⇒ Relevant only to network totals.

Other network characteristic summaries are of interest as well ... especially, the degree distribution!

**Findings:** Sampling can potentially render observed degree distributions *highly* unrepresentative of actual degree distributions ... and in ways particularly *unhelpful* to the problem of characterizing heterogeneous distributions.

We'll briefly look at a mix of older and newer results in this area.

## Estimating Degree Frequencies

---

Frank (1971) has looked at the problem of estimating the degree frequencies  $\{f_d\}_{d \geq 0}$ .

Let  $f_d$  and  $f_d^*$  be the true and observed frequencies of degree  $d$  vertices in  $G$  and  $G^*$ , respectively.

## Estimating Degree Frequencies

Frank (1971) has looked at the problem of estimating the degree frequencies  $\{f_d\}_{d \geq 0}$ .

Let  $f_d$  and  $f_d^*$  be the true and observed frequencies of degree  $d$  vertices in  $G$  and  $G^*$ , respectively.

Under induced subgraph sampling,

$$\mathbb{E}[f_d^*] = \sum_{d'=0}^{N_v-1} P(d, d') f_d' ,$$

where 
$$P(d, d') = \binom{d'}{d} \binom{N-1-d'}{n-1-d} / \binom{N_v-1}{n-1} .$$

## Estimating Degree Frequencies (cont)

---

- A method-of-moments approach suggests substituting  $f_d^*$  for  $\mathbb{E}[f_d^*]$  on the LHS.
- In general, however, this yields an under-determined system of equations for  $\{f_d\}$ .
- If sensible, a set of constraints can rectify this e.g.,

$$f_d = 0, \text{ for all } d \geq n .$$

- But this is problematic for scale-free settings!

Note: In addition, variance formulas problematic.

## Degree Distribution Shape

---

Sampling also can fundamentally affect our inference of the basic shape of the degree distribution  
i.e., **homogeneous vs heterogeneous**.

We'll quickly summarize three seminal sets of results in this area.

- Lakhina *et al.* (2003)
- Han *et al.* (2005)
- Stumpf *et al.* (2005)

**Experiment: Simulating traceroute in the Internet**

- $G$  an Erdos-Renyi random graph.
- Equip  $G$  with edge weights  $w = 1 \pm \epsilon$ .
- Define a routed path from node  $i$  to node  $j$  to be the shortest path wrt  $\{w\}$ .
- Randomly sample  $n_S$  source nodes  $S = \{s_1, \dots, s_{n_S}\}$  and  $n_T$  target nodes  $T = \{t_1, \dots, t_{n_T}\}$ .  
(E.g.,  $n_S = 1, 5, \text{ or } 10$ ;  $n_T = 1000$ ;  $N \approx 100000$ )
- Obtain  $G^*$  through traceroute sampling.

## Lakhina *et al.* (cont.)

**Results:**  $G^*$  exhibits a power-law-like degree distribution  
... yet  $G$  has a Poisson degree distribution!

## Lakhina *et al.* (cont.)

**Results:**  $G^*$  exhibits a power-law-like degree distribution  
... yet  $G$  has a Poisson degree distribution!

Follow-up work by Clauset and Moore (2005) and Achlioptas *et al.* (2005) confirm and refine this using analytical arguments.

E.g.,

- For  $G$  a power-law graph, traceroute-like sampling can produce a power-law  $G^*$  whose exponent significantly underestimates that of  $G$ .
- Equivalency of exponents can be managed with  $n_S \sim \bar{d}(G)$ .

**Experiment: Simulating Y2H Experiments in Biology**

- $G$  from one of either Erdos-Renyi, Exponential, Scale-free, or Truncated Normal random graph models. E.g.,  $N = 6000$  or  $10000$ , as in yeast or worm proteomes.
- Randomly sample a fraction of nodes as ‘bait’, and a fraction of their neighbors as ‘prey’.
- Observe the corresponding edges between bait and prey nodes, and let  $G^*$  be the union of these nodes and edges.

## Han *et al.* (cont.)

---

### Results:

- Low-coverage sampling produced  $G^*$  with degree distributions of a power-law-like form.
- Non-trivial range of sampling rates can re-produce a set of four topological characteristics observed *in silico*.

(See Thomas *et al.* 2003 for related work.)

## Stumpf *et al.*

**Model:**  $G$  a random graph;  $G^*$  produced by induced subgraph sampling, following Bernoulli( $p$ ) random sampling of nodes.

**Question:** When are the PGFs of  $G$  and  $G^*$  in the same family?

## Stumpf *et al.*

**Model:**  $G$  a random graph;  $G^*$  produced by induced subgraph sampling, following Bernoulli( $p$ ) random sampling of nodes.

**Question:** When are the PGFs of  $G$  and  $G^*$  in the same family?

**Results:**

- True when (negative) binomial in distribution.  
⇒ Includes E-R and geometric (exponential) as special cases.
- Does *not* include power-law distributions.  
⇒ Low- to medium-connectivity nodes most affected.

# Agenda for the Talk

---

**Goal:** Examine the implications of sampling on the inference of network graph characteristics, and describe existing work addressing these implications.

1. Establish context and notation.
2. Network Sampling and Estimation
  - (a) Horvitz-Thompson Estimation for Totals
  - (b) Network Sampling Designs and Estimates of Totals
  - (c) Beyond Totals
3. **Estimation of Network Size**
  - (a) Estimation of Group Size / Species Problems
  - (b) Extended Example: Estimating the Size of the Internet

## Estimating the Size of a Group

---

Estimation of the total number of members  $N$  in a group is a special type of estimation problem.

Distinguish between two (related!) problems:

- Estimating the size of a population.
- Estimating the number of 'species' in the population.

We'll look at the generic version of each of these, followed by an analogous version in the context of networks.

## Estimating the Size of a Population

---

Suppose we sample  $n$  units from the population  $\mathcal{U}$  and wish to know the size  $\tau = N_{\mathcal{U}} = |\mathcal{U}|$ .

## Estimating the Size of a Population

---

Suppose we sample  $n$  units from the population  $\mathcal{U}$  and wish to know the size  $\tau = N_{\mathcal{U}} = |\mathcal{U}|$ .

If we knew the collection  $\{\pi_i\}$ , we could use

$$\hat{\tau}_{\pi} = \sum_{i=1}^{N_{\mathcal{U}}} \frac{S_i}{\pi_i} .$$

## Estimating the Size of a Population

Suppose we sample  $n$  units from the population  $\mathcal{U}$  and wish to know the size  $\tau = N_{\mathcal{U}} = |\mathcal{U}|$ .

If we knew the collection  $\{\pi_i\}$ , we could use

$$\hat{\tau}_{\pi} = \sum_{i=1}^{N_{\mathcal{U}}} \frac{S_i}{\pi_i} .$$

Often not realistic, since knowledge of  $\pi_i$ 's frequently derives from knowledge of  $N_{\mathcal{U}}$  e.g.,

$$\pi_i^{SRS} = \frac{n}{N_{\mathcal{U}}} .$$

## Capture-Recapture Methods

---

An alternative approach is to use two rounds of sampling i.e., for 'capture' and 'recapture'.

## Capture-Recapture Methods

---

An alternative approach is to use two rounds of sampling i.e., for 'capture' and 'recapture'.

### Basic Idea:

- Sample  $n_1$  units from  $\mathcal{U}$  and 'mark' them.
- Sample  $n_2$  units from  $\mathcal{U}$ ;  
denote the number found to be marked by  $m$ .
- Equating the proportions of marked units in the second sample and in the population i.e.,  $m/n_2 = n_1/\tau$ , suggests

$$\hat{\tau} = \frac{n_1}{(m/n_2)} .$$

## How Large is a 'Hidden' Population?

---

In 'hidden' populations, individuals typically do not wish to expose themselves to view (e.g., socially sensitive and/or elicit activities, homeless, etc.).

Frank and Snidjers (1995) propose to use snowball sampling and capture-recapture principles to estimate the size of a hidden population.

**Key Idea:** By accounting for how many individuals in the first wave are identified both (i) among themselves, and (ii) by individuals in the initial sample, we mimic capture-recapture sampling.

## Estimating the Size of 'Hidden' Populations

---

- Let  $G = (V, E)$  be a directed graph.
- An arc from  $i$  to  $j$  indicates that, if asked, person  $i$  would identify person  $j$  as a member of the hidden pop.
- Sample initial set of vertices  $V_0^*$  wrt Bernoulli( $p_0$ ).
- Sample first wave snowball sample, yielding additional vertices  $V_1^*$ .
- Let  $G^* = (V^*, E^*)$ , where  $V^* = V_0^* \cup V_1^*$  and  $E^*$  are the arcs revealed by the first wave.

## Estimating the Size of 'Hidden' Populations (cont.)

Frank and Snijders show that

$$\begin{aligned}\mathbb{E}(N) &= N_v p_0 \\ \mathbb{E}(M_1) &= (N_e - N_v) p_0^2 \\ \mathbb{E}(M_2) &= (N_e - N_v) p_0 (1 - p_0) ,\end{aligned}$$

where

- $N = |V_0^*|$
- $M_1 =$  number of arcs within  $V_1^*$
- $M_2 =$  number of arcs from  $V_0^*$  to  $V_1^*$

Method-of-moments yields

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

## Species Estimation

---

Suppose instead that ‘copies’ of units are observed i.e.,

- Individual lions, tigers, and bears (oh my!)
- Individual words of an author’s vocabulary

**Problem Statement:** Given  $L^*$  species observed in a sample of  $n$  units, estimate the total number  $L \geq L^*$  of species in the population.

## Species Estimation (cont.)

---

**Key Issue:** The nature of inclusion probabilities for each species  
... *particularly* for those species *not* seen!

## Species Estimation (cont.)

---

**Key Issue:** The nature of inclusion probabilities for each species  
... *particularly* for those species *not* seen!

- Under SRS, and species of roughly equal memberships, the problem can be treated essentially with reasoning in the spirit of capture-recapture.

(See *JASA* review by Bunge and Fitzpatrick.)

## Species Estimation (cont.)

---

**Key Issue:** The nature of inclusion probabilities for each species  
... *particularly* for those species *not* seen!

- Under SRS, and species of roughly equal memberships, the problem can be treated essentially with reasoning in the spirit of capture-recapture.
- Typically sampling intensities are unequal, often highly so.

(See *JASA* review by Bunge and Fitzpatrick.)

## Species Estimation (cont.)

---

**Key Issue:** The nature of inclusion probabilities for each species  
... *particularly* for those species *not* seen!

- Under SRS, and species of roughly equal memberships, the problem can be treated essentially with reasoning in the spirit of capture-recapture.
- Typically sampling intensities are unequal, often highly so.
- Methods include
  - Parametric modeling of inclusion probabilities
  - Estimation of coverage probabilities

(See *JASA* review by Bunge and Fitzpatrick.)

## Internet 'Species'

---

As a massive, self-organizing system, the topology of the Internet is largely unknown in its entirety.

Even basic characteristics, such as  $N_v = |V|$ ,  $N_e = |E|$ , and  $\{f_d\}$  are not known with any certainty.

## Internet 'Species'

As a massive, self-organizing system, the topology of the Internet is largely unknown in its entirety.

Even basic characteristics, such as  $N_v = |V|$ ,  $N_e = |E|$ , and  $\{f_d\}$  are not known with any certainty.

Key Observation: Under `traceroute` sampling, estimation of  $N_v$ ,  $N_e$ , and degrees are all species problems ...

... and potentially quite difficult!

We'll look at work of Viger *et al.* (2007), studying the problem of estimating  $N_v$ .

# How 'Big' is the Internet?

---

**Goal:** Estimation of  $N_v$ .

## How 'Big' is the Internet?

---

**Goal:** Estimation of  $N_v$ .

Can argue that

$$N_v = 1 + \frac{\mathbb{E}[b]}{\ell - 1},$$

where

- $\mathbb{E}[b]$  is the average vertex betweenness on  $G$
- $\ell$  is the average shortest path between vertices

## How 'Big' is the Internet?

**Goal:** Estimation of  $N_v$ .

Can argue that

$$N_v = 1 + \frac{\mathbb{E}[b]}{\ell - 1},$$

where

- $\mathbb{E}[b]$  is the average vertex betweenness on  $G$
- $\ell$  is the average shortest path between vertices

**Idea:**

- Parametric model for  $P(b) = \#\{i \in V : b_i = b\}/N_v$
- Estimation of  $N_v$  through estimation of  $\mathbb{E}[b]$ .

## Modeling Vertex Betweenness

---

Consider a mixture model

$$P(b) = \pi P_1(b) + (1 - \pi) P_2(b) ,$$

where

- $P_1(b)$  is supported on  $[1, b_{min})$
- $P_2(b)$  is supported on  $[b_{min}, b_{max}]$
- $P_2(b) = b^{-\beta} / K$

# Modeling Vertex Betweenness

---

Consider a mixture model

$$P(b) = \pi P_1(b) + (1 - \pi) P_2(b) ,$$

where

- $P_1(b)$  is supported on  $[1, b_{min})$
- $P_2(b)$  is supported on  $[b_{min}, b_{max}]$
- $P_2(b) = b^{-\beta} / K$

Estimation of

$$\mathbb{E}[b] = \pi \mathbb{E}_1[b] + (1 - \pi) \mathbb{E}_2[b]$$

requires estimation of  $\pi$  and each component mean.

## Difficulties w/ Parametric Approach

---

Unfortunately, what we know of `traceroute` sampling suggests this approach will be problematic.

## Difficulties w/ Parametric Approach

---

Unfortunately, what we know of `traceroute` sampling suggests this approach will be problematic.

- Estimation of  $\mathbb{E}_1[b]$  and  $\pi$  require information on nodes with low betweenness i.e., precisely those nodes we are unlikely to see.

## Difficulties w/ Parametric Approach

---

Unfortunately, what we know of `traceroute` sampling suggests this approach will be problematic.

- Estimation of  $\mathbb{E}_1[b]$  and  $\pi$  require information on nodes with low betweenness i.e., precisely those nodes we are unlikely to see.
- Estimation of  $\mathbb{E}_2[b]$  requires knowledge of  $\pi$ , and additionally is likely to be unstable, due to  $\beta \approx 2$ .

# 'Leave-One-Out' Estimator: Overview

---

**Idea:** Information on unseen nodes gained through rate of return per target node.

**Assumptions:** Low marginal rate of return from any single target node; simple random sampling of targets.

Formal argument leads to

$$\hat{N}_{vL1O} \approx (n_S + n_T) + \frac{N_v^* - (n_S + n_T)}{1 - w^*},$$

where  $w^*$  is the fraction of target nodes not discovered by traces to any other target.

## 'Leave-One-Out' Estimator: Details

Define  $V_{ij}^*$  = vertices on path from  $i$  to  $j$  (includes  $i$  and  $j$ ).

Let  $V_{(-j)}^* = \bigcup_i \bigcup_{j' \neq j} V_{ij'}^*$  and  $N_{(-j)}^* = |V_{(-j)}^*|$ .

Write

$$\mathbb{E}[X] = \frac{n_T(N_v - \mathbb{E}[N_{(-)}^*])}{N_v - n_S - n_T + 1}$$

where  $X = \sum_j \delta_j$ , *i.e.* how many targets are not discovered by routes to any other target. Rewrite

$$N_v = \frac{n_T \mathbb{E}[N_{(-)}^*] - (n_S + n_T - 1) \mathbb{E}[X]}{n_T - \mathbb{E}[X]}$$

## Details (cont.)

Estimate  $\mathbb{E}[N_{(-)}^*]$  by

$$\bar{N}_{(-)}^* = (1/n_T) \sum_j N_{(-j)}^* .$$

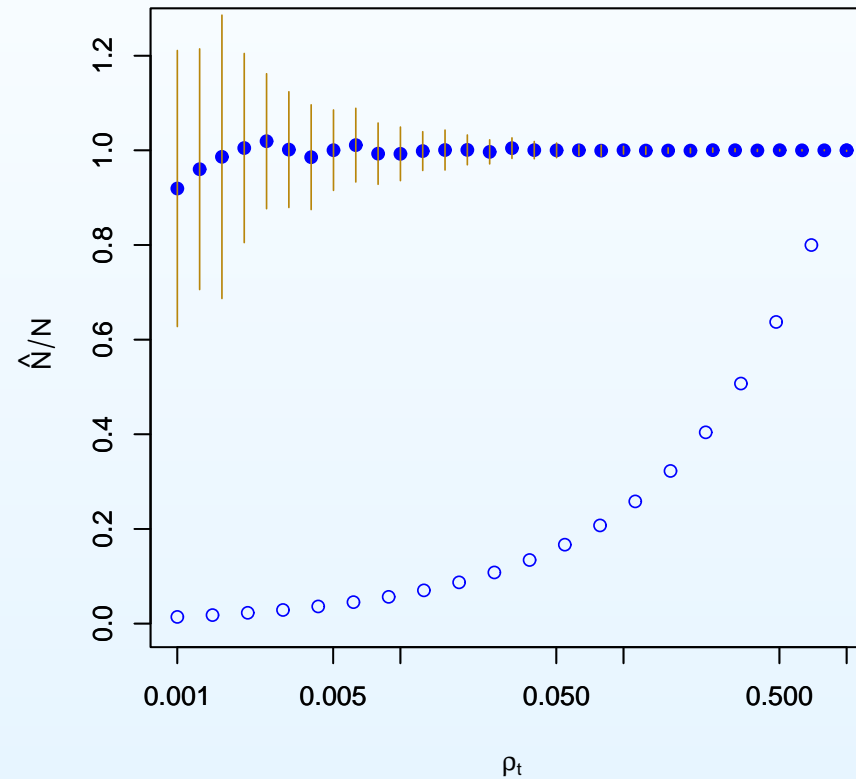
Estimating  $\mathbb{E}[X]$  by  $X$  can be unstable in the denominator.  
Instead, estimate  $(n_T - \mathbb{E}[X])^{-1}$  directly.

If relative yield of new vertices from any single or pair of targets is modest,

$$\frac{n_T + 1}{n_T + 1 - X}$$

is approximately unbiased for  $(n_T - \mathbb{E}[X])^{-1}$ .

# Numerical Results



Comparison of  $\hat{N} = \hat{N}_v$  (filled circles) and  $\hat{N} = N_v^*$  (open circles), as estimators of  $N_v$ , for various values of target sampling density  $\rho_t$ .

## A Small Empirical Study

---

**Goal:** Compare estimates of  $N$  from `ping` and `traceroute`.

## A Small Empirical Study

**Goal:** Compare estimates of  $N$  from ping and traceroute.

### **Ping:**

- 'Ping's sent to  $n = 3,726,773$  of the  $2^{32}$  possible IP addresses, from a single source.
- 61,246 valid responses received  
⇒ 1.64% response rate
- $\hat{N}_{ping} = 2^{32} \times 0.0164 \approx 70,583,787$  alive addresses

## A Small Empirical Study (cont.)

---

### Traceroute:

- Traceroutes run from the  $n_S = 1$  source to the  $n_T = 61,246$  responding IP addresses.
- $\hat{N}_{L1O} \approx 72,296,221$  alive addresses

Relative difference of only

$$\frac{\hat{N}_{L1O} - \hat{N}_{ping}}{\hat{N}_{ping}} \approx 0.024 .$$

# Closing Thoughts

## Closing Thoughts

---

- It's critical that the broader community in this area recognize the implications of sampling when characterizing networks through summary statistics.

## Closing Thoughts

---

- It's critical that the broader community in this area recognize the implications of sampling when characterizing networks through summary statistics.
- There is already a nontrivial existing literature in the statistical theory/methods of network graph sampling and inference.

## Closing Thoughts

---

- It's critical that the broader community in this area recognize the implications of sampling when characterizing networks through summary statistics.
- There is already a nontrivial existing literature in the statistical theory/methods of network graph sampling and inference.
- The current needs of the community go beyond what is available at this time.

## Closing Thoughts

---

- It's critical that the broader community in this area recognize the implications of sampling when characterizing networks through summary statistics.
- There is already a nontrivial existing literature in the statistical theory/methods of network graph sampling and inference.
- The current needs of the community go beyond what is available at this time.
- Solutions in this area are likely to be nontrivial, and frequently design-dependent.

## Additional Topics

---

- Other existing results e.g., estimation of number of subgraph counts, connectivity, etc.
- Model-based methods.
- Bayes and empirical Bayes methods.
- Adaptive sampling designs and inference.
- Testing.
- Measurement error.

## References

---

Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer, New York. (See Chapter 5.)

Achlioptas, D., Clauset, A., Kempe, D., and Moore, C. (2005). On the bias of traceroute sampling. *STOC '05*.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association*, 88, 364-373.

Clauset, A. and Moore, C. (2005). Accuracy and scaling phenomena in Internet mapping. *PRL* 94, 018701.

Dall'Asta, L., Alvarez-Hamelin, I., Barrat, A., Vázquez, A., and Vespignani, A. (2006). Exploring networks with traceroute-like probes: Theory and simulations. *Theoretical Computer Science*, 355, 6-24.

Frank, O. (1971). *Statistical Inference in Graphs*. PhD Thesis, Stockholm University.

Frank, O. (1977a). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

Frank, O. (1977b). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

## References (cont.)

---

- Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:1, 53-67.
- Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 20, 572-579.
- Han, J-D J., Dupuy, D., Bertin, N., Cusick, M.E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interactions networks. *Nature Biotechnology*, 23:7, 839-844.
- Lakhina, A., Byers, J.W., Crovella, M., and Xie, P. (2003). Sampling biases in IP topology measurements. *Proceedings of the IEEE Infocom 2003*.
- Stumpf, M.P.H., Wiuf, C., and May, R.M. (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102:12, 4221-4224.
- Thomas, A., Cannings, R., Monk, N.A.M., and Cannings, C. (2003). On the structure of protein-protein interaction networks. *Biochemical Society Transactions*, 31:6, 1491-6.
- Thompson, S.K. (1992). *Sampling*. Wiley & Sons.
- Viger, F., Barrat, A., Dall'Asta, L., Zhang, C-H., and Kolaczyk, E.D. (2007). What is the real size of a sampled network? The case of the Internet. *Physical Review E*, 75, 056111.