
Sampling Networks and the Inference of Network Characteristics

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.

- Establish context and notation.
- Extended example: Degree distributions.
- Background on statistical sampling theory.
- Sampling and estimation for network graphs.

Some Notation

Let

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

Some Notation

Let

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

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

Examples of Network Summaries

Examples of $\eta(\mathcal{G})$ include

- The number of nodes $N = |\mathcal{V}|$
- The number of links $M = |\mathcal{E}|$
- The degree d_i of a node $i \in \mathcal{V}$
- The fraction $f(k)$ of nodes $i \in \mathcal{V}$ with degree $d_i = k$
- The clustering coefficient \bar{C} of \mathcal{G}
- Etc.

A Natural Question . . .

Question: How representative of $\eta(\mathcal{G})$ is $\eta(\mathcal{G}^*)$?

Intuition: Given a sample x_1, \dots, x_n from a distribution F with mean μ_F , the sample mean \bar{x} is increasingly representative of μ_F .

A Natural Question . . .

Question: How representative of $\eta(\mathcal{G})$ is $\eta(\mathcal{G}^*)$?

Intuition: Given a sample x_1, \dots, x_n from a distribution F with mean μ_F , the sample mean \bar{x} is increasingly representative of μ_F .

Answer: Often $\eta(\mathcal{G}^*)$ is a poor representative of $\eta(\mathcal{G})$!

A Natural Question . . .

Question: How representative of $\eta(\mathcal{G})$ is $\eta(\mathcal{G}^*)$?

Intuition: Given a sample x_1, \dots, x_n from a distribution F with mean μ_F , the sample mean \bar{x} is increasingly representative of μ_F .

Answer: Often $\eta(\mathcal{G}^*)$ is a poor representative of $\eta(\mathcal{G})$!

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

Accurate Estimation

In order to do better, we need to incorporate the *sampling mechanism* and the effects of

- random sampling, and
- measurement error.

Focus today primarily on effects of random sampling.

Perspective one of ‘design-based’ inference (i.e., 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.

- Establish context and notation.
- **Extended example: Degree distributions.**
- Background on statistical sampling theory.
- Sampling and estimation for network graphs.

Extended Example

- Motivation: Degree distributions.
- Three Case Studies:
 - Lakhina *et al.* 2003
 - Han *et al.* 2005
 - Stumpf *et al.* 2005
- Lee *et al.* (Beyond degree distributions)

Illustration: Degree distributions

Since \mathcal{G}^* is a subgraph of \mathcal{G} , clearly $N^* = |\mathcal{V}^*|$ under-estimates $N = |\mathcal{V}|$.

Illustration: Degree distributions

Since \mathcal{G}^* is a subgraph of \mathcal{G} , clearly $N^* = |\mathcal{V}^*|$ under-estimates $N = |\mathcal{V}|$.

But the degree distribution $\{f_k\}$ must behave more like a set of averages, and hence be accurate ... right?

Illustration: Degree distributions

Since \mathcal{G}^* is a subgraph of \mathcal{G} , clearly $N^* = |\mathcal{V}^*|$ under-estimates $N = |\mathcal{V}|$.

But the degree distribution $\{f_k\}$ must behave more like a set of averages, and hence be accurate ... right?

You would think so ... but not necessarily!

Depends on

- the type of sampling
- the rate of sampling
- the network \mathcal{G}
- etc.

Case Study I: Lakhina *et al.*

Experiment: Simulating `traceroute` in the Internet

- \mathcal{G} an Erdos-Renyi random graph.
- Equip \mathcal{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., $n_S = 1, 5, \text{ or } 10$; $n_T = 1000$; $N \approx 100000$)
- Observe the corresponding $n_S \times n_T$ paths, and let \mathcal{G}^* be the union of these paths.

Lakhina *et al.* (cont.)

Results: \mathcal{G}^* exhibits a power-law-like degree distribution
... yet \mathcal{G} has a Poisson degree distribution!

Lakhina *et al.* (cont.)

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

Case Study II: Han *et al.*

Experiment: Simulating Y2H Experiments in Biology

- \mathcal{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 \mathcal{G}^* be the union of these nodes and edges.

Han *et al.* (cont.)

Results:

- Low-coverage sampling produced \mathcal{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.)

Case Study III: Stumpf *et al.*

Model: \mathcal{G} a random graph; \mathcal{G}^* produced by Bernoulli(p) random sampling of nodes.

Question: When are the PGFs of \mathcal{G} and \mathcal{G}^* in the same family?

Case Study III: Stumpf *et al.*

Model: \mathcal{G} a random graph; \mathcal{G}^* produced by Bernoulli(p) random sampling of nodes.

Question: When are the PGFs of \mathcal{G} and \mathcal{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.

Beyond Degree Distributions

Lee and colleagues looked at the empirical performance of the naive estimator $\hat{\eta}(\mathcal{G}) = \eta(\mathcal{G}^*)$.

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	= = ↑	↑ ↓ ↑	↓ ↓ ↑	↓ ↓ ↓

Table 1: Direction of bias for vertex (**red**), edge (**green**), and snowball (**blue**) sampling.

Agenda for the Talk

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

- Establish context and notation.
- Extended example: Degree distributions.
- **Background on statistical sampling theory.**
- Sampling and estimation for network graphs.

Background on Statistical Sampling

- Basic notation and terminology.
- Estimation of averages and totals
 - Naive estimator
 - Horvitz-Thompson estimator
- Estimating the size of group
 - Estimating the size of a population
 - Estimating the number of species

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 μ that is unbiased i.e.,

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

where the ‘ E ’ is expectation wrt the random sampling mechanism.

- Evaluate or estimate $\text{var}(\hat{\mu})$.

Estimation: A First Attempt

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

Estimation: A First Attempt

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

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

Then

$$E[\bar{y}_n] = (1/n) \sum_{i=1}^{N_U} y_i \pi_i .$$

$\Rightarrow \bar{y}_n$ is unbiased iff $\pi_i = n/N_U$

Estimation: A First Attempt

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

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

Then

$$E[\bar{y}_n] = (1/n) \sum_{i=1}^{N_U} y_i \pi_i .$$

$\Rightarrow \bar{y}_n$ is unbiased iff $\pi_i = n/N_U$

Key Point: The π 's are n/N_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 μ is $\hat{\mu}_\pi = (1/N_U)\hat{\tau}_\pi$, where

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

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 μ is $\hat{\mu}_\pi = (1/N_U)\hat{\tau}_\pi$, where

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

for

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

Caveat Emptor: π_i 's can be nontrivial to compute.

Horvitz-Thompson (cont.)

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

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

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 $\pi_i \propto y_i$.

(See Thompson's *Sampling* for more in this area.)

Estimating the Size of a Group

Estimation of the total number τ members in a group is a special case.

Distinguish between two (related!) problems:

- Estimating the size of a population.
- Estimating the number of ‘species’.

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 π_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

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

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.

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.

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.)

Agenda for the Talk

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

- Establish context and notation.
- Extended example: Degree distributions.
- Background on statistical sampling theory.
- **Sampling and estimation for network graphs.**

Sampling and Estimation for Network Graphs.

- Network summaries as graph totals and averages
- Examples from the literature:
 - Vertex/Edge Sampling
 - Snowball Sampling
- New Results: Internet ‘Species’ Estimation.

Network Sampling & Summaries Revisited

Many common network summary statistics can be expressed as ‘totals’ or ‘averages’ of appropriately defined variables y .

Network Sampling & Summaries Revisited

Many common network summary statistics can be expressed as ‘totals’ or ‘averages’ of appropriately defined variables y .

Note: Need potentially three sets of variables

i.e., $y_i^{(v)}$, $y_{i,j}^{(e)}$, and $y_{i,j}^{(a)}$ on vertices, edges, and arcs.

Network Sampling & Summaries Revisited

Many common network summary statistics can be expressed as ‘totals’ or ‘averages’ of appropriately defined variables y .

Note: Need potentially three sets of variables

i.e., $y_i^{(v)}$, $y_{i,j}^{(e)}$, and $y_{i,j}^{(a)}$ on vertices, edges, and arcs.

Examples:

- Size N of a vertex set \mathcal{V}
- Dyad-based statistics
E.g., number of edges, arcs, or mutual arcs.
- Triad-based statistics
E.g., number of transitive triples.
- Frequency of nodes with degree d .

Caveat Emptor

Two Important Points!

Caveat Emptor

Two Important Points!

- Inclusion probabilities π , necessary for H-T estimators, may be for nodes or edges or ... !
Potentially non-trivial to compute.

Caveat Emptor

Two Important Points!

- Inclusion probabilities π , 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., $y_i^{(v)} = d_i$

Vertex/Edge Sampling

Design:

- Take a simple random sample (without replacement) of n vertices (edges);
- Observe all vertices (edges) sampled and their incident edges (vertices).

Note: For vertex sampling, distinguish between observation of

1. incident edges within the sample, or
2. all incident edges (a.k.a. ‘star sampling’).

Examples of Vertex/Edge Sampling

- Survey of individuals in a population for their friendships, cooperation, communication, etc.

Examples of Vertex/Edge Sampling

- Survey of individuals in a population for their friendships, cooperation, communication, etc.
- Bait/Prey experiments in proteomics.

Examples of Vertex/Edge Sampling

- Survey of individuals in a population for their friendships, cooperation, communication, etc.
- Bait/Prey experiments in proteomics.
- Collection of phone call records by NSA.

Estimating Edge Totals

- Let $S = \{i_1, \dots, i_n\}$ be a SRS of nodes from \mathcal{V}

Estimating Edge Totals

- Let $S = \{i_1, \dots, i_n\}$ be a SRS of nodes from \mathcal{V}
- Let $\tau = \sum_{(i,j) \in \mathcal{V}^{(2)}} y_{i,j}$.

Estimating Edge Totals

- Let $S = \{i_1, \dots, i_n\}$ be a SRS of nodes from \mathcal{V}
- Let $\tau = \sum_{(i,j) \in \mathcal{V}^{(2)}} y_{i,j}$.
- Observing incident edges among $i, j \in S$ means

$$\pi_{i,j} = \Pr\{(i, j) \text{ is sampled}\} = \frac{n(n-1)}{N(N-1)} .$$

Estimating Edge Totals

- Let $S = \{i_1, \dots, i_n\}$ be a SRS of nodes from \mathcal{V}
- Let $\tau = \sum_{(i,j) \in \mathcal{V}^{(2)}} y_{i,j}$.
- Observing incident edges among $i, j \in S$ means

$$\pi_{i,j} = \Pr\{(i, j) \text{ is sampled}\} = \frac{n(n-1)}{N(N-1)} .$$

- The Horvitz-Thompson estimator for τ is

$$\hat{\tau} = \frac{N(N-1)}{n(n-1)} \sum_{(i,j) \in S^{(2)}} y_{i,j} .$$

Estimating Edge Totals (cont)

- Generalizes for arbitrary $\pi_{i,j}$'s.
- Variance formulas and unbiased estimators follow from standard H-T methodology.

See Frank (1977a).

Estimating Edge Totals (cont)

- Generalizes for arbitrary $\pi_{i,j}$'s.
- Variance formulas and unbiased estimators follow from standard H-T methodology.

See Frank (1977a).

- For $y_{i,j} = 1$ or 0 , and SRS, variance formulas simplify.

See Frank (1978) for this and extensions to triad sums.

Estimating Degree Frequencies

Although degree counts (frequencies) too are totals (averages), their estimation is a good deal more challenging.

Let $f(k)$ and $f^*(k)$ be the true and observed frequencies of degree k nodes in \mathcal{G} and \mathcal{G}^* , respectively.

Estimating Degree Frequencies

Although degree counts (frequencies) too are totals (averages), their estimation is a good deal more challenging.

Let $f(k)$ and $f^*(k)$ be the true and observed frequencies of degree k nodes in \mathcal{G} and \mathcal{G}^* , respectively.

Under our basic vertex sampling design,

$$E[f^*(k)] = \sum_{k'=0}^{N-1} P(k, k') f(k') ,$$

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

Est. Degree Frequencies (cont)

- Substituting $f^*(k)$ for $E[f^*(k)]$ on the LHS yields an under-determined system of equations for $\{f(k)\}$.
- If sensible, a set of constraints can rectify this e.g.,

$$f(k) = 0, \text{ for all } k \geq n .$$

(Problematic for scale-free settings?)

- Variance formulas problematic.

(See e.g., Frank 1971.)

Snowball Sampling

- Let $S_0 = \{i\} \subseteq \mathcal{V}$ be chosen through i.i.d. Bernoulli(p) sampling.
- Let $S_1 =$ all new vertices ‘reachable’ from S_0 in one hop.
- Iterate until no new nodes ‘reachable’, yielding S_0, S_1, \dots, S_K .
- Let $\mathcal{V}^* = \cup_j S_j$ and let \mathcal{E}^* be the set of edges used.

Note: One-hop ‘reachable’ nodes will vary with context.

(Originally due to Goodman (1961).)

Examples of Snowball Sampling

- Surveys using ‘who do you know’ approaches.

Common with hidden populations e.g., sexual contacts, drug users, homeless, etc.

Examples of Snowball Sampling

- Surveys using ‘who do you know’ approaches.
Common with hidden populations e.g., sexual contacts, drug users, homeless, etc.
- WWW ‘spiders’ (?)

Estimating Graph Totals

- In principle, Horvitz-Thompson estimators may again be used.

(Note: Randomness is only wrt the initial Bernoulli sampling.)

- Calculation of inclusion probabilities π require certain local ‘observability’ conditions not always satisfied.
- See Frank (1977b) for details in the case of 1-wave snowball sampling.

Estimating Population Size

A natural task with ‘hidden’ populations is estimation of $N = |\mathcal{V}|$.

Frank and Snijder (1995) propose the following estimator:

$$\hat{N} = \frac{n}{\hat{p}} \quad \text{where} \quad \hat{p} = \frac{k}{k + m}$$

for

- $m = |S_1|$
- $k =$ the number of vertices in S_0 connected by at least one arc to another vertex in S_0 .

Also proposed are Horvitz-Thompson estimators, model-based estimators, and resampling-based procedures for estimating variances.

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

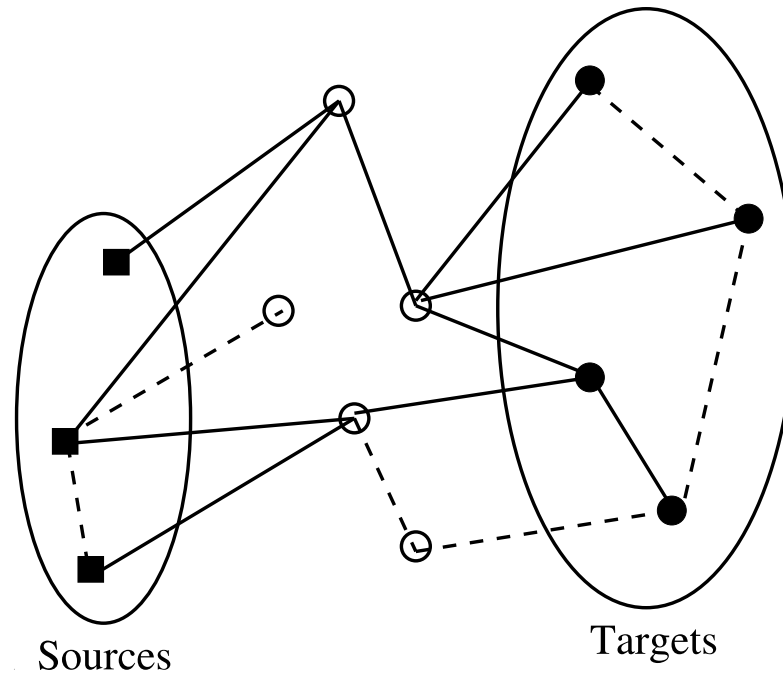
Design:

- 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 **Inclusion Probabilities**

Dall'Asta *et al.* 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

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

for vertex and edges, respectively, where

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

Warning: Internet Species Ahead!

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

Even basic characteristics, such as $N = |\mathcal{V}|$, $M = |\mathcal{E}|$, and $\{f(k)\}$ are not known with any certainty.

Warning: Internet Species Ahead!

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

Even basic characteristics, such as $N = |\mathcal{V}|$, $M = |\mathcal{E}|$, and $\{f(k)\}$ are not known with any certainty.

Key Observation: Estimation of N , M , and degrees are all species problems . . .

. . . and potentially quite difficult!

How 'Big' is the Internet?

Goal: Estimation of N .

How 'Big' is the Internet?

Goal: Estimation of N .

Can argue that

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

where

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

How 'Big' is the Internet?

Goal: Estimation of N .

Can argue that

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

where

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

Idea:

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

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

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

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

$$E[b] = \pi E_1[b] + (1 - \pi) E_2[b]$$

requires estimation of π 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 $E_1[b]$ and π 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 $E_1[b]$ and π require information on nodes with low betweenness i.e., precisely those nodes we are unlikely to see.
- Estimation of $E_2[b]$ requires knowledge of π , 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

$$\hat{N}_{L1O} \approx (n_S + n_T) + \frac{N^* - (n_S + n_T)}{1 - w^*},$$

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 $\delta_j = I\{t_j \notin V_{(-j)}^*\}$.

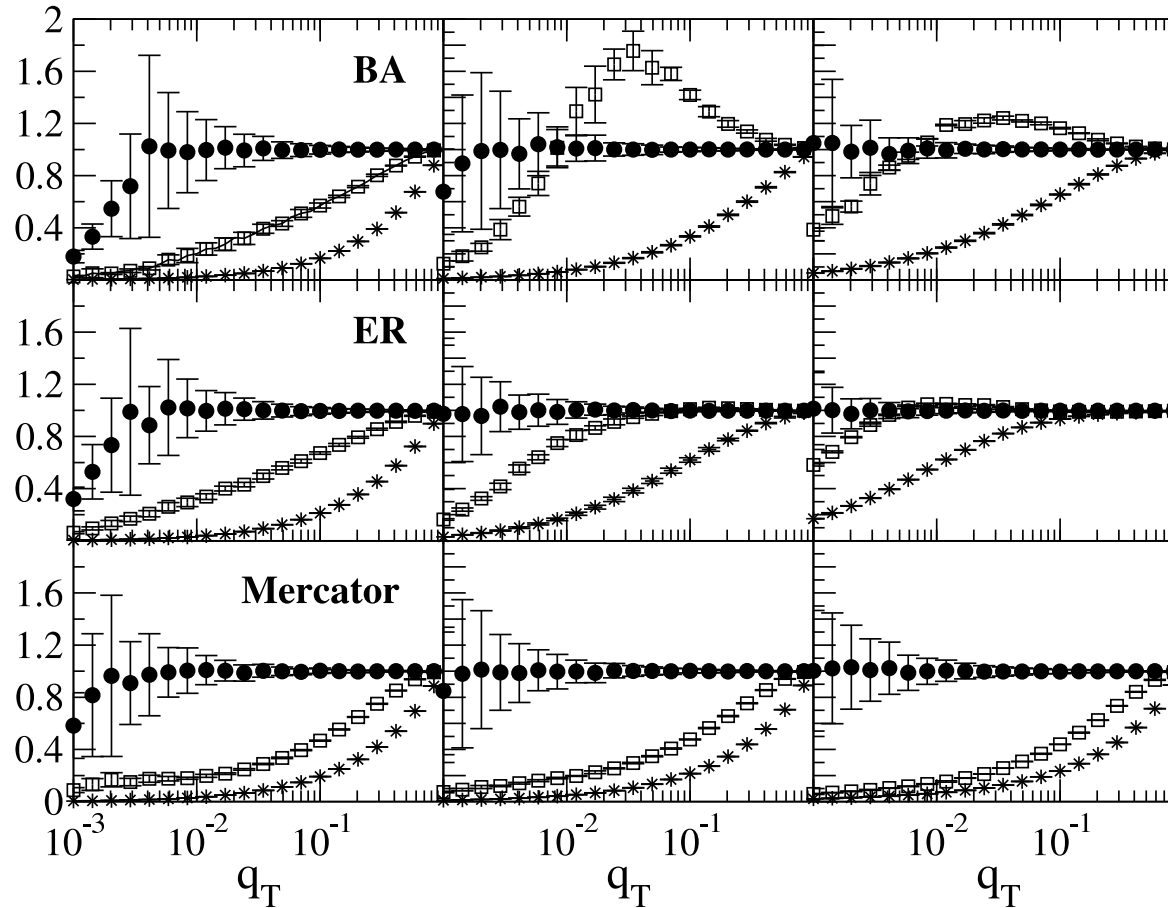
Define $N_{(-j)}^* = |V_{(-j)}^*|$. Write

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

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

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

Numerical Results



\hat{N}/N versus $q_T = n_T/N$

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

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
 \Rightarrow 1.64% response rate
- $\hat{N}_{ping} = 2^{32} \times 0.0164 \approx 70,583,787$ alive addresses

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

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 Cited

- 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.
- 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.

References Cited (cont.)

- 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. (2005). Network inference from traceroute measurements: Internet topology 'species'. <http://arxiv.org/abs/cs.NI/0510007/>