Statistical Analysis of Network Data

Lecture 2: Network Sampling

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu

BOSTON UNIVERSITY

Outline

BOSTON UNIVERSITY

Focus of Lectures

Remaining lectures:

- L1 Introduction, Background, and Descriptive Statistics (1.5hrs)
- L2 Network Sampling (1hr)
- L3 Network Modeling (1.5hrs)
- L4 Additional Topics in Modeling/Analysis (1.5hr)

In this lecture we'll concentrate on sampling as it relates to networks.

Two general directions in this area:

- when the network itself is the focus of study;
- when the network is used only as a tool for studying a population.

We will focus on the first of these two directions.

Outline

BOSTON UNIVERSITY

Network Sampling: 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 . . . right?

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 distinction is important!

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

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 $\mathrm{cl}(G)$
- Etc.

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

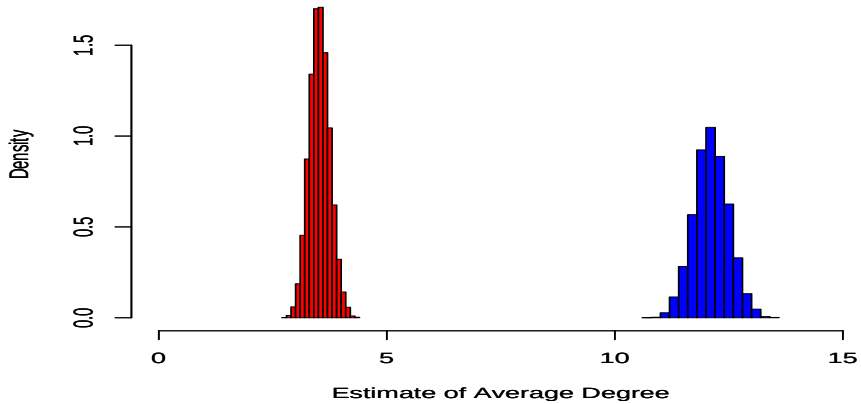
We conduct a small sampling experiment for illustration.

- Let G be a network of protein interactions in yeast[1].
- Take $\eta(G) =$ 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

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

[1] Corresponds to *S. Cerevisiae*; taken from BioGRID.

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 in this lecture primarily on effects of random sampling.

Perspective one of 'design-based' inference[2].

[2] As opposed to 'model-based' inference.

Outline

BOSTON UNIVERSITY

Background on Classical Sampling

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

  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 $\{i_1, \ldots, i_n\}$ from $\mathcal{U}$
- Observe the value y_{i_k} for $k = 1, \ldots, n$
- Form an estimator $\hat{\mu}$ of μ 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?

Let $\pi_i = \Pr\{$ Unit i 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 π'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.

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

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

for

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

Caveat Emptor: π_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 $\pi_{ij} = \Pr\{$ Units i and j 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 $\mathcal{P}(i,j)$,

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

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

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

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)

True . . . but caveat emptor

- Inclusion probabilites 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., $y_i = d_i$

Outline

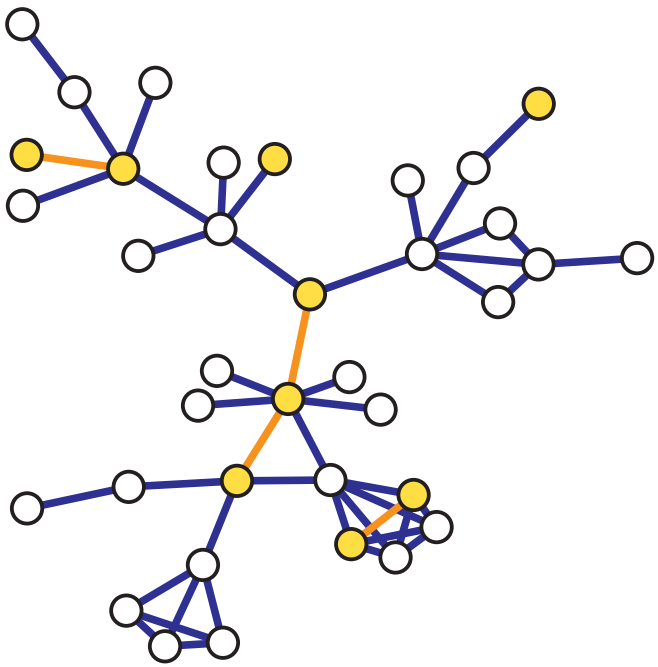
BOSTON UNIVERSITY

Four Common Network Sampling Designs

We'll look at four common network sampling designs and their inclusion probabilities π_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.

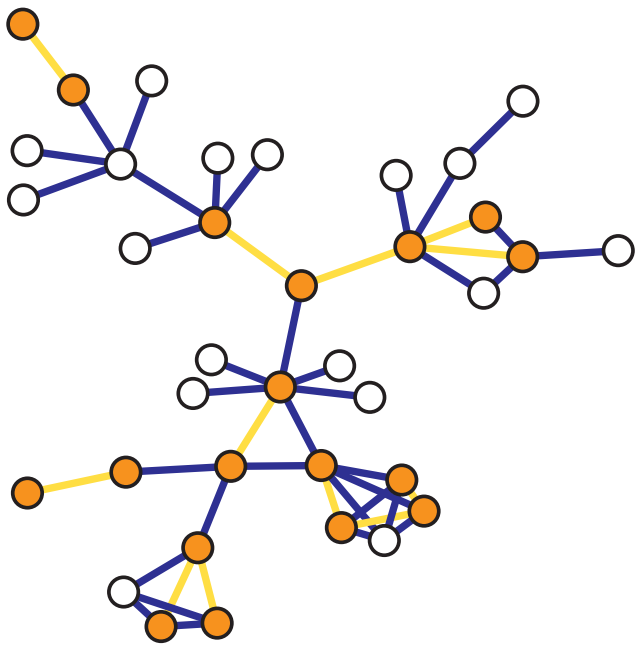
Induced Subgraph Sampling (cont.)

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

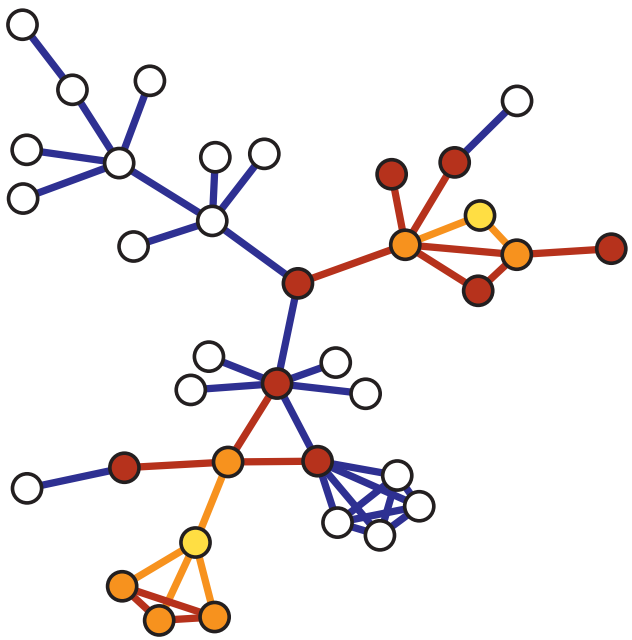
$$\pi_{\{i,j\}} = n/N_e \ .$$

Vertex inclusion probabilities take the form

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

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.

BOSTON UNIVERSITY

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

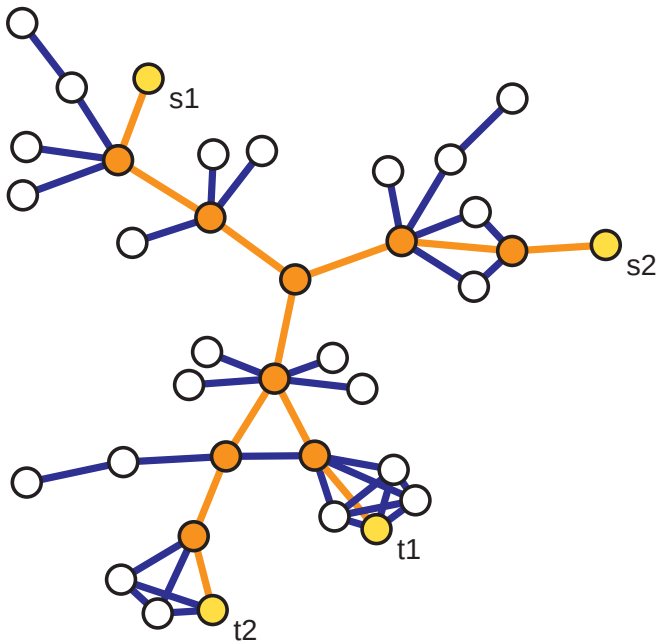
Design:

- Randomly select
  - a set of source nodes $S = \{s_1, \ldots, s_{n_S}\}$
  - a set of target nodes $T = \{t_1, \ldots, t_{n_T}\}$
- 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 $\{i,j\}$
- $\rho_S = n_S/N$; $\rho_T = n_T/N$

Outline

BOSTON UNIVERSITY

Estimation of Other Network Characteristics

Classical sampling theory rests heavily on Horvitz-Thompson framework.
$\Rightarrow$ 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*[3].

[3] See, for example, Lakhina *et al.* (2003), Clauset and Moore (2005), Achlioptas *et al.* (2005), Stumpf *et al.* (2005), and Han *et al.* (2005)

Impact of Sampling on Degree Distribution

Under a variety of sampling designs, the following holds:

$$E[\mathbf{N}^*] = P\mathbf{N} \ , \tag{1}$$

where

- $\mathbf{N} = (N_0, N_1, ..., N_M)$: the true degree vector, for
  N_i: the number of vertices with degree i in the original graph
- $\mathbf{N}^* = (N_0^*, N_1^*, ..., N_M^*)$: the observed degree vector, for
  N_i^*: the number of vertices with degree i in the sampled graph
- P is an $M+1$ by $M+1$ matrix operator, where
  M = maximum degree in the original graph

Estimating Degree Distribution: An Inverse Problem

Ove Frank (1978) proposed solving for the degree distribution by an unbiased estimator of N, defined as

$$\hat{\mathbf{N}}_{\text{naive}} = P^{-1}\mathbf{N}^* \quad . \tag{2}$$

There are two problems with this simple solution:

1. The matrix P is typically not invertible in practice.
2. The non-negativity of the solution is not guaranteed.

BOSTON UNIVERSITY

An Illustration

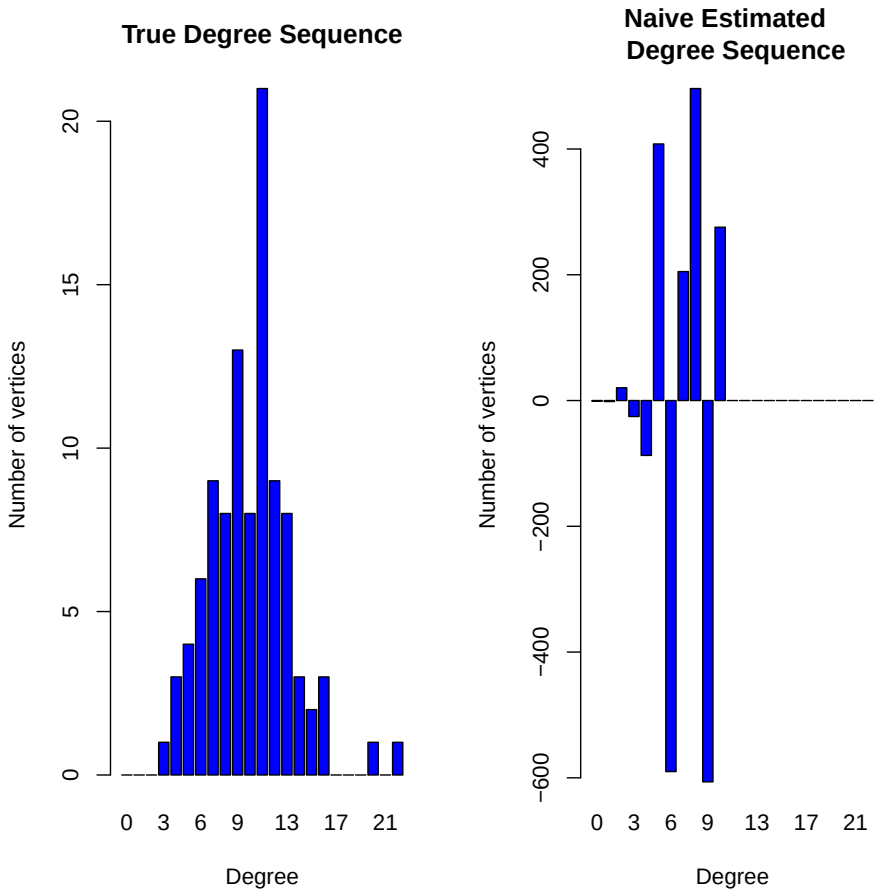


Figure : Left: ER graph with 100 vertices and 500 edges. Right: Naive estimate of degree distribution, according to equation (2). Data drawn according to induced subgraph sampling with sampling rate $p = 60\%$.

A Modern Variant: Constrained, Penalized WLS

We have recently proposed[4] a penalized weighted least squares with additional constraints.

$$\begin{aligned} \underset{\mathbf{N}}{\text{minimize}} \quad & (P\mathbf{N} - \mathbf{N}^*)^T C^{-1} (P\mathbf{N} - \mathbf{N}^*) + \lambda \cdot \text{pen}(\mathbf{N}) \\ \text{subject to} \quad & N_i \geq 0, \ i = 0, 1, \ldots M \\ & \sum_{i=0}^{M} N_i = n_v \ , \end{aligned} \tag{3}$$

where

- $C = \text{Cov}(\mathbf{N}^*)$,
- $\text{pen}(\mathbf{N})$ is a penalty on the complexity of $\mathbf{N}$,
- λ is a smoothing parameter, and
- n_v is the total number of vertices of the true graph.

[4] Zhang, Y., Kolaczyk, E.D., and Spencer, B.D. (2013). Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks . Revised for *Annals of Applied Statistics*. arxiv-1305.4977

Application to Online Social Networks

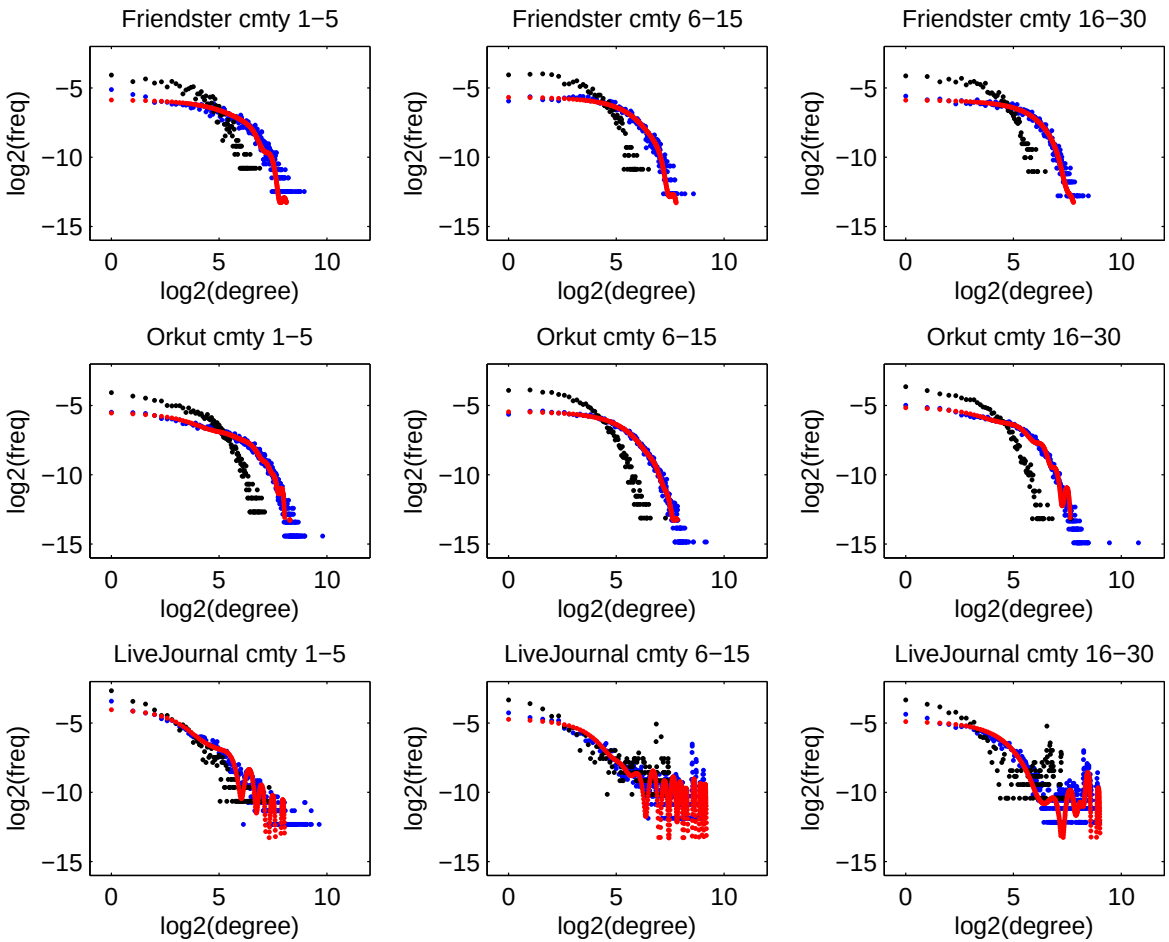


Figure : Estimating degree distributions of communities from Friendster, Orkut and Livejournal. Blue dots represent the true degree distributions, black dots represent the sample degree distributions, red dots represent the estimated degree distributions. Sampling rate=30%. Dots which correspond to a density $< 10^{-4}$ are eliminated from the plot.

Estimating Approximate Epidemic Thresholds: Friendster

- Moments of degree distributions can be used to obtain bounds of the network's epidemic threshold τ_c.
- An approximate threshold is give by the inverse of the largest eigenvalue λ_1 of the adjacency matrix (Mieghem, Omic, & Kooij '09).
- Simple bounds for λ_1 are

$$M_1 \leq \sqrt{M_2} \leq \lambda_1 \leq \sqrt{|E|} \quad (4)$$

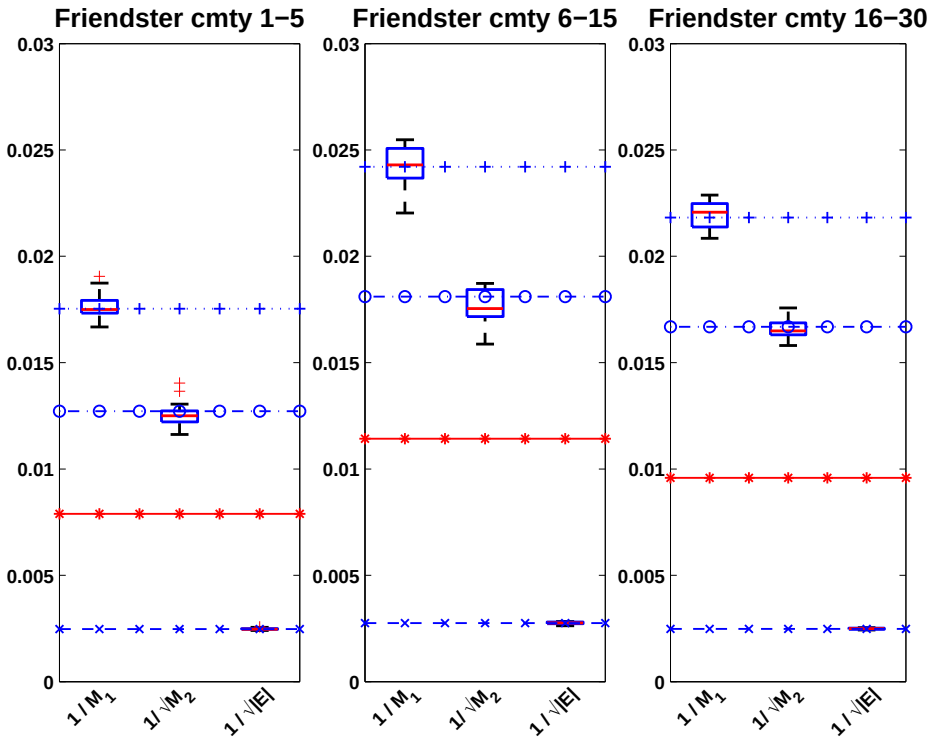


Figure : Estimated bounds for epidemic threshold in Friendster, based on 20 samples. Four horizontal lines are the true values for $\frac{1}{M_1}$, $\frac{1}{\sqrt{M_2}}$, $\frac{1}{\lambda_1}$ and $\frac{1}{\sqrt{|E|}}$ from top to bottom.

Additional Results: Orkut and LiveJournal

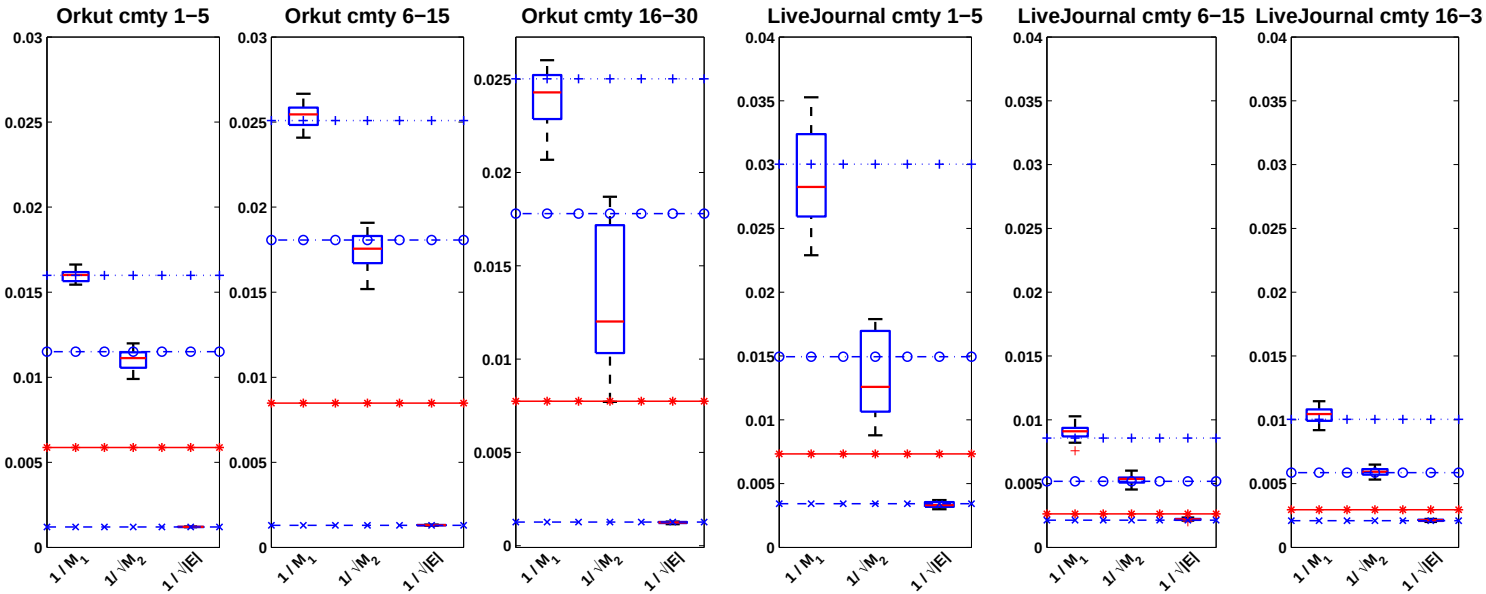


Figure : Estimated bounds for epidemic threshold in Orkut and LiveJournal, based on 20 samples. Four horizontal lines are the true values for $\frac{1}{M_1}$, $\frac{1}{\sqrt{M_2}}$, $\frac{1}{\lambda_1}$ and $\frac{1}{\sqrt{|E|}}$ from top to bottom.

Outline

BOSTON UNIVERSITY

Other Uses of Networks in Sampling

It may not always be the case that obtaining a network is the goal of a sampling study.

But networks can still be used in an often-advantageous manner for obtaining data in more standard survey situtations.

Examples include

- Estimating the size of 'hidden' populations
- "How many X do you know?"

Wrapping Up (cont.)

Remaining lectures:

- L1 Introduction, Background, and Descriptive Statistics (1.5hrs)
- L2 Network Sampling (1hr)
- L3 Network Modeling (1.5hrs)
- L4 Additional Topics in Modeling/Analysis (1.5hr)

BOSTON UNIVERSITY

References

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

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.

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.

BOSTON UNIVERSITY