Estimating network degree distributions from sampled networks: An inverse problem

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu

Network Graphs

It is common to represent networks – i.e., systems of inter-connected elements – with a graph $G = (V, E)$, of vertices $v \in V$ and edges $\{u, v\} \in E$ between them.

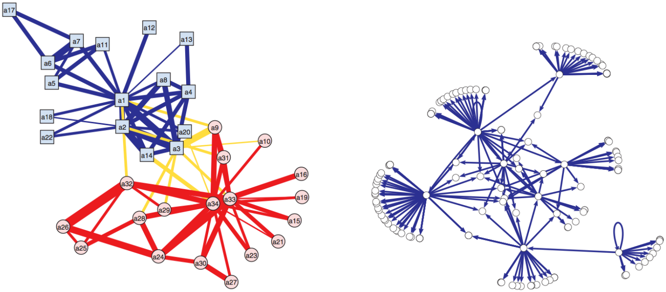


Figure : Zacharys karate club network (left) and AIDS Blog Network (right)

Network Sampling: Motivation

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?

BOSTON UNIVERSITY

Interpretation: Two Scenarios

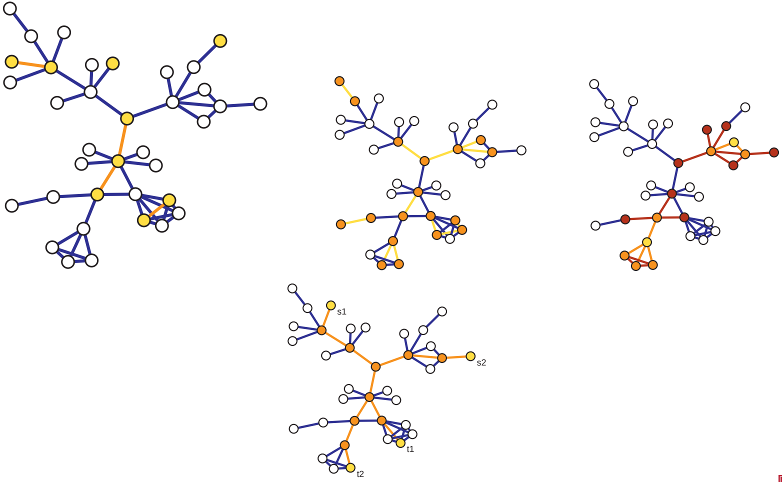
With respect to what frame of reference are the network characteristics interpreted?

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 Common Network Sampling Designs



Caveat emptor . . .

Completely ignoring sampling issues is equivalent to using 'plug-in' estimators.

The resulting bias(es) can be both substantial and unpredictable!

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

Lee *et al* (2006): Entries indicate direction of bias for vertex (red), edge (green), and snowball (blue) sampling.

The Degree Distribution

- The *degree* of a vertex[1] is the number of edges it shares with other vertices.
- The *degree distribution* is given by the relative frequency of these degrees over the whole network.
- As such, degree distributions are considered one of the most fundamental summary characteristics of a graph.
- **Our Objective**: Given a sub-graph $G^* \subset G$ observed through random sampling, estimate the degree distribution of G.

[1] For simplicity, we consider only undirected graphs. Extension to directed graphs is straightforward.

Some Notation

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.

An Illustration

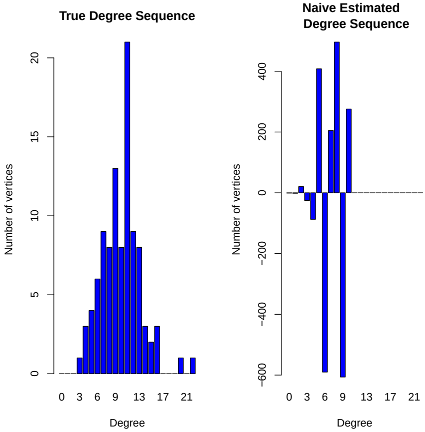


Figure : Left: ER graph with 100 vertices and 500 edges. Right: Naive estimate of degree distribution, according to equation (2). Data drawn according to induced subgraph sampling with sampling rate $p = 60\%$.

Also . . . Degree Distributions Can Take Many Forms!

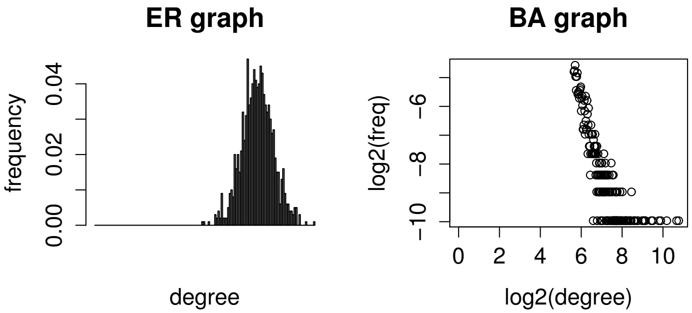


Figure : Erdős-Rényi model (left) and Barabási-Albert model (right)

Our Contributions

- Characterization of the problem as an *ill-posed linear inverse problem*.

- Development of a constrained, penalized least-squares estimator.
- Smoothing parameter selection through Monte Carlo SURE.
- Illustration through simulation and application to social media data.

BOSTON UNIVERSITY

Sampling Design

Our focus is on the contexts where the matrix P fully depends on the sampling design.

Designs of interest include

- Ego-centric and one-wave snow-ball sampling,
- Induced and incident subgraph sampling,
- Random walk and other exploration-based methods.

Characterization Through the SVD

The *singular value decomposition* can be used to better understand the nature of the operator P in our linear inverse problem.

Let $P = UDV^T$, where $D = \text{diag}(d_0, d_1, \cdots, d_M)$ is a diagonal matrix of singular values, and $U = (\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_M)$, $V = (\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_M)$ are orthogonal matrices of the left- and right-singular vectors, respectively.

Then

$$\hat{\mathbf{N}}_{\text{naive}} = \sum_{i=0}^{M} \left[\frac{1}{d_i} \mathbf{u}_i^T \mathbf{N}^* \right] \mathbf{v}_i \quad (3)$$

decomposes the naive estimator (2) into a linear combination of the right singular vectors of P.

Ego-centric and One-wave Snow-ball Sampling

For *ego-centric sampling*, the operator P is a diagonal matrix with the sampling rate p at each diagonal position, i.e.,

$$P_{\text{ego}}(i,j) = \begin{cases} p & \text{for } i = j = 0, 1, \cdots, M \\ 0 & \text{for } i, j = 0, \cdots, M; \quad i \neq j \end{cases} . \tag{4}$$

The operator P for *one-wave snow-ball sampling* is

$$P_{\text{snow}}(i,j) = \begin{cases} 1 - (1-p)^{i+1} & \text{for } i = j = 0, 1, \cdots, M \\ 0 & \text{for } i, j = 0, \cdots, M; \quad i \neq j \end{cases} . \tag{5}$$

Ego-centric and One-wave Snow-ball Sampling (Cont.)

- In both cases,
  - the singular values are equal to the diagonal elements, and
  - both the left and right singular vectors are just the canonical basis vectors.
- P_{ego} is not ill-conditioned at all, since $P_{\text{ego}} = I \times p$.
- The condition number of P_{snow} is equal to

$$\frac{P_{\text{snow}}(M,M)}{P_{\text{snow}}(0,0)} = \frac{1-(1-p)^{M+1}}{1-(1-p)} = \frac{1-(1-p)^{M+1}}{p} \quad , \tag{6}$$

  In the case where p is fixed, as M increases, the condition number is upper bounded by $\frac{1}{p}$.
  On the other hand, if $Mp = o(1)$, the condition number $\sim (M+1)$.

BOSTON UNIVERSITY

Induced and Incident Subgraph Sampling

The P matrix for *induced subgraph sampling* is

$$P_{\text{ind}}(i,j) = \begin{cases} \binom{j}{i} p^{i+1}(1-p)^{j-i} & \text{for } 0 \leq i \leq j \leq M \\ 0 & \text{for } 0 \leq j < i \leq M \end{cases}, \tag{7}$$

while that for *incident subgraph sampling*[2] is

$$P_{\text{inc}}(i,j) = \begin{cases} \binom{j}{i} p^{i}(1-p)^{j-i} & \text{for } 1 \leq i \leq j \leq M \\ 0 & \text{for } 0 \leq j < i \leq M \end{cases}. \tag{8}$$

[2] For incident subgraph sampling the index i starts from 1, because there are no isolated vertices in the sample.

Induced and Incident Subgraph Sampling (Cont.)

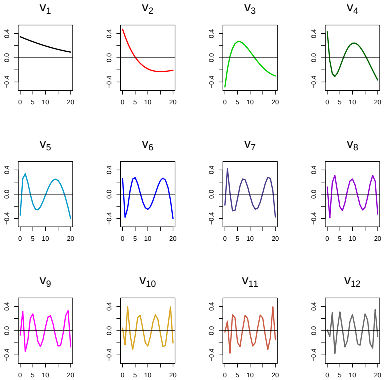


Figure : Right singular vectors: maximum degree $M = 20$, sampling rate $p = 0.2$

Induced and Incident Subgraph Sampling (Cont.)

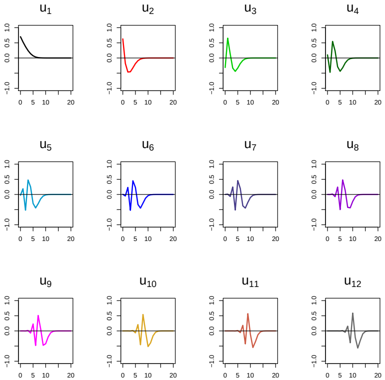


Figure : Left singular vectors: maximum degree $M = 20$, sampling rate $p = 0.2$

Induced and Incident Subgraph Sampling (Cont.)

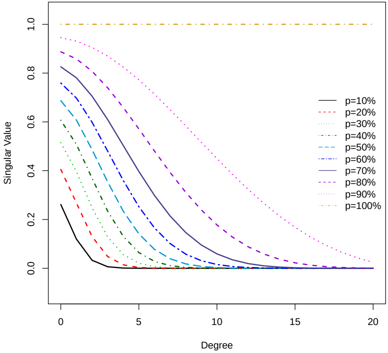


Figure : Singular values decay under Induced Subgraph sampling. $M = 20$.

Induced and Incident Subgraph Sampling (Cont.)

- While it would be desirable to have an analytical expression for the singular vectors under induced/incident subgraph sampling, we are unable to produce one.
- However, it is possible to produce expressions for the eigenfunctions of P_{ind}, as solutions to the non-symmetric eigen-decomposition $P_{ind} = \tilde{U}\Lambda\tilde{U}^{-1}$.

Random Walk and Other Exploration-based Methods

- If we consider a random walk sampling over a non-bipartite, connected, undirected graph, once the steady state is reached, it shares an important property with incident subgraph sampling with SRS of edges, in that both sample edges uniformly at random (Ribeiro and Towsley, 2010).
- Thus

$$P_{\mathrm{RW}}(i,j) = \begin{cases} \binom{j}{i}\binom{n_e-j}{n_e^*-i}\binom{n_e}{n_e^*}^{-1} & \text{for } 1 \leq i \leq j \leq M \\ 0 & \text{for } 0 \leq j < i \leq M \end{cases} . \tag{9}$$

where n_e is the total number of edges in the true network, n_e^* is the number of edges selected in the sample.

- With respect to the nature of the inverse problem that we study here, we may categorize this sampling plan with the induced and incident subgraph sampling plans described above.

BOSTON UNIVERSITY

A Regression-based Perspective

N^* can be thought of as a *'noisy'* observation of N.

Our numerical and analytical work suggests two possible models:

- *Normal Model:*

$$\mathbf{N}^* = P\mathbf{N} + \epsilon \tag{10}$$

- *Poisson Model*

$$\mathbf{N}^* = Pois(P\mathbf{N}) \tag{11}$$

Our goal then becomes one of recovering **N** through regression.

BOSTON UNIVERSITY

Modeling the 'Noise'

For *ego-centric sampling*, a vertex is observed to have degree k if and only if the vertex is selected through Bernoulli sampling and also has degree k in the true graph.

Therefore

$$N_k^* = \sum_{\{u: d_u = k\}} I\{u \in V^*\} \ , \tag{12}$$

Thus the distribution of the N_k^* is that of $M+1$ independent binomials, i.e. $N_k^* \sim \mathrm{Bin}(p, N_k)$.

$\Longrightarrow$ Nonconstant variance a concern,
especially for heterogeneous degree distributions.

BOSTON UNIVERSITY

Modeling the 'Noise' (Cont.)

- For one-wave snowball sampling, the representation (12) still applies. However, the indicator functions are not independent.
- For induced-subgraph sampling, we can write

$$N_k^* = \sum_{r=k}^{M} \sum_{u=1}^{n_v} I\{u \in V^*, d_u^* = k, d_u = r\} \quad . \tag{13}$$

- Under these two sampling methods, a Chen-Stein argument shows that the Poisson model is a good approximation under low sampling rate.

BOSTON UNIVERSITY

Solving An Ill-posed Inverse Problem

- These observations suggest approaching the estimation of $\mathbf{N}$ as an ill-posed linear inverse problem.
- Inverse problems are well-studied in literature, with contributions from mathematics, statistics, signal/image processing, geology, etc.
- Penalized least-squares solutions are the most common approach. Need to match
  1. loss-function to noise, and
  2. penalty function to nature of object to be recovered.

We pursue a constrained, penalized weighted least-squares approach.

BOSTON UNIVERSITY

Constrained, Penalized WLS

We use penalized weighted least squares with additional constraints.

$$\begin{aligned} \underset{\mathbf{N}}{\text{minimize}} \quad & (P\mathbf{N} - \mathbf{N}^*)^T C^{-1} (P\mathbf{N} - \mathbf{N}^*) + \lambda \cdot \text{pen}(\mathbf{N}) \\ \text{subject to} \quad & N_i \geq 0, \; i = 0, 1, \ldots M \\ & \sum_{i=0}^{M} N_i = n_v \quad , \end{aligned} \tag{14}$$

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.

BOSTON UNIVERSITY

Penalty Function

We assume a smooth true degree distribution, and therefore adopt a penalty of the form

$$\|D\mathbf{N}\|_2^2 \ ,$$

where the matrix D represents a second-order differencing operator, i.e.,

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & \ldots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -2 & 1 \end{bmatrix} \ . \qquad (15)$$

This choice, in the discrete setting, is analogous to the use of a Sobolev norm with nonparametric function estimation in the continuous setting.

BOSTON UNIVERSITY

Smoothing Parameter Selection

- *Cross validation* appears not to work well in this setting. Why?
  - The elements in the observed degree vector are not *i.i.d.*
- *Stein's Unbiased Risk Estimation (SURE)* estimates MSE for i.i.d. Gaussian case. However, in our setting
  - the elements in the observed degree vector are not *i.i.d.* , and
  - the operator P is rank deficient .
- Our Solution:
  - Yonina C. Eldar (2008) extended *SURE* to general exponential families.
  - We define a weighted mean square error (WMSE) in the observation space as

$$WMSE(\hat{\mathbf{N}}, \mathbf{N}) = E\left[(P\mathbf{N} - P\hat{\mathbf{N}})^T C^{-1} (P\mathbf{N} - P\hat{\mathbf{N}})\right] \quad . \qquad (16)$$

Smoothing Parameter Selection (Cont.)

- An unbiased estimate of *MSE* is given by

$$\widehat{WMSE}(\hat{\mathbf{N}}, \mathbf{N}) = (P\mathbf{N})^T C^{-1} P\mathbf{N} + (P\hat{\mathbf{N}})^T C^{-1} P\hat{\mathbf{N}} + 2\left\{\text{Trace}\left(P\frac{\partial \hat{\mathbf{N}}}{\partial \mathbf{N}^*}\right)\right\} - 2(P\hat{\mathbf{N}})^T C^{-1}\mathbf{N}^* \, .$$

- The *Monte-Carlo* technique proposed by Ramani, Blu, and Unser '08 can be used to compute $\text{Trace}\left(P\frac{\partial \hat{\mathbf{N}}}{\partial \mathbf{N}^*}\right)$.

Approximating div: Principles

Denote the solution to the optimization problem in (14) as $\hat{\mathbf{N}} = f_\lambda(\mathbf{N}^*)$, a function of $\mathbf{N}^*$, indexed by λ.

Let b be a vector with zero mean, covariance matrix I (that is independent of $\mathbf{N}^*$) and bounded higher order moments. Then under mild conditions,

$$\text{div} \equiv \text{Trace}\left(P\frac{\partial \hat{\mathbf{N}}}{\partial \mathbf{N}^*}\right) = \lim_{\epsilon \to 0} E_b\left\{b^T P\left(\frac{f_\lambda\left(\mathbf{N}^* + \epsilon \mathbf{b}\right) - f_\lambda\left(\mathbf{N}^*\right)}{\epsilon}\right)\right\}\quad . \tag{17}$$

Approximating div: Algorithm

Let $\mathbf{b_i}$ be the realization of $\mathbf{b}$ at each simulation.

The algorithm for estimating div$=$ Tr $\left(P\frac{\partial \hat{\mathbf{N}}}{\partial \mathbf{N}^*}\right)$ and computing of $\widehat{WMSE}$ for a given $\lambda = \lambda_0$ and fixed ϵ is as follows:

1. $\mathbf{y} = \mathbf{N}^*$
2. For $\lambda = \lambda_0$, evaluate $f_\lambda(\mathbf{y})$; $i = 1$; div $= 0$
3. Build $\mathbf{z} = \mathbf{y} + \mathbf{b_i}$; Evaluate $f_\lambda(\mathbf{z})$ for $\lambda = \lambda_0$
4. div=div$+\frac{1}{\epsilon}\mathbf{b_i}^T P(f_\lambda(\mathbf{z}) - f_\lambda(\mathbf{y}))$; $i = i + 1$
5. If $(i \leq K)$ go to Step 3; otherwise evaluate sample mean: div $=$ div$/K$ and compute $\widehat{WMSE}(\lambda_0)$ using eqn (17).

BOSTON UNIVERSITY

Simulation Study: Ego-Centric Sampling

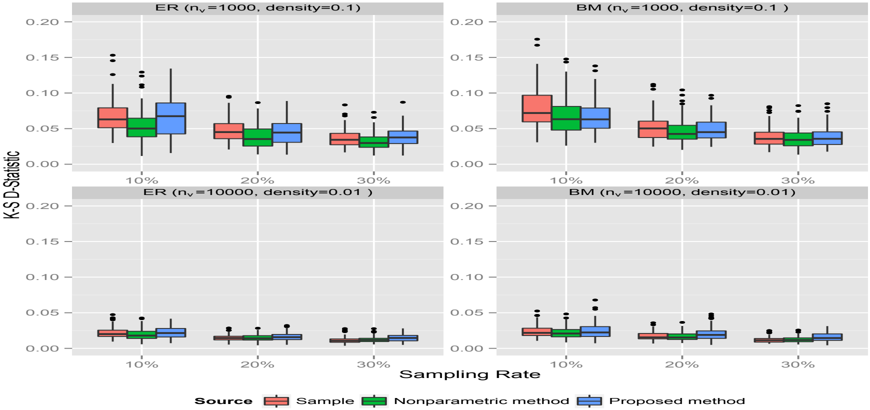


Figure : Simulation results for ego-centric sampling. Error measured by K-S D-Statistic.

Simulation Study: One-Wave Snowball Sampling

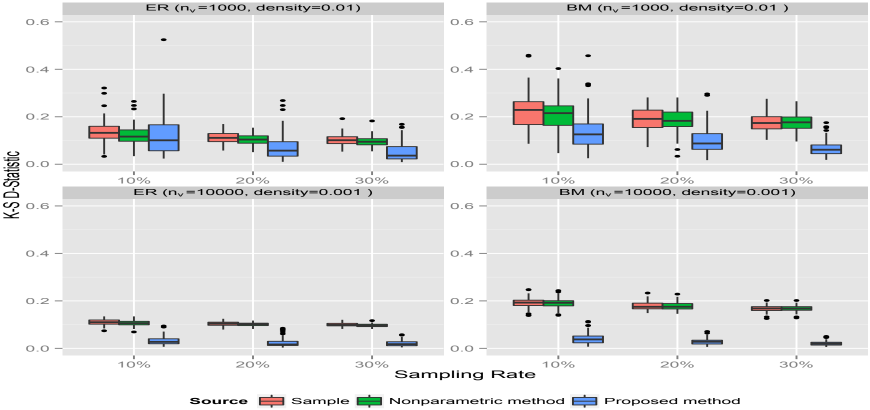


Figure : Simulation results for one-wave snowball sampling. Error measured by K-S D-Statistic.

Simulation Study: Induced Subgraph Sampling

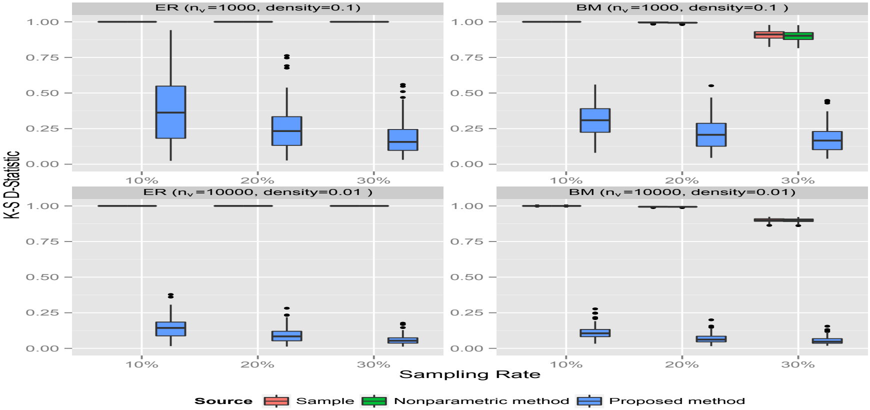


Figure : Simulation results for induced subgraph sampling. Error measured by K-S D-Statistic.

Application to Online Social Networks

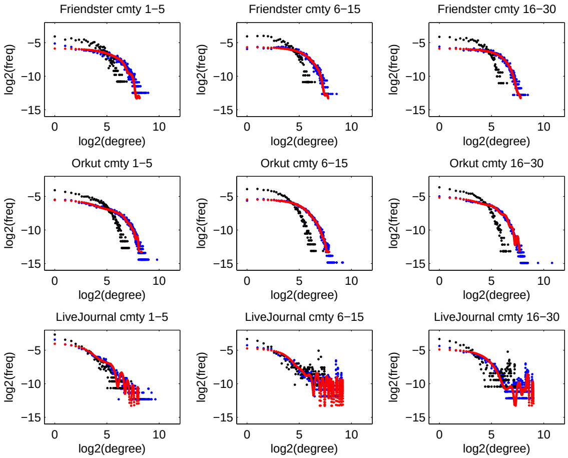


Figure : Estimating degree distributions of communities from Friendster, Orkut and Livejournal. Blue dots represent the true degree distributions, black dots represent the sample degree distributions, red dots represent the estimated degree distributions. Sampling rate=30%. Dots which correspond to a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

Approximating an Epidemic Threshold

Moments of degree distributions can be used to obtain bounds of the network's epidemic threshold τ_c, which is relevant to viral marketing in online social networks, etc.

- For infection rate β and cure rate δ, an effective spreading rate $\tau = (\beta/\delta) > \tau_c$ means the virus persists and a nontrivial fraction of the nodes are infected, whereas for $\tau \leq \tau_c$ the epidemic dies out.
- This threshold is shown to equal the inverse of the largest eigenvalue λ_1 of the network's adjacency matrix in (Mieghem, Omic and Kooij, 2009) using mean field theory.
- We can bound λ_1 using functions of the first and second moments of the degree distribution $M1, M2$, and the total number of edges $n_e = |E|$.The relationship is,

$$M_1 \leq \sqrt{M_2} \leq \lambda_1 \leq U \quad , \tag{18}$$

where $U = (2 * n_e(n_v - 1)/n_v)^{1/2}$.

We estimate these bounds using our estimated degree distributions.

Friendster

Our method estimates the bounds pretty much right on target, whereas using the sampled data is way off.

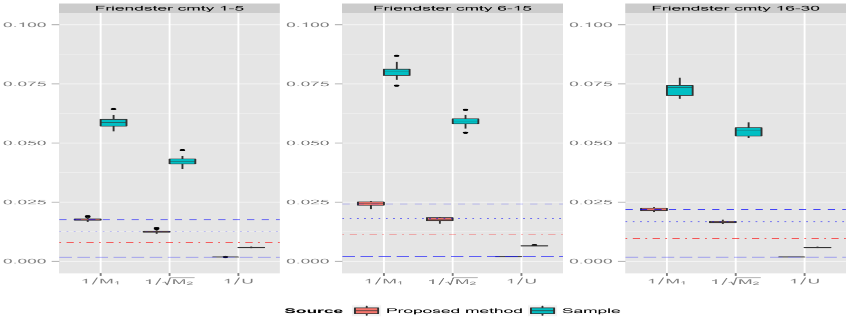
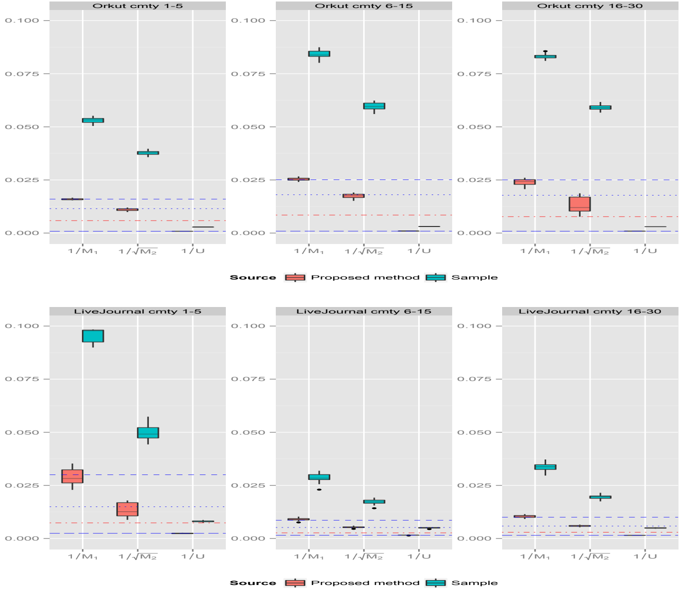


Figure : Estimated bounds for epidemic threshold in Friendster, based on 20 samples. Four horizontal lines are the true values for $\frac{1}{M_1}$, $\frac{1}{\sqrt{M_2}}$, $\frac{1}{\lambda_1}$ and $\frac{1}{U}$ from top to bottom.

Orkut and LiveJournal



Final Thoughts

- Original proposed solution to this problem was 35 years ago.
- Key insight allowing new progress is observing connection to ill-posed linear inverse problems, and leveraging modern machinery.
- Ongoing work includes:
    - Theoretical characterization of performance.
    - Extension to estimation of node degrees (much harder!)
    - Generalization to adaptive sampling plans

Thank you!

This is a joint work with

Yaonan Zhang
Department of Mathematics and Statistics, Boston University
Bruce Spencer
Department of Statistics, Northwestern University

This work is supported by **NSF** and **AFOSR**.