Statistical Analysis of Network Data

Lecture 2 – Network Sampling & Modeling

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu

BOSTON UNIVERSITY

Outline

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.

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, or
   - an ensemble of networks.

The distinction is important!

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

Plan for this Lecture

In this lecture we will discuss

1. Network sampling
   A mix of classical and recent / ongoing work.
2. Network modeling
   - Mathematical / probabilistic modeling
   - Statistical modeling

BOSTON UNIVERSITY

Outline

BOSTON UNIVERSITY

Common Network Sampling Designs

Suppose that we observe a graph G^* that is a subgraph of some true underlying graph G, via sampling of the network[1]

Viewed from the perspective of classical statistical sampling theory, the network sampling design is important.

Examples include

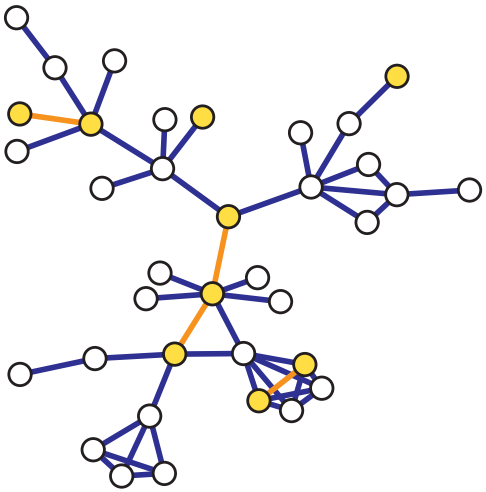
- Induced Subgraph Sampling
- Incident Subgraph Sampling
- Snowball Sampling
- Link Tracing

[1] That is, we take a design-based perspective.

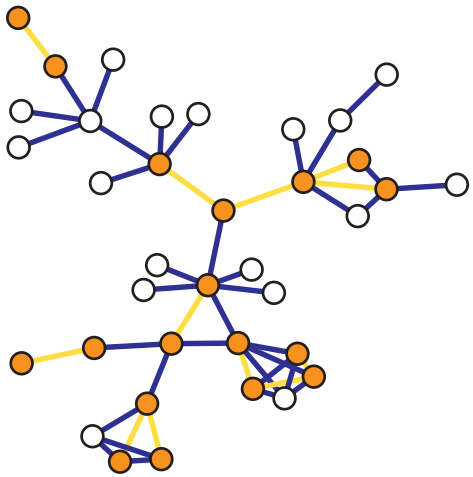
BOSTON UNIVERSITY

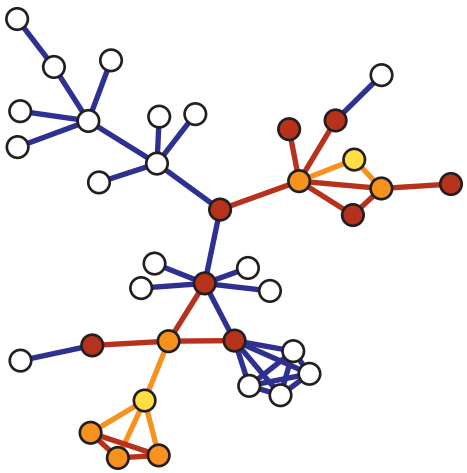
Common Network Sampling Designs (cont.)

Induced Subgraph Sampling

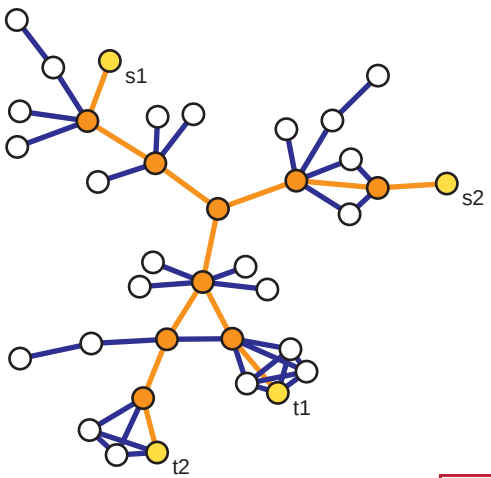


Incident Subgraph Sampling





Snowball Sampling



Traceroute Sampling

BOSTON UNIVERSITY

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 induced subgraph (red), incident subgraph (green), and snowball (blue) sampling.

BOSTON UNIVERSITY

Accounting for Sampling Design

Accounting for sampling design can be non-trivial.

Classical work goes back to the 1970's (at least), with contributions of Frank and colleagues, based mainly on Horvitz-Thompson theory.

More recent resurgence of interest, across communities, has led to additional studies using both classical and modern tools.

See Kolaczyk (2009), Chapter 5.

Illustration: 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)[2], studying the problem of estimating N_v.

[2] 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

A 'Leave-One-Out' Estimator

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

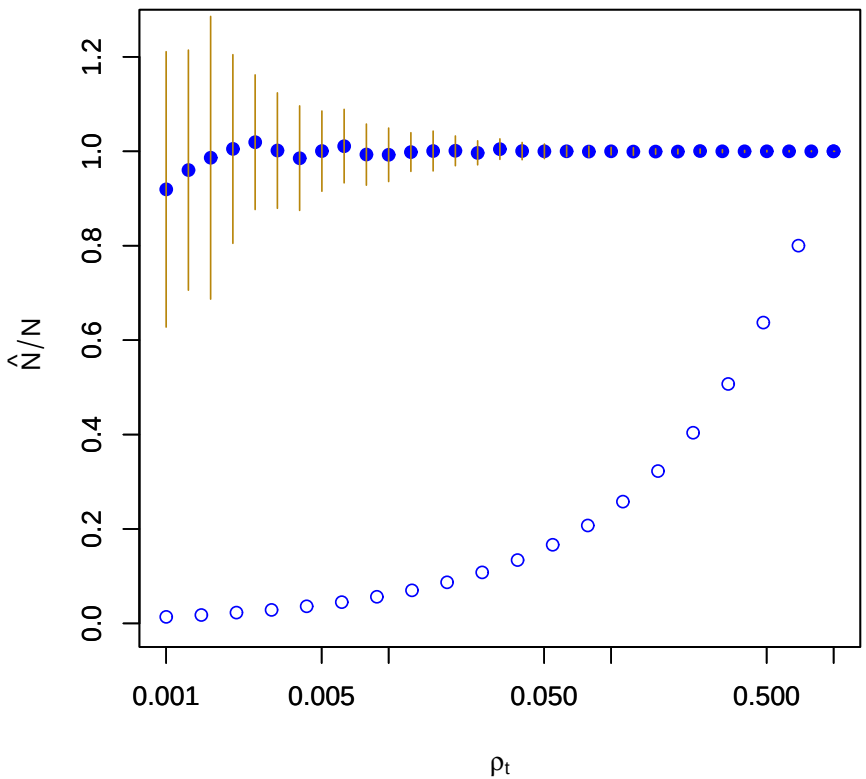
Assumptions: Low marginal rate of return from any single target node; simple random sampling of targets.

Formal argument, based on leave-one-out principles, 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.

Numerical Illustration on Internet Data



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 ρ_t.

Estimaging Degree Distribution Under Sampling: An Inverse Problem

More ambitious is to estimate the full degree distribution of G, from sampled G^*, rather than just a single numerical summary of G.

For many sampling designs, this problem can be usefully formulated as a . . . potentialy ill-posed . . . linear inverse problem[3], since

$$E[N(S)] = PN. \tag{1}$$

where

- $N = (N_1, N_2, ..., N_M)$: the true degree vector
- $S \subset V$ is a set of sampled vertices
- $N(S) = (N_1(S), N_2(S), ..., N_M(S))$: the observed degree vector
- P: $M + 1$ by $M + 1$ matrix capturing sampling effects.

[3] Zhang, Kolaczyk, and Spencer (2013). Manuscript.

BOSTON UNIVERSITY

Sampling Design

Operator P is assumed to depend fully on the sampling design, where

$$P(i,j) \quad = \quad \text{probability that a vertex with degree } j \text{ in the original graph is selected and has degree } i \text{ in the subgraph.}$$

We explore:

- Induced Sub-graph Sampling: vertices V^* are selected by Bernoulli(p), all edges between sampled vertices are observed.

$$P(i,j) = \binom{j}{i} p^{i+1}(1-p)^{j-i} \tag{2}$$

- Other designs of interest include: Induced Subgraph sampling under SRS, Incident Subgraph Sampling, Star Sampling, One-wave Snowball, Random Walk.

BOSTON UNIVERSITY

Frank (1978)

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

$$\hat{N} = P^{-1}N(S). \tag{3}$$

There are 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.

Constrained Penalized LS

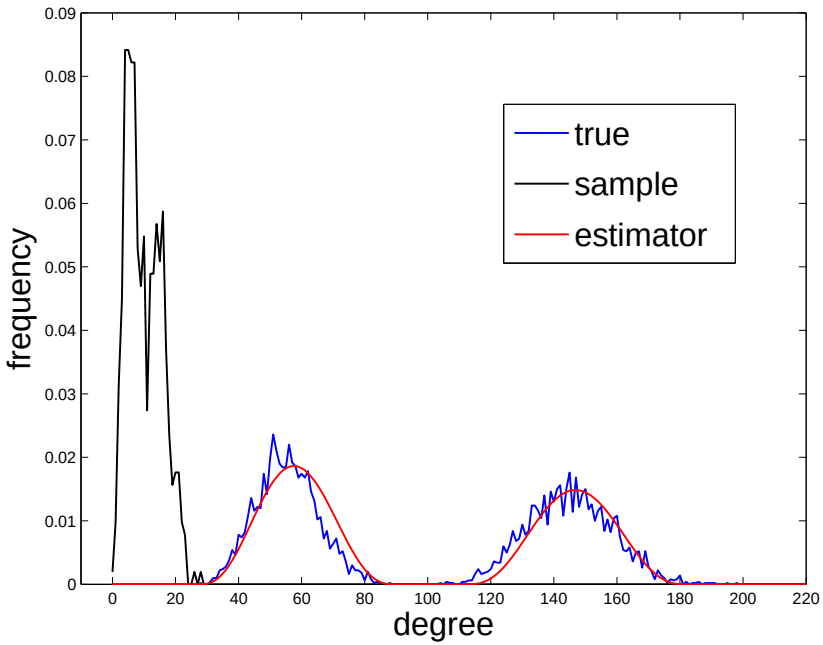
- We use penalized Weighted Least Squares with additional constraints of non-negativity and total number of vertices.
- Let C be the covariance matrix of $N(S)$, N_v be the total number of vertices in the true graph

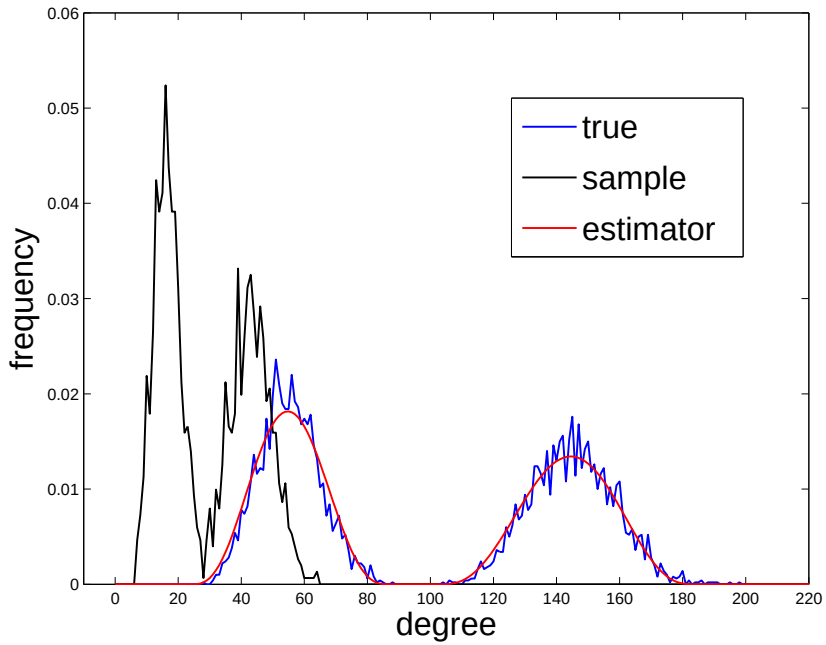
$$\begin{aligned} \underset{N}{\text{minimize}} \quad & (PN - N(S))^T C^{-1} (PN - N(S)) + \lambda \cdot \text{penalty}(N) \\ \text{subject to} \quad & N_i \geq 0, \ i = 0, 1, \ldots M \\ & \sum_{i=0}^{M} N_i = N_v. \end{aligned} \tag{4}$$

- Selection of λ through Monte Carlo SURE

BOSTON UNIVERSITY

Illustration: Sampling a Stochastic Block Model





Two-module stochastic block models, $N_v = 5000$, with density 0.08, under 10% (left) and 30% (right) sampling.

BOSTON UNIVERSITY

Outline

Why Model Networks?

Generally speaking, a network graph model takes the form

$$\{ \mathbb{P}_\theta(G), G \in \mathcal{G} : \theta \in \Theta \}$$

Models used to

- propose/study mechanisms for producing observed characteristics
- test 'significance' of observed characteristic(s)
- assess association between network structure and node/edge attributes
- create toy versions of network structures for modeling network-indexed processes (e.g., epidemics, flows, etc.)

BOSTON UNIVERSITY

Types of Network Models

Broadly speaking, we can think of two classes of network models:

1. Mathematical/Probabilistic Models: Emphasis is on constructions that allow mathematical analysis, e.g., of structural properties.
2. Statistical Models: Emphasis is on constructions that allow for statistical fitting and interpretation.

We'll take a (quick!) look at examples of each.

Random Graphs

Generally 'random graph' is used to refer to graphs drawn uniformly from some collection/ensemble $\mathcal{G}$.

Common to distinguish between

- classical, and
- generalized.

Main advantage is ability (with work!) to derive various properties of graphs in the ensemble.

Classical Random Graphs

In a series of seminal papers, Erdös and Rényi established and explored the model where

- $\mathcal{G}_{N_v,N_e} = \{G = (V, E) : |V| = N_v \text{ and } |E| = N_e\}$, and
- $\mathbb{P}(G) = \binom{N}{N_e}^{-1}$, for each $G \in \mathcal{G}_{N_v,N_e}$

where $N = \binom{N_v}{2}$ is the total number of distinct vertex pairs.

Such graphs can, under appropriate conditions,

- have a connected component of order $O(N_v)$;
- have a Poisson degree distribution (asymptotically);
- have small diameter (i.e., $O(\log N_v)$);
- be sparse and have low clustering.

BOSTON UNIVERSITY

Generalized Random Graphs

Classical random graph models put equal mass on all G of a fixed order N_v and size N_e.

Generalized random graph models equip $G \in \mathcal{G}$ with other characteristics.

Most common choice: a fixed degree sequence[4]

$$\{d_{(1)}, \ldots, d_{(N_v)}\} \ .$$

Note: Given N_v, this fixes N_e as well, since $\bar{d} = 2N_e/N_v$.

[4] Or, asymptotically equivalent, a fixed expected degree distribution.

BOSTON UNIVERSITY

Using Random Graph Models in Statistics

Random graph models are useful as more than just theoretical play-things.

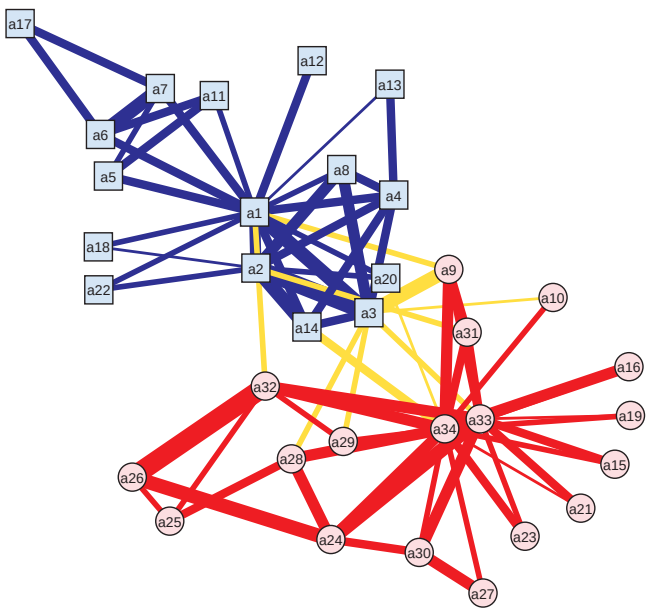
Within statistics, they've been used for

- model-based inference in network sampling[5]
  (e.g., estimation of the size of 'hidden' populations)
- assessment of significance of network graph characteristics

[5] See Chapter 6.2.4.1 in the Kolaczyk (2009).

BOSTON UNIVERSITY

Example: Clustering in the Karate Club Network



The fraction of connected triples that close to form triangles, say $\mathrm{cl}_T(G)$, is called the *clustering coefficient* (or *transitivity*) of G.

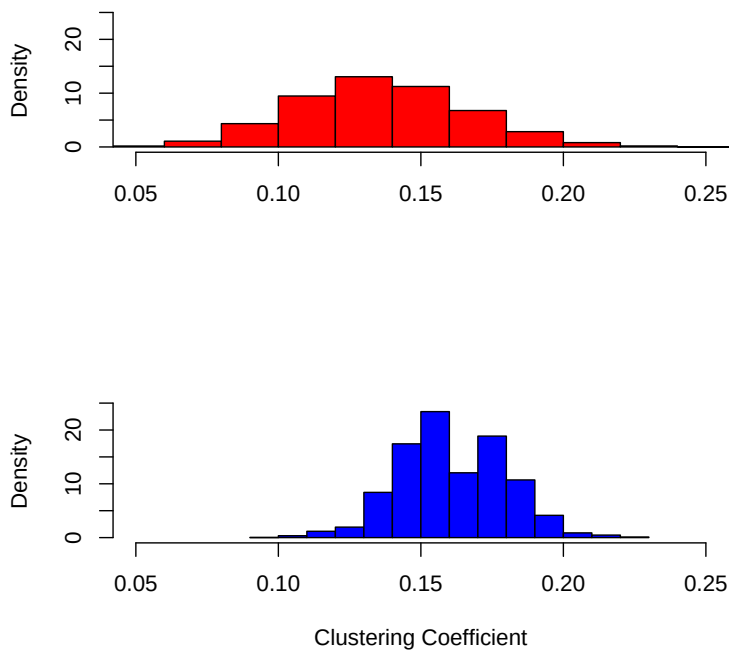
In the Karate Club network, $\mathrm{cl}_T(G) = 0.2557$.

Question: Is this 0.2557 significant?

Answer: Uhm ... 'significant' in what sense?!

BOSTON UNIVERSITY

Example: Clustering in the Karate Club Network (cont.)



Can use random graph models to create reference distributions against which to compare an observed network summary statistic.

Consider

1. $\mathcal{G}$ has fixed $N_v = 34$ and $N_e = 78$,
2. $\mathcal{G}$ has fixed $N_v = 34$ and same degree distribution as Karate club

Results: $\text{cl}_T(G) \geq \text{cl}_T(G^{Obs})$ only 3 and 0 times, out of 10,000.

BOSTON UNIVERSITY

Other Classes of Mathematical/Probabilitistic Random Graphs

A major shift in focus in network modeling, from classical to modern era, is the movement towards models explicitly designed to mimic observed 'real-world' properties.

Canonical examples include

- Small-world models
- Network-growth models (e.g., preferential attachment, copying, etc.)

All still, however, maintain the somewhat 'toy-like' nature we've seen.

BOSTON UNIVERSITY

Mathematical/Probabilistic vs. Statistical Network Models

The models seen so far serve various useful purposes, but arguably come up short as statistical models.

> *"A good [statistical network graph] model needs to be both estimable from data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best explanation of the data."*
>
> *Robins & Morris (2007)*

Probabilistic vs. Statistical Network Models (cont.)

Roughly speaking, there are network-based versions of three canonical classes of statistical models:

1. regression models (i.e., ERGMs)
2. mixed effects models (i.e., latent variable models)
3. mixture models (i.e., stochastic blocks models)

We will take a brief look at each.

Exponential Random Graph Models

Let $G = (V, E)$ be a random graph, and $\mathbf{Y} = [Y_{ij}]$, the corresponding random adjacency matrix.

An *exponential random graph model (ERGM)* for $\mathbf{Y}$ has the form

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) \exp\left\{\sum_H \theta_H \, g_H(\mathbf{y})\right\} \quad ,$$

where

(i) each H is a *configuration*, which is defined to be a set of possible edges among a subset of the vertices in G;

(ii) $g_H(\mathbf{y}) = \prod_{y_{ij} \in H} y_{ij}$, and is therefore either one if the configuration H occurs in $\mathbf{y}$, or zero, otherwise;

(iii) a non-zero value for θ_H means that the Y_{ij} are dependent for all pairs of vertices $\{i, j\}$ in H, conditional upon the rest of the graph; and

(iv) $\kappa = \kappa(\theta)$ is a normalization constant,

$$\kappa(\theta) = \sum_{\mathbf{y}} \exp\left\{\sum_H \theta_H \, g_H(\mathbf{y})\right\} \quad .$$

ERGMs (cont.)

ERGMs have a natural appeal, have been developed over decades, and are particularly popular in the social network literature.

See the tutorial by Johan Koskinen tomorrow for a full development!

Note, however, that there are various challenges associated with ERGMs:

- Normalizing constant $\kappa(\theta)$ generally unknown.
- Model degeneracy a major concern (e.g., placing disproportionately large mass on a few extreme models)
- Model-fitting (i.e., by MCMC-based methods) requires some care.
- Usual accompanying asymptotic theory (e.g., for Gaussian confidence intervals or chi-square testing) does not extend immediately[6]

[6] See Kolaczyk, E.D. and Krivitsky, P.N. (2012). On the question of effective sample size in network modeling. arxiv-1112.0840, for a start on such things.

BOSTON UNIVERSITY

Latent Space Models

Question: What if the structure in a network is being driven by some 'simpler' form of relationship?

- Convenient to conceptualize this idea in the form of a *latent space*.
- Likelihood of two nodes being linked is then tied to their 'distance' in this latent space.
- Motivated by Hoover/Aldous' representation theorem for exchangeable (infinite) random graphs.

Key contributions in this area by Nowicki & Snijders, and Hoff and colleagues.

BOSTON UNIVERSITY

Latent Space Models (cont)

Challenge: Latent spaces indexed by variables that are . . . well . . . latent, and hence unobserved!

Need to be inferred, whether of primary or secondary interest.

We'll look at a general framework proposed by Hoff, based on latent eigen-spaces, which includes certain previously proposed models.

Latent Eigenspace Model

Consider the modeling of $\mathbf{Y} = [Y_{ij}]$, indicating presence/absence of edges in a graph G, as a *classification problem*.

A standard logistic regression classifier in this context might be based on models of the form

$$\log\left[\frac{\mathbb{P}_\beta\left(Y_{ij}=1|\mathbf{Z}_{ij}=\mathbf{z}\right)}{\mathbb{P}_\beta\left(Y_{ij}=0|\mathbf{Z}_{ij}=\mathbf{z}\right)}\right] = \beta^T\mathbf{z} \ ,$$

where

- $\mathbf{Z}_{ij}$ is a vector of explanatory variables indexed in the unordered pairs $\{i,j\}$, and
- β is a vector of regression coefficients, assumed common to all pairs.

Problem: This type of model assumes Y_{ij}'s conditionally independent, given $\mathbf{Z}_{ij}$'s. Unlikely to be true here!

BOSTON UNIVERSITY

Latent Eigenspace Model (cont.)

Solution: Introduce latent variables $\mathbf{M} = [M_{ij}]$, and assume conditional independence of Y_{ij}'s given both $\mathbf{Z}_{ij}$'s *and* M_{ij}'s.

Let

$$\mathbf{M} \;=\; \mathbf{U}^T \Lambda \mathbf{U} + \mathbf{E} \;\;,$$

be an (unknown) random, symmetric $N_v \times N_v$ matrix.

Then replace the standard classification model with

$$\log \left[\frac{\mathbb{P}_\beta \left(Y_{ij} = 1 | \mathbf{Z}_{ij} = \mathbf{z}, M_{ij} = m \right)}{\mathbb{P}_\beta \left(Y_{ij} = 0 | \mathbf{Z}_{ij} = \mathbf{z}, M_{ij} = m \right)} \right] = \beta^T \mathbf{z} + m \;\;.$$

Fitting Latent Eigenspace Models

The latent variable matrix $\mathbf{M}$ is intended to capture effects of network structural characteristics or processes not already described by the observed explanatory variables $\mathbf{Z}_{ij}$.

Additional distributional assumptions needed for β and components of $\mathbf{M}$.

Hoff proposes

- $\mathbf{U}$ uniform on the space of all $N_v \times N_v$ orthonormal matrices, and
- $\text{diag}(\Lambda)$, $\mathbf{E}$, and β multivariate Gaussian.

MCMC used to simulate for posterior-based inference.

Illustration: Predicting Protein Protein Interactions

- Networks of protein-protein interactions are of fundamental interest in computational biology.
- Measurement of such interactions is subject to error (i.e., by FP and FN).
- In addition, assessment of protein pairs is not necessarily exhaustive.
- As a result, this has become a canonical context within which to develop and assess methods for *network link prediction*.

BOSTON UNIVERSITY

Predicting PPI (cont.)

In joint work with Xiaoyu Jiang[7], we propose a latent eigen-probit model with link uncertainty (LEPLU), consisting of

- eigen-probit mixed effects network model
- inputs from STRING database as fixed effects (e.g., co-expression, database information, text mining, etc.)
- error model for FP / FN observations

[7] Jiang, X. and Kolaczyk, E.D. (2012). A latent eigen-probit model with link uncertainty for prediction of protein-protein interactions. *Statistics in Biosciences Special Issue on Networks,* 4:1, 84-104 .

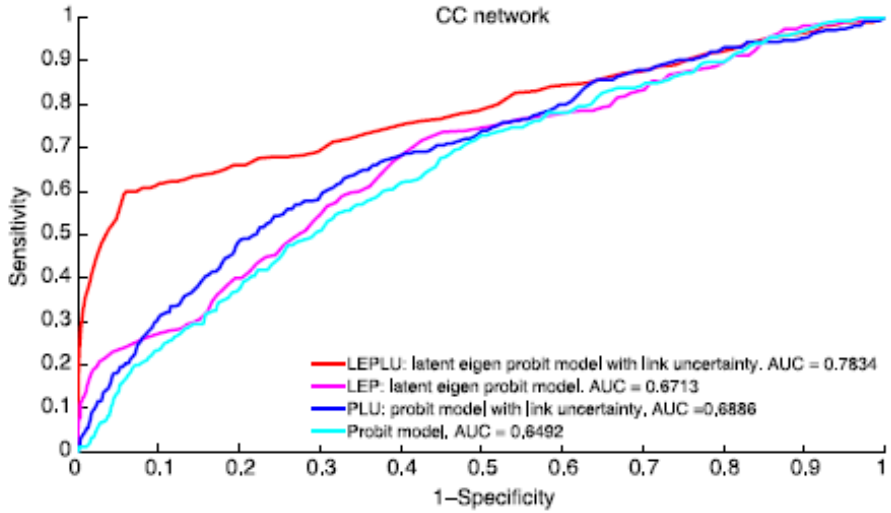
LEPLU Model for Predicting PPI

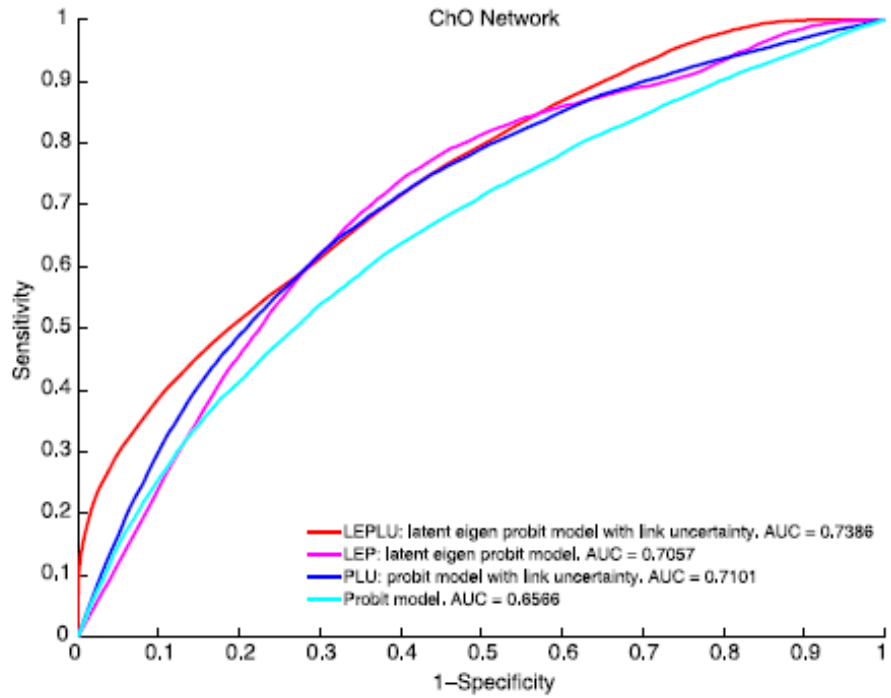
Table 1 Posterior estimate (95% credibility intervals) for parameters in LEPLU

Parameters	CC network	ChO network
$\beta_{\text{neighborhood}}$	–	0.0841
	–	(0.0074, 0.1607)
$\beta_{\text{gene fusion}}$	0.4791	0.4224
	(0.4292, 0.5291)	(0.3755, 0.4693)
$\beta_{\text{co-occurrence}}$	1.9652	3.6858
	(1.8267, 2.1037)	(3.6264, 3.7453)
$\beta_{\text{co-expression}}$	0.0842	1.4370
	(0.0362, 0.1322)	(1.3855, 1.4886)
β_{database}	0.3630	0.5833
	(0.3195, 0.4065)	(0.5372, 0.6294)
$\beta_{\text{textmining}}$	0.2875	1.1937
	(0.2405, 0.3345)	(1.1493, 1.2380)
λ_1	1.6253×10^{-6}	-5.6708×10^{-6}
	$(-5.1648 \times 10^{-6}, 8.1452 \times 10^{-6})$	$(-2.3705 \times 10^{-7}, 1.2364 \times 10^{-6})$
λ_2	−0.0275	−0.0110
	(−0.0277, −0.0274)	(−0.0111, −0.0108)
g_{FP}	0.4884	0.0040
	(0.4883, 0.4885)	(0.0028, 0.0052)
g_{FN}	0.0159	0.1109
	(0.0144, 0.0173)	(0.1096, 0.1122)

BOSTON UNIVERSITY

LEPLU Model for Predicting PPI (cont.)





BOSTON UNIVERSITY

Stochastic Block Models

Stochastic block models explicitly parameterize the notion of

- groups/modules, with
- different rates of connections between/within.

They are a hierarchical version of an ERGM, in which one level assigns group membership and the second level is a block model.

Often used in community detection, where the focus is solely on the inference of group membership, resulting in a graph partitioning algorithm.

More generally, when modeling we may be interested in the parameters, the groups, or both!

BOSTON UNIVERSITY

Stochastic Block Models: Definition

This is a generative model, where

1. Each vertex independently belongs to a group k with probability π_k, where $\sum_{k=1}^{K} \pi_k = 1$.
2. For vertices $i, j \in V$, with $i \in C_k$ and $j \in C_{k'}$, the probability that $\{i, j\} \in E$ is $P_{k,k'}$.

Note that the probability that there is no edge between i and j is

$$1 - \sum_{1 \le k,k' \le K} \pi_k \pi_{k'} P_{k,k'} \quad .$$

Stochastic Block Models: Inference

Stochastic block models are defined up to parameters

- $\{\pi_k\}_{k=1}^K$ and
- $\{P_{k,k'}\}_{1 \le k,k' \le K}$.

Suggests, conceptually, thinking of parallel sets of observations

- latent class indicators $\mathbf{I} = \{\{I_{i \in C_k}\}_{k=1}^K\}_{i \in V}$, and
- adjacency matrix $\mathbf{A} = (A_{ij})$

We observe $\mathbf{A}$ but not the $I_{i \in C_k}$.

Traditionally, therefore, inference uses the expectation-maximization (EM) algorithm; more recent work includes the use of variational methods.

Focus of much theoretical development as well.

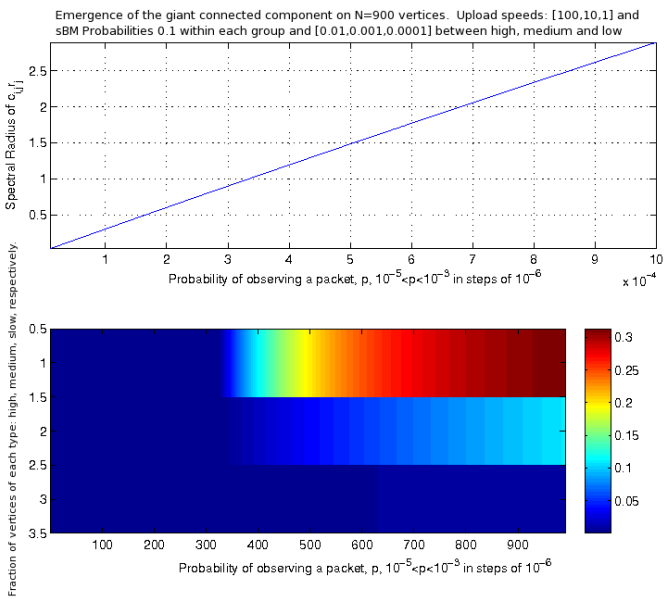
BOSTON UNIVERSITY

Illustration: Packet Sampling & P2P Networks

- In 2007, Ledoux et al. performed an experiment on 40 peers in a Bittorrent P2P Network that showed that upload traffic stratify peers into three groups: high, medium, and slow upload speeds.
- Detection of a connection between peers occurs by the observation of sending and receiving of a unit of information known as a packet.
- The ability to detect a packet therefore affects the underlying network topology and hence the inference of the groups themselves. We wish to understand these effects.
- Since the results of Ledoux et al. show that the underlying network partitions into three distinct groups, we assume that the underlying network structure without packet sampling is that of a stochastic block model [sBM].
- We then study the impact of Internet packet sampling on our ability to accurately observe the topology of this network.

BOSTON UNIVERSITY

Packet Sampling & P2P Networks: Results



Analysis and computation are used to quantify effect of 'thinning' on

- Observability of nodes
- Observed degree distribution
- Numbers/size of groups observed.

Left: We obtain an explicit solution for the fraction of vertices expected to be visible in each group, for a given packet sampling rate, as a function of when the spectral radius of a certain 3×3 symmetric matrix exceeds 1.

BOSTON UNIVERSITY

Looking Ahead . . .

Tomorrow's lecture:

- Network processes