E6885 Network Science Lecture 5:
Network Estimation and Modeling

Ching-Yung Lin, Dept. of Electrical Engineering, Columbia University

October 7th, 2013

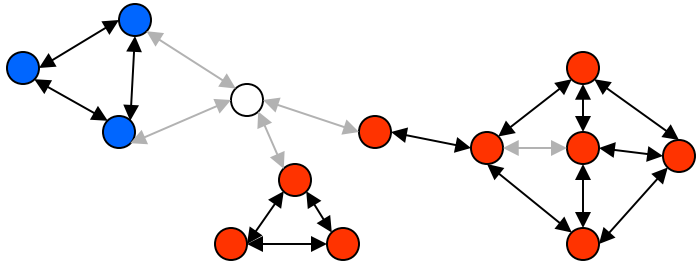
Course Structure

Class Date	Lecture	Topics Covered
09/09/13	1	Overview of Network Science
09/16/13	2	Network Representation and Feature Extraction
09/23/13	3	Network Paritioning, Clustering and Visualization
09/30/13	4	Network Analysis Use Case
10/07/13	5	Network Sampling, Estimation, and Modeling
10/14/13	6	Graph Database
10/21/13	7	Network Tolopogy Inference and Prediction
10/28/13	8	Graphical Model and Bayesian Networks
11/11/13	9	Final Project Proposal Presentation
11/18/13	10	Dynamic and Probabilistic Networks
11/25/13	11	Information Diffusion in Networks
12/02/13	12	Impact of Network Analysis
12/09/13	13	Large-Scale Network Processing System
12/16/13	14	Final Project Presentation

Network Sampling and Estimation

Why Network Sampling and Estimation is important?

- Frequently, only a portion of the nodes and edges is observed in a complex system.
- What is the outcome?
  - Are the network characteristics measurements from a subgraph representing the whole network?
  - What's the difference between principles of statistical sampling theory and sampling of graphs?



Definitions

- Population graph: $G = (V,E)$
- Sampled graph: $G^* = (V^*,E^*)$
  - In principle, G^* is a subgraph of G.
  - Error may exist in assessing the existence of vertices or edges, through observations.

- Assume we are interested in a particular characteristic of G: $\boldsymbol{\eta(G)}$
  - For instance, $\eta(G)$ is:
    - Number of edges of G
    - Average degree
    - Distribution of vertex betweenness centrality scores
    - Attributes of vertices such as the proportion of men with more female than male friends in a social network.
  - Are we able to get a good estimate of $\eta(G)$, say, $\hat{\eta}$ from G^* ?

Estimation based on sampled graph?

- How accurate if we directly use:

$$\hat{\eta} = \eta(G^*)$$

  - This implicity is used in many network study that asserts the properties of an observed network graph are indicative of those same properties for the graph of the network from which the data were sampled.

  - Statistic sampling theories: means, standard deviations, quantiles, etc., are measurement of individual's properties.

Examples

- Supposed that the characteristic of interest is the average degree of a graph G,

$$\eta(G) = (1/N_v)\sum_{i \in V} d_i$$

- Let the sample graph G^* be based on the n vertices, $V^* = \{i_1, \ldots, i_n\}$, and denote its observed degree sequence by $\{d_i^*\}_{i \in V^*}$

$$\hat{\eta} = \eta(G^*) = (1/n)\sum_{i \in V^*} d_i^*$$

- Begin with a simple random sample *without replacement*.

- Scenario 1: for each vertex $i \in V^*$, we observe all edges $\{i, j\} \in E$

- Scenario 2: for each vertex $i \in V^*$, we only observe all edges $\{i, j\} \in E$ when $i, j \in V^*$

Estimated average degree of the previous example

- 5,151 vertices and 31,201 edges. Original average degree 12.115.
- Sample ➔ n =1,500.

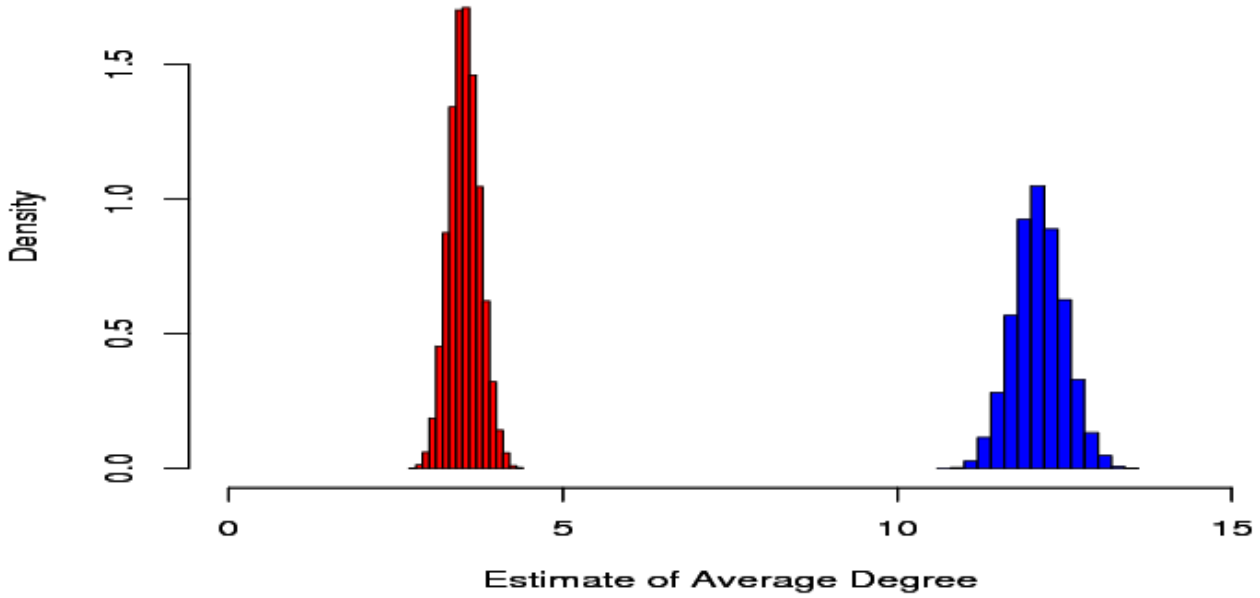


Fig. 5.1 Histograms of estimated average degree in the yeast protein interaction network, based on sampling under Design 1 (blue) and Design 2 (red), over 10,000 trials.

- Scenario 1: (mean, std) = (12.117, 0.3797). Scenario 2: (mean, std) = (3.528, 0.2260).
- A typical adjustment of Scenario 2 is: $d_i^* \approx n \cdot d_i / N_v$ ➔ mean* = 12.115

Choices of Sampling Designs

- "Statistical Properties of Sampled Networks": Physical Review, Lee, Kim and Jeong, 2006
- "Effect of sampling on topology predictions of protein-protein interaction networks, "Han et. al., Nature Biotechnology, 2005.
- In principle, this shall depend on:
  - The topology of the graph G.
  - The characteristics of $\eta(G)$
  - The nature of the sampling design

Estimation for Totals

- Suppose we have:
  - Population $\Omega = \{1, \ldots, N_u\}$,and y_i is the attribute value of interest.
  - Let $\tau = \sum_i y_i$ and $\mu = \tau / N_u$ be the total and average values of the y's in the population.
  - Let $S = \{i_1, \ldots, i_n\}$ be a sample of *n* units from Ω
- In the canonical case in which S is chosen by drawing n units uniformly from Ω, *with replacement*, a nature estimate of *μ* is:

$$\bar{y} = (1/n) \sum_{i \in S} y_i$$

and $\hat{\tau} = N_u \bar{y}$. These estimates are unbiased (i.e., $E(\bar{y}) = \mu$ and $E(\bar{\tau}) = \tau$).

Estimation for Totals (cont'd)

- The variances of these estimators take the forms:

$$V(\bar{y}) = \sigma^2 / n \qquad V(\bar{\tau}) = N_u^2 \sigma^2 / n$$

where σ^2 is the variance of the values y in the full population Ω.

- In practice:
  - Seldom simple random sample with replacement. Some units are more likely than others to be included. E.g.:
    - Marketing
    - Census
  - *Unequal Probability Sampling*

Horvitz-Thompson Estimation for Totals

- The Horvitz-Thompson estimator – through the use of weighted-averaging.

- Suppose that, under a given sampling design, each unit $i \in \Omega$ has probability π_i of being included in a sample of size n.

- Let S be the set of distinct units in the sample. Then the Horvitz-Thompson estimate of the total τ takes the form:

$$\hat{\tau}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} \qquad \text{and} \qquad \hat{\mu}_\pi = (1/N_u)\hat{\tau}_\pi$$

Let Z be a set of binary random variables, which is 1 if unit i is in S, and zero otherwise. Since,

$$E(\hat{\tau}_\pi) = E(\sum_{i \in S} \frac{y_i}{\pi_i}) = E(\sum_{i \in \Omega} \frac{y_i}{\pi_i} Z_i) = \sum_{i \in \Omega} \frac{y_i}{\pi_i} E(Z_i)$$

and $E(Z_i) = P(Z_i = 1) = \pi_i$ ➔ $\hat{\mu}_\pi$ Is an unbiased estimate of μ

Horvitz-Thompson Estimation for Totals (cont'd)

- Variance of the estimator can be expressed as:

$$V(\hat{\tau}_\pi) = \sum_{i \in \Omega} \sum_{j \in \Omega} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$$

- Its estimation is an unbiased fashion by the quantity

$$\hat{V}(\hat{\tau}_\pi) = \sum_{i \in \Omega} \sum_{j \in \Omega} y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right)$$

assuming $\pi_{ij} > 0$ for all pairs i,j.

Simple Random Sampling Without Replacement

- Consider the case of sampling without replacement.

$$\pi_i = \frac{\binom{N_u - 1}{n-1}}{\binom{N_u}{n}} = \frac{n}{N_u} \qquad \text{and} \qquad \pi_{ij} = \frac{n(n-1)}{N_u(N_u - 1)}$$

- Then the Horvitz-Thompson estimates of the total and mean have the form:

$$\hat{\tau}_\pi = N_u \bar{y} \qquad \text{and} \qquad \hat{\mu}_\pi = \bar{y}$$

- The variance may be shown to be

$$N_u(N_u - n)\sigma^2 / n$$

Compare to $V(\bar{\tau}) = N_u^2 \sigma^2 / n$ while with replacement

Probability Proportional to Size Sampling

- For instance, sampling household based on people.
- Sampling is done with replacement.
- If the probability is directly proportional to the value c_i of some characteristics.

$$\pi_i = 1 - (1 - p_i)^n \quad \text{where} \quad p_i = c_i / \sum_i c_i$$

The Horvitz-Thompson estimators are more appropriate than the sample mean.

Estimation of Group Size

- Many real cases, the size of the population is unknown.
- The capture-recapture estimators.
- The simplest version of capture-recapture involves two stages of simple random sampling without replacement, yielding two samples, say S_1 and S_2.
- Stage 1:
  - the sample S_1 of size n_1 is taken.
  - Mark all the units in S_1 .
  - All units are returned.
- Stage 2:
  - Take another sample of size n_2. Then the estimation:

$$\hat{N}_u^{(c/r)} = \frac{n_2}{m} n_1$$

where m is the number of intersection

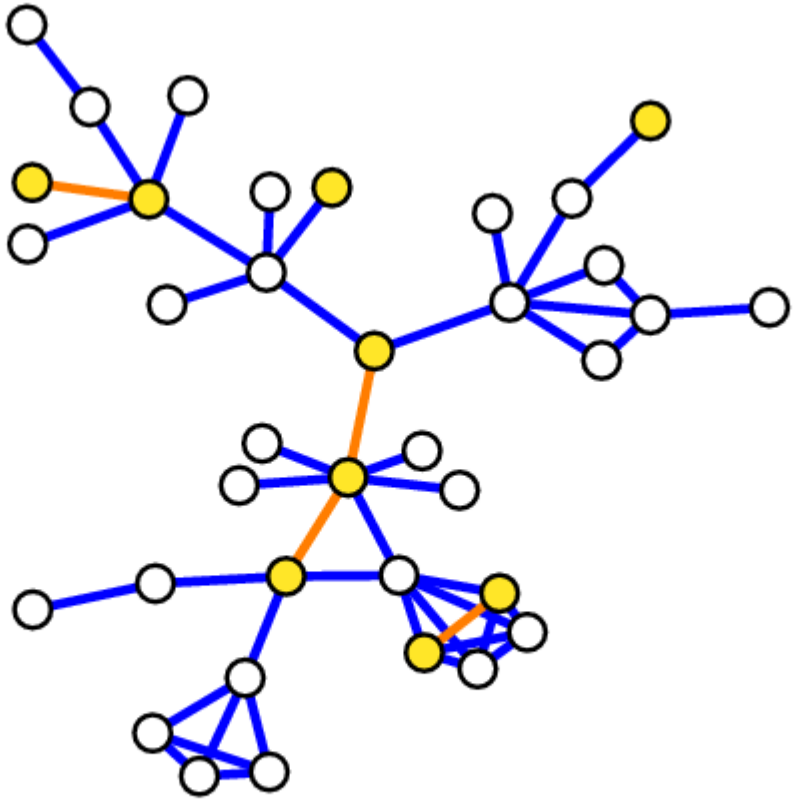
Common Network Graph Sampling Designs

- Procedures:
  - Selection Stage: two inter-related sets of united being sampled.
  - Observation Stage

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

Induced Subgraph Sampling

- Random sample of vertices in a graph and observing their induced subgraph



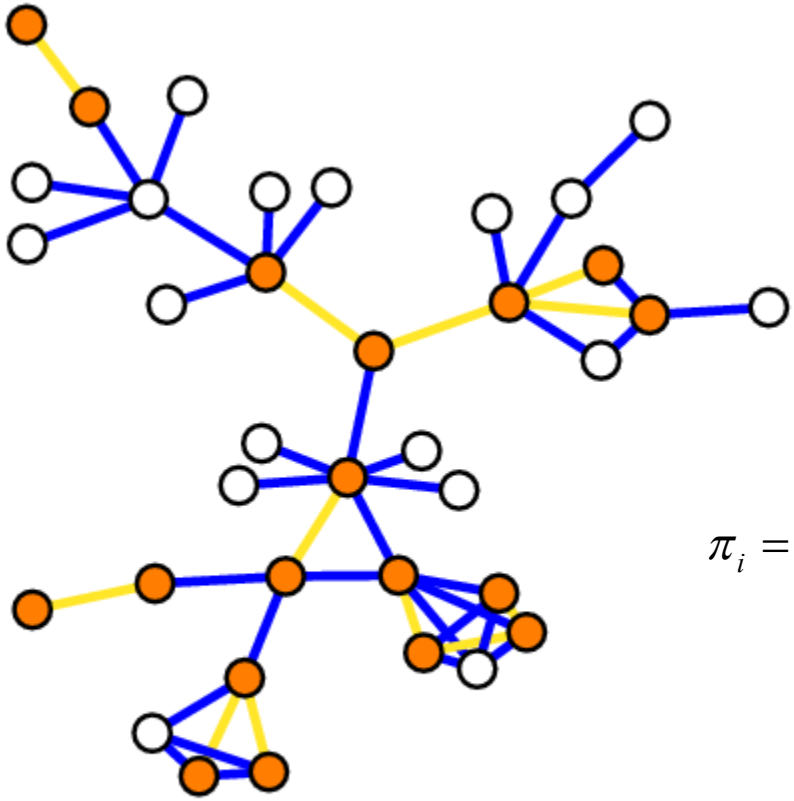
$$\pi_i = \frac{n}{N_V}$$

$$\pi_{i,j} = \frac{n(n-1)}{N_V(N_V-1)}$$

Fig. 5.2 Schematic illustration of induced subgraph sampling. Selected nodes are shown in yellow, while observed edges are shown in orange.

Incident Subgraph Sampling

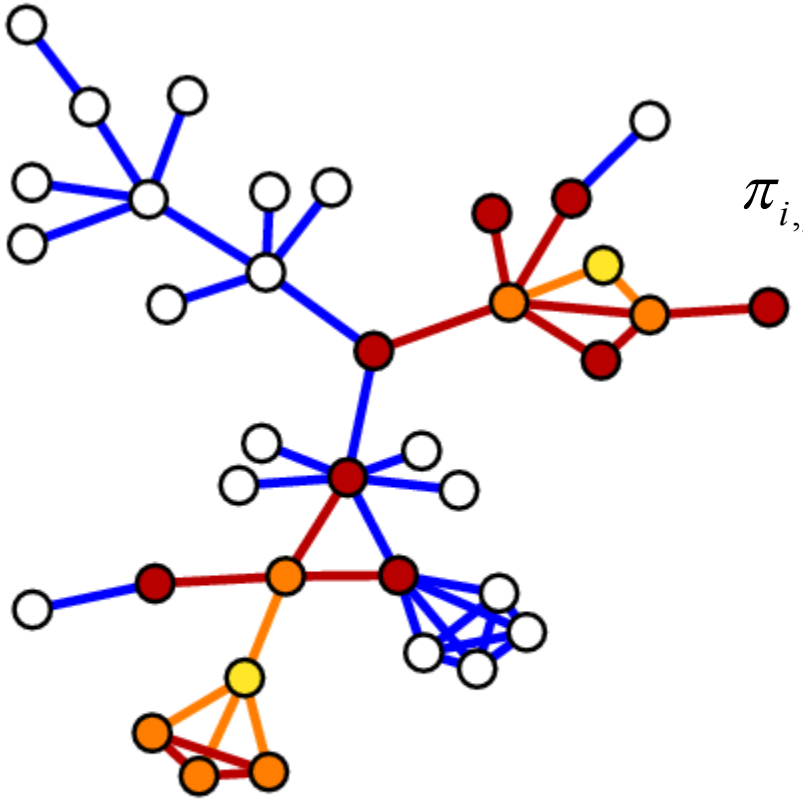
- Uniform sampling based on edges



$$\pi_i = \begin{cases} 1 - \dfrac{\binom{N_e - d_i}{n}}{\binom{N_e}{n}} & , n \le N_e - d_i \\ 1 & , n > N_e - d_i \end{cases}$$

Fig. 5.3 Schematic illustration of incident subgraph sampling. Selected edges are shown in yellow, while observed nodes are shown in orange.

Star and Snowball Sampling



$$\pi_{i,j} = \sum_{L \subseteq N_i^+} (-1)^{|L|+1} \cdot \frac{\binom{N_v - |L|}{n - |L|}}{\binom{N_v}{n}}$$

$$\pi_{i,j} = 1 - \frac{\binom{N_v - 2}{n}}{\binom{N_v}{n}}$$

Fig. 5.4 Schematic illustration of two-stage snowball sampling. Nodes selected in the initial sampling are shown in yellow, while edges and nodes observed in the first and second waves of sampling are shown in orange and brown, respectively.

Link Tracing

- After selection of an initial sample, some subset of the edges from vertices in this sample are traced to additional vertices

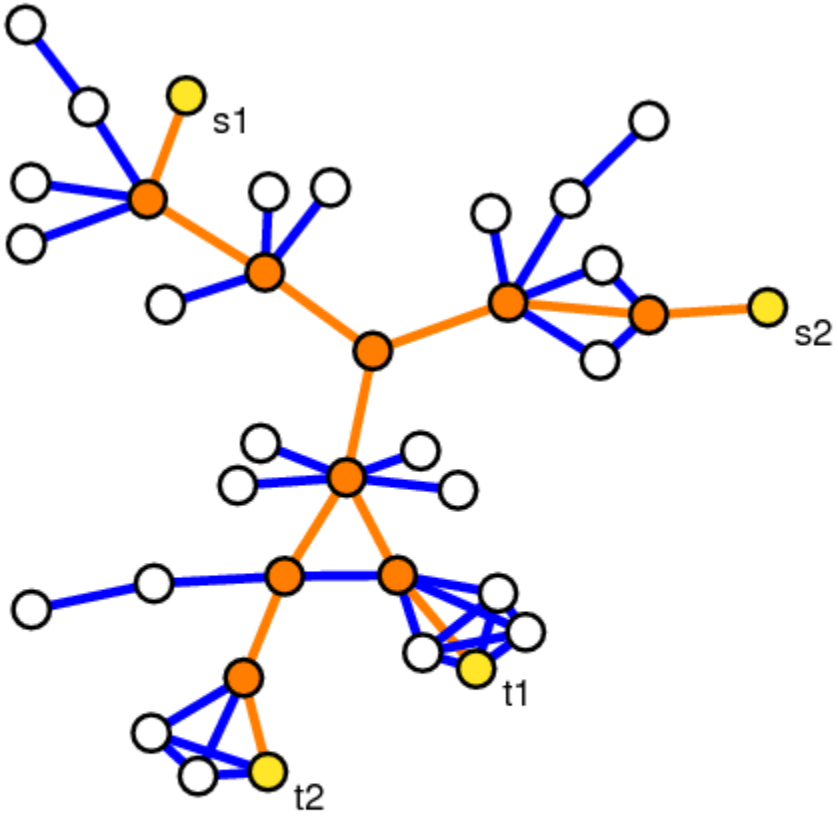
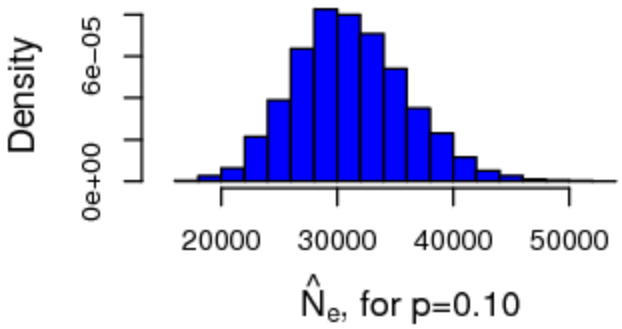
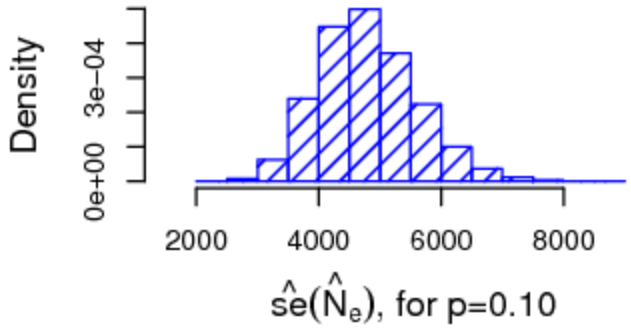
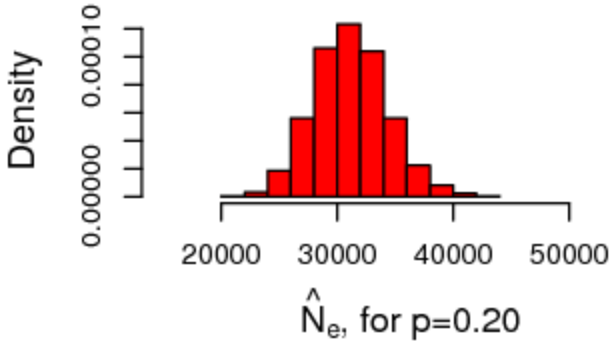


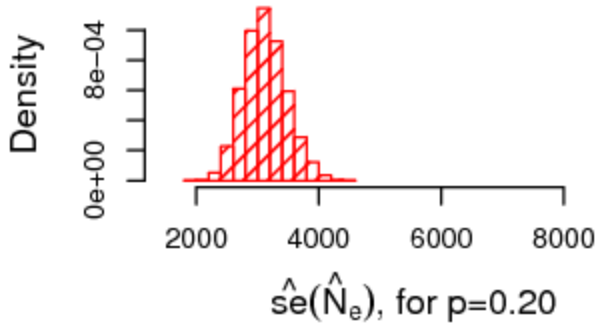
Fig. 5.5 Schematic illustration of the `traceroute` version of link-tracing. Selected source nodes $\{s_1, s_2\}$ and target nodes $\{t_1, t_2\}$ are shown in yellow, while nodes and edges observed on traces from sources to targets are shown in orange.

Estimation of the number of edges. Example with different p by induced subgraph sampling









Estimation of the number of edges. Example with different p by induced subgraph sampling (cont'd)

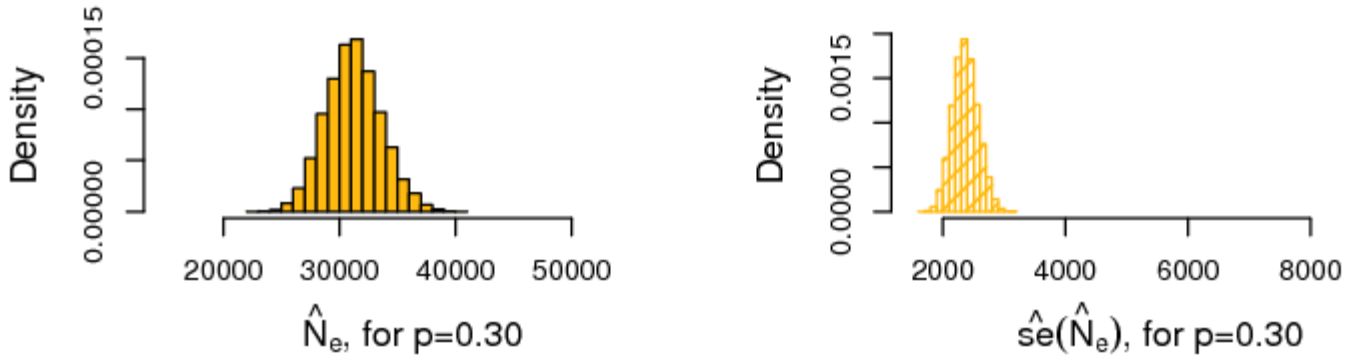


Fig. 5.6 Histograms of estimates $\hat{N}_e$ (left) of $N_e = 31,201$, as well as estimated standard errors (right), in the yeast protein interaction network, under induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.10$ (blue), 0.20 (red), and 0.30 (yellow). Results based on 10,000 trials.

Histograms of estimates of the number of triangles, connected triples and clustering coefficients.

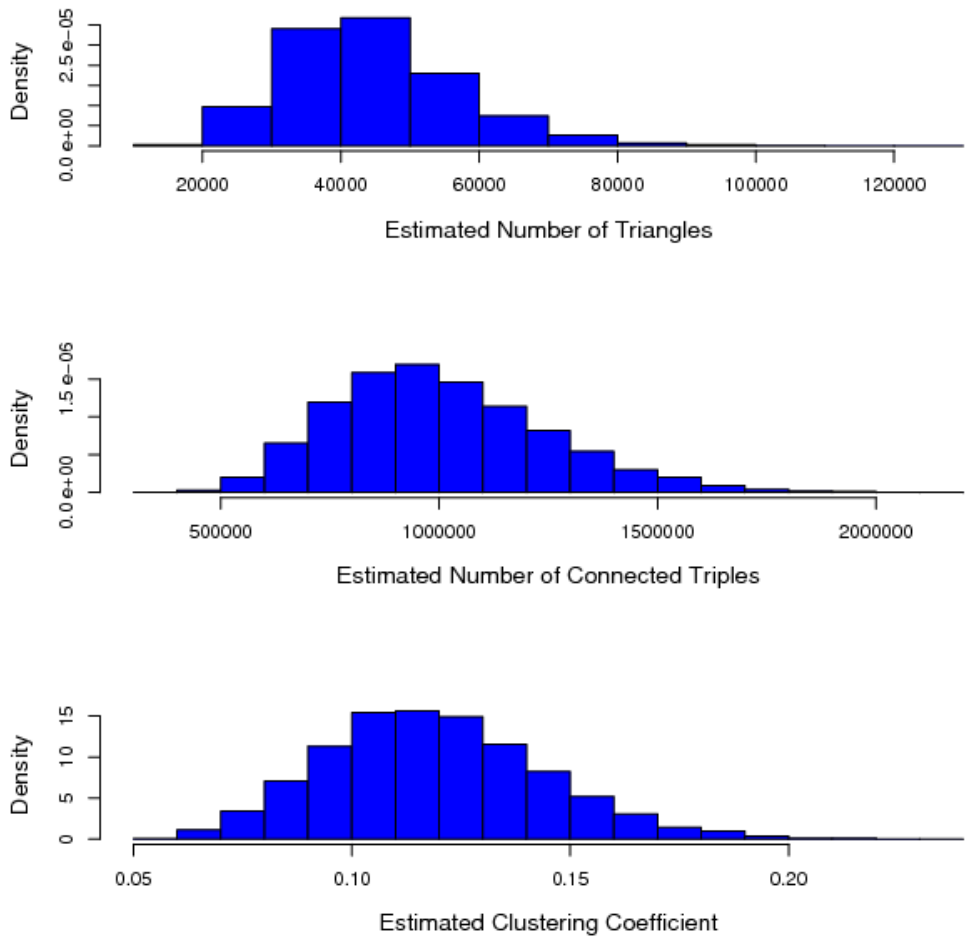


Fig. 5.7 Histograms of estimates $\hat{\tau}_{\triangle}(G)$ (top), $\hat{\tau}_3^{\dagger}(G)$ (middle), and $\widehat{\mathrm{cl}}_T(G)$ (bottom) in the yeast protein interaction network, under induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.20$. True values being estimated were $\tau_{\triangle}(G) = 44,858$, $\tau_3^{\dagger}(G) = 1,006,575$, and $\mathrm{cl}_T(G) = 0.1179$. Results based on 10,000 trials.

Random Graph Models

Network Graph Models

- Modeling of random network graphs.

$$\left\{ \mathrm{P}_\theta(G), G \in \Gamma : \theta \in \Theta \right\}$$

P_θ : probability distribution on Γ

Γ : a collection of possible graphs

Θ : a collection of parameters

Usage of network graph models

- In practice, network graph models are used for a variety of purposes.
- Study of proposed mechanisms for the emergence of certain commonly observed properties in real-world networks
- Or, the testing for significance of a pre-defined characteristics in a given network graph.

Random Graph Models

- The term 'Random Graph Model' typically is used to refer to a model specifying a collection Γ and a uniform probability $P(\cdot)$ over Γ .

- Random graph models are arguably the most well-developed class of network graph models, due to:

  – comparatively simpler nature of these models

  – This nature allows for the precise analytical characterization of many of the structural summary measures (in Chapter 4).

Model-based estimation vs. Design-based Estimation

- Model-Based Estimation vs. Design-based Estimations in Network Graphs.
  - Design-based: inference is based entirely on the random mechanism by which a subset of elements were selected from the population to create the sample.
  - Model-based: a model is given by the analyst that specifies a relationship between the sample and the population.
- In recent decades, the distinction between these two approaches has become more blurred.
- Consider the task of estimating a given characteristic $\eta(G)$ of a network graph G, based on a sampled version of that graph, G^*.
- In Chapter 5, we used 'design-based' perspective.
- If we augment this perspective to include model-based component, then G is assumed to be randomly from a collection and inference needs to consider it.

Example – Assessing Significance in Network Graphs

- Suppose we have a graph derived from observations: G^{obs}
- We are interested in accessing whether the value $\eta(G^{obs})$ is 'significant', in the sense of being somehow unusual or unexpected.

- Formally, a random graph model is used to create a reference distribution which, under the accompanying assumption of uniform likelihood of elements in Γ, takes the form:

$$P_{\eta,\Gamma(t)} = \frac{\#\{G \in \Gamma : \eta(G) \leq t\}}{|\Gamma|}$$

- If $\eta(G^{obs})$ is found to be sufficiently unlikely under this distribution, this is taken as evidence against the hypothesis that G^{obs} is a uniform draw from Γ.

- How best to choose Γ is a practical issue of some importance.

Classical Random Graph Models

- Erdos and Renyi models (1959):
  - A simple model that places equal probability on all graphs of a given order and size.

- A collection Γ_{N_v,N_e} of all graphs $G=(V,E)$ with $|V|=N_v$ and $|E|=N_e$.

- Assign probability $P(G)=\binom{N}{N_e}^{-1}$ to each $G\in\Gamma_{N_v,N_e}$, where $N=\binom{N_v}{2}$ is the total number of distinct vertex pairs.

- The key contribution of Erdos and Renyi was to develop a foundation of formal probabilistic results concerning the characteristics of graphs G drawn randomly from Γ_{N_v,N_e}

Classical Random Graph Models

- Gilber Model (1959):
- A collection $\Gamma_{N_v,p}$ of all graphs $G=(V,E)$ with $|V|=N_v$.
- Assign edge independently to each pair of distinct vertices with probability $p \in (0,1)$
- When p is an appropriately defined function of N_v and N_e : $e \cdot N_v$, these two classes of models are essentially equivalent for large N_v .

Example

- Let $p = \frac{c}{N_v}$

- Then, if $c>1$, with high probability G will have a single connected component consisting of $\alpha_c N_v$ vertices, for some constant α_c depending on c, with the remaining components having only on the order of $O(\log Nv)$ vertices.

- If $c<1$, then all components will have on the order of $O(\log Nv)$ vertices, with high probability G will consist entirely of a large number of very small, separate components.

Classic random graphs have distributions that are concentrated

- Classical random graphs have distributions that are concentrated, with exponentially decaying tails.

$$(1-\varepsilon)\frac{c^d e^{-c}}{d!} \le f_d(G) \le (1+\varepsilon)\frac{c^d e^{-c}}{d!}$$

$f_d(G)$: the proportion of vertices with degree d.

$$p = \frac{c}{N_v}$$

- For large Nv, G will have a degree distribution that is like a Poisson distribution with mean c.

Are Classical Random Graphs Practical?

- Classical random graphs do not have the broad degree distribution observed in many large-scale real-world network.
- They do not display much clustering.

- On the other hand, these graphs do possess the smal-world property. The diameter can be shown to vary like $O(\log Nv)$.

Comparing Random Graph Models with Small World Graphs

- Small World graphs start from the lattice structure, which can be shown a high level of clustering. The clustering coefficient is roughly ¾ for r large. This model begin with a set of Nv vertices, arranged in a periodic fashion, and join each vertex to r of its neighbors to each side.
- Add a few randomly rewired edges.

4

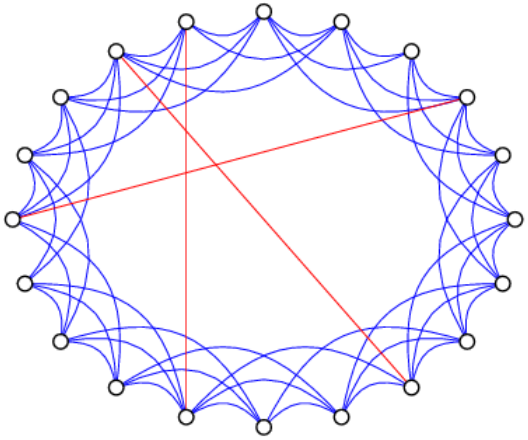


Fig. 6.4 Example of a Watts-Strogatz 'small-world' network graph. Blue edges pertain to the ori inal underlying lattice; red edges are rewired.

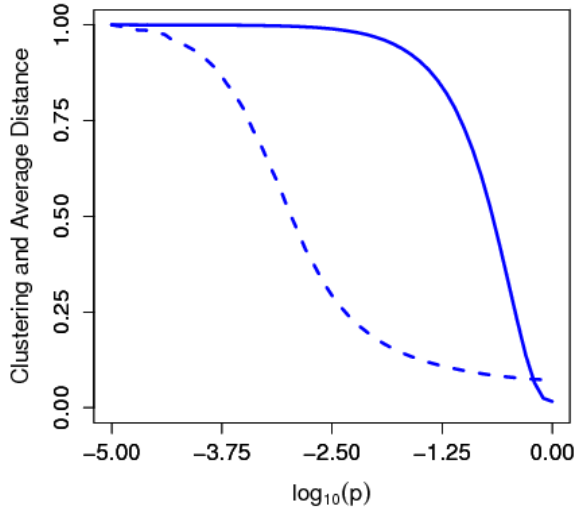


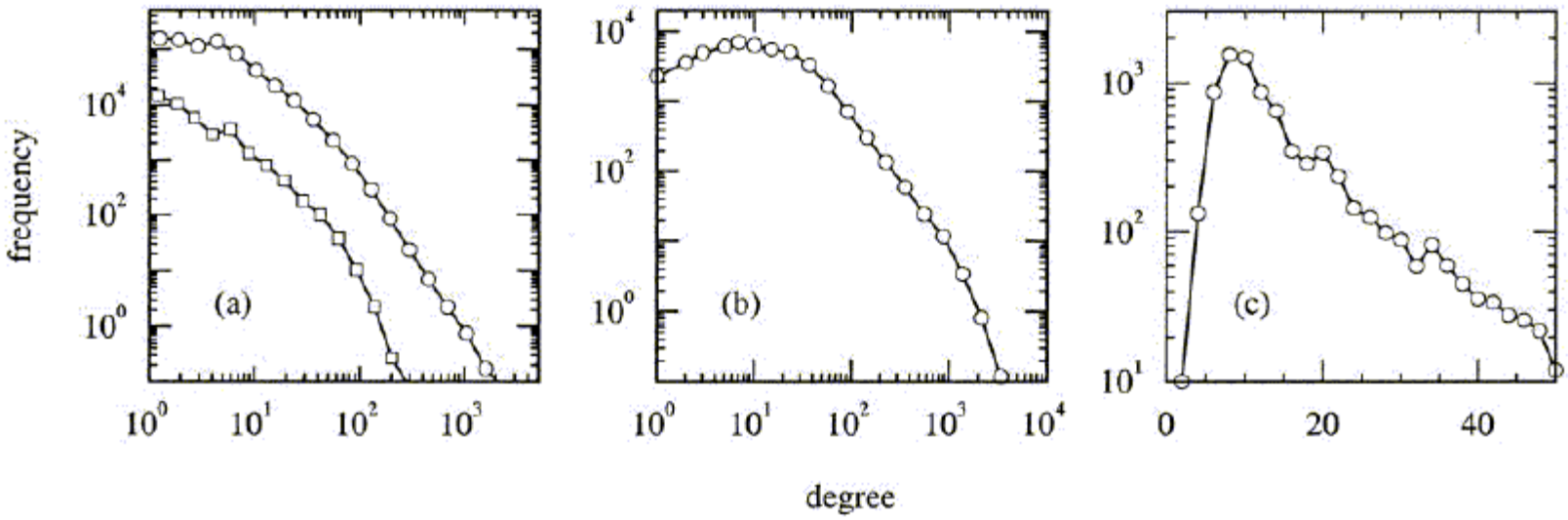
Fig. 6.5 Plot of the clustering coefficient cl(G) (solid) and average geodesic distance $\bar{l}$ (dashed), as a function of the rewiring probability p for a Watts-Strogatz small-world model. Results are averages based on 1,000 simulation trials.

Network Growth Models

- Network grows over time
- Preferential Attachment Models:
  - 'The rich get richer' principle.
  - Simon (1955) proposed a class of models that produced such broad, skewed distributions.
  - Price (1965) took this idea and applied it in creating a model for the manner in which networks of citations for document sin the literature grow.
  - Barabasi and Albert's model (1999) – a network growth model for undirected graphs.

Some examples of Degree Distribution

- (a) scientist collaboration: biologists (circle) physicists (square), (b) collaboration of move actors, (d) network of directors of Fortune 1000 companies



Power-Law Model

- Barabasi-Albert model:
  - Start with an initial graph $G^{(0)}$ of $N_v^{(0)}$ vertices and $N_e^{(0)}$ edges.
  - At Stage t=1,2,…, the current graph $G^{(t-1)}$ is modified to create a new graph $G^{(t)}$ by adding a new vertex of degree $m \geq 1$, where the m new edges are attached to m different vertices in $G^{(t-1)}$, and the probability that the new vertex will be connected to a given vertex v is given by $$\frac{d_v}{\sum_{v' \in V} d_{v'}}$$
  - At each stage, m existing vertices are connected to a new vertex in a manner preferential to those with higher degrees.
  - After t iterations, the resulting graph G will have $N_v^{(t)} = N_v^{(0)} + t$ vertices and $N_e^{(t)} = N_e^{(0)} + tm$ edges.
  - In the time as t tends to infinity, the graph G have degree distributions that tend to a power-law form $d^{-\alpha}$, with $\alpha = 3$.

Copying Models

- More common in biochemical networks, rather the WWW.
- Gene duplication is at the heart of nature's observed tendency of 're-use' biological information in evolving the genomes of living orgamisms.

- Chung et. al. (2003):
  - Beginning with an initial graph $G^{(0)}$.
  - Graphs $G^{(t)}$ are constructed from their immediate predecessors, $G^{(t-1)}$, by the addition of a new vertex, say v, that is connected to some randomly chosen subset of neighbors of a randomly chosen existing vertex, say u.
  - A vertex u is chosen from $G^{(t-1)}$ uniformly at random, and then the new vertex v is joined with each of the neighbors of u independently with probability p.
  - The degree distribution will tend to a power-law form, with exponent α satisfying the equation $p(\alpha - 1) = 1 - p^{\alpha - 1}$
  - When p=1, each new vertex is connected to $G^{(t-1)}$ by fully duplicating the edges of the randomly selected vertex u.

Fitting Network Growth Models

- Predicting – making informal comparisons between certain characteristics of an observed network and the graph resulting from such models.

- Example Wiuf duplication-attachment models – calculating a univariate likelihood function for a network (e.g., interactions among 2,368 proteins).
- However, there are a number of open issues, such as the methodology to be scaled up effectively to more complicated contexts, such as involving multivariate parameters, larger networks, more realistic network growth models, etc.

Exponential Random Graph Models

- Robins and Morris: "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."
- A potential set of such models are the "Exponential Random Graph Models" – ERGM models.

Link Inference

Network Topology Inference

- Link Inference – how to decide a link?
  - User-specified decisions and rules
  - Domain / expert knowledge

- Validation:
  - Whether the network graph representation is accurate?
  - What accuracy ideally can be expected, given the available measurable info?
  - Whether there are other similar representations that are equally or almost as accurate?
  - How robust the representation is to small changes in the measurements?
  - How useful the representation is as a basis for other purposes?

Statistical Inference on Graph

- Take our goal to select an appropriate number of graph that "best" captures the underlying state of the system.
- Based on the information in the data, as well as:
  - Any other prior information
  - Using techniques of statistical modeling and inference.

- Unfortunately, there is now no single coherent body of formal results on inference problems of this type…

➔ different forms, different context, different goals….

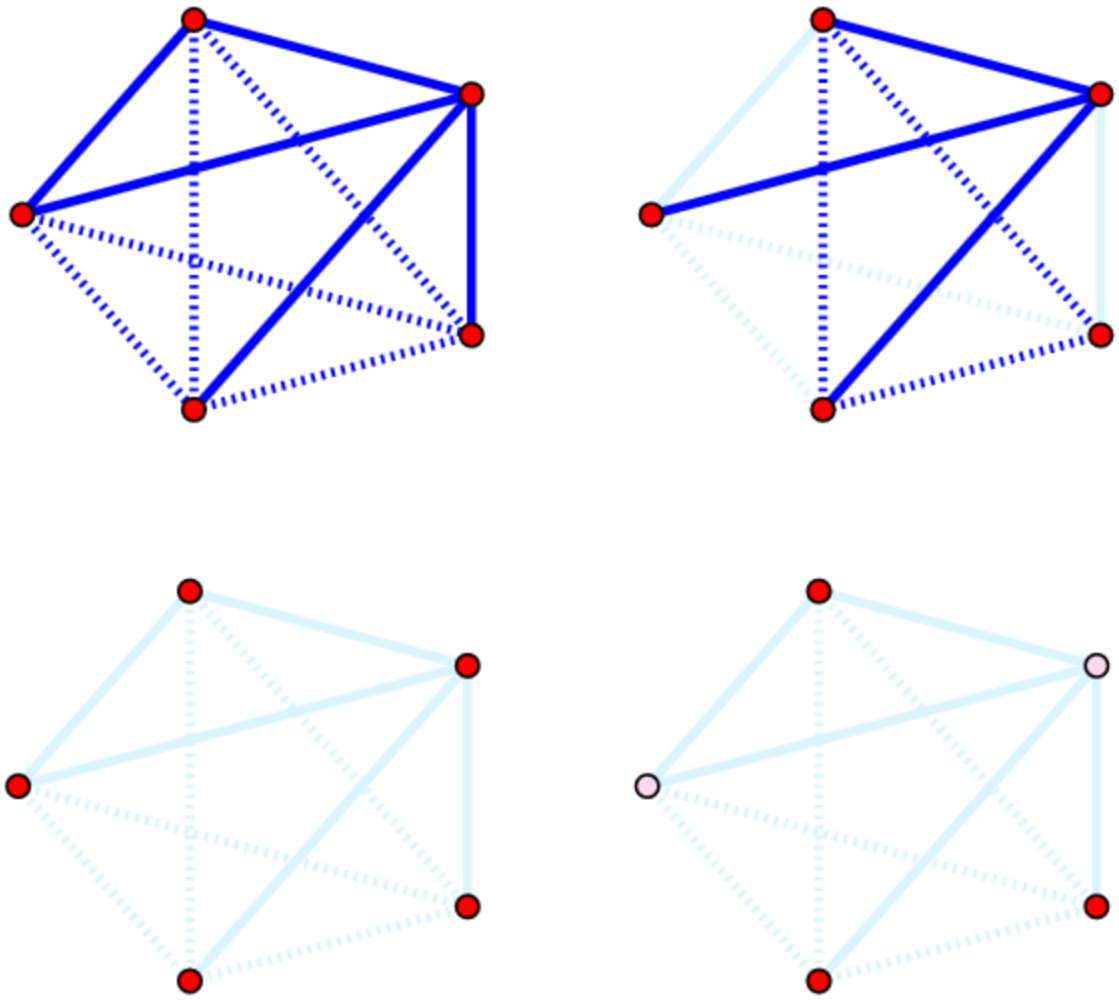
Canonical problems

- Link prediction:
  - Whether a pair of vertices does or does not have an edge between them.
  - Assume knowledge of all the vertices and the status of some of the edges/non-edges of the graph, and seeks to infer the rest of the edges.

- Inference of Association Networks:
  - Edges are defined by a “nontrivial” level of association between certain characteristics of the vertices.
  - No knowledge of edge status anywhere in the network graph.

- Tomographic Network Inference:
  - Measurements are available only at vertices that are somehow at the ‘perimeter’ of the network.
  - Infer the presence or absence of both edges and vertices in the ‘interior’.
  - Involves measurements at only a particular subset of vertices

Visual characterization of three types of network topology inference problems



Link Prediction

- Let $G = (V, E)$ be an undirected random network graph.
- $V^{(2)}$ is the set of distinct unordered pairs of vertices.
- It is the union of:
  - The set E of edges in G,
  - The set $V^{(2)} \setminus E$ of non-edges in G.
- The presence or absence of an edge is observed only for some subset of pairs, say $V^{(2)}_{obs}$. For the remaining pairs, say $V^{(2)}_{miss} = V^{(2)} \setminus V^{(2)}_{obs}$, this information is missing.

Link Prediction (cont'd)

- Let $\mathbf{Y}$ be the random $N_v \times N_v$ binary adjacency matrix.
- Denote $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{miss}$ the entries of $\mathbf{Y}$.
- The problem of link prediction is to predict the entries in $\mathbf{Y}^{miss}$, given the values $\mathbf{Y}^{obs} = \mathbf{y}^{obs}$ and possibly various vertex attribute variables $\mathbf{X} = \mathbf{x} \in R^{N_v}$
- Sometimes edges are missing, because of the difficulty in observation in a certain point in time.
- Sometimes edges are missing because of sampling.

A common assumption – missing at random

- Hoff (2007); Taskar et. al. (2004):
  - Missing information on edge presence/absence is *missing at random*.
    - The probability that an edge variable is observed depends only on the values of those other edge variables observed and not on its own value.
  - If edge variables are missing with probability depending upon themselves, this is called *informative missingness*.

- An example of information missingness in biological networks:
  - Network derived from databases (summarizing known results)
  - Bias toward 'positive' results in literatures.

The prediction model

- Given an appropriate model for $\mathbf{X}$ and $(\mathbf{Y}^{obs}, \mathbf{Y}^{miss})$, we might aim to jointly predict the elements of $\mathbf{Y}^{miss}$ based on a model for

$$P(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs} = \mathbf{y}^{obs}, \mathbf{X} = \mathbf{x})$$

➔Use the network models, e.g., as discussed in the previous 2 lectures.

➔ This approach simplify matters considerably in many ways

Informal Scoring Methods

- Scoring methods:
  - Somewhat less formal than the model-based methods
  - Popular and can be effective.
  - A score function can be incorporated into the model-based methods as explanatory variables.

- With scoring methods, for each pair of vertices i and j whose edge status is unknown, a score $s(i,j)$ is computed.

- A set of predicted edges may then be returned either by applying a threshold s^* to those scores, or by ordering them and keeping those pairs with the top n^* values.

Informal Scoring Methods (cont'd)

- A simple score, inspired by the small-world principle, is negative the shortest-path distance between i and j,

$$s(i,j) = -dist_{G^{obs}}(i,j)$$

Jaccard coefficient and other coefficients

- Scores based on comparison of the observed neighborhoods N_i^{obs} and N_j^{obs} of i and j in G^{obs} including the number of common neighbors:

$$s(i,j) = | N_i^{obs} \cap N_j^{obs} |$$

- A standardized version of this value, called the Jaccard coefficient,

$$s(i,j) = \frac{| N_i^{obs} \cap N_j^{obs} |}{| N_i^{obs} \cup N_j^{obs} |}$$

- A variation of this idea due to Liben-Nowell and Kleinberg (2003), weighting more heavily those common neighbors that are themselves not highly connected

$$s(i,j) = \sum_{k \in N_i^{obs} \cap N_j^{obs}} \frac{1}{\log | N_k^{obs} |}$$

Questions?