Network Data Sampling and Estimation

Hui Yang and Yanan Jia

September 25, 2014

Introduction

Introduction

- Population graph: network graph $G = (V, E)$
- Sampled graph: $G^{\star} = (V^{\star}, E^{\star})$
- A characristic of network graph G: $\eta(G)$
- Estimation of $\eta(G)$: $\hat{\eta}(G)$
- Example: Estimate the average degree of a network graph G by the average degree of a sampled graph $G^{\star}$, $\hat{\eta}(G) = \eta(G^{\star})$, is this a proper estimator? Depend on the sampling design!

Network Sampling Designs

Induced and Incident Subgraph Sampling

Induced and Incident Subgraph Sampling

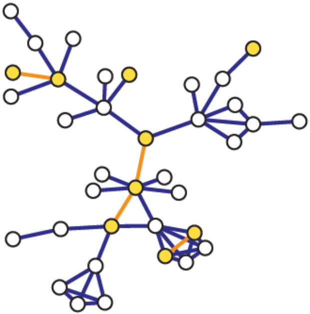


Figure : Induced Subgraph Sampling, vertices→edges

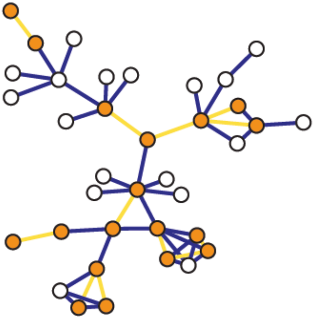


Figure : Incident Subgraph Sampling, edges→vertices

Induced and Incident Subgraph Sampling

$$\pi_i = \frac{n}{N_v} \quad \text{and} \quad \pi_{\{i,j\}} = \frac{n(n-1)}{N_v(N_v-1)}$$

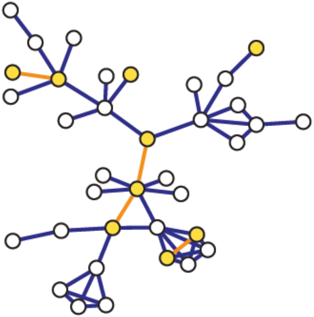


Figure : Induced Subgraph Sampling, vertices→edges

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

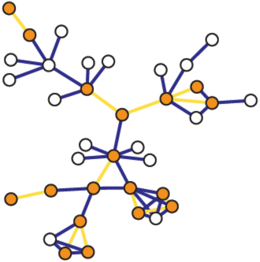


Figure : Incident Subgraph Sampling, edges→vertices

Star and Snowball Sampling

Unlabeled Star Subgraph Sampling

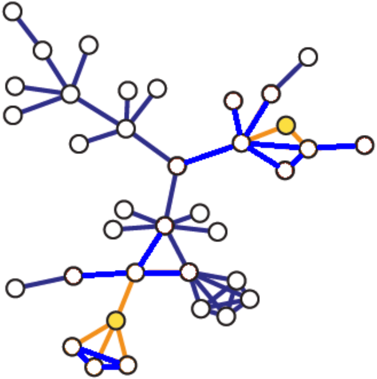


Figure : Unlabeled Star Subgraph Sampling

Labeled Star (One-stage Snowball) Subgraph Sampling

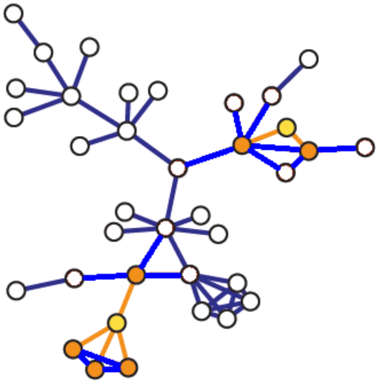


Figure : Labeled Star Subgraph Sampling

Two-stage Snowball Subgraph Sampling

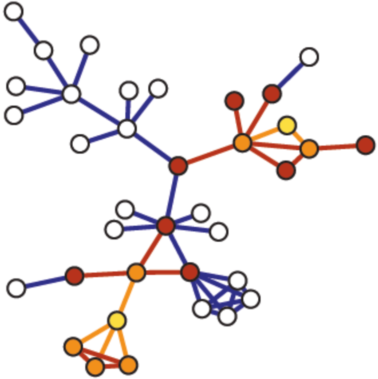


Figure : Two-stage Snowball Subgraph Sampling

Inclusion Probabilities of Star Subgraph Sampling

unlabeled star sampling

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

$$\pi_{\{i,j\}} = 1 - \mathbb{P}(\text{ neither } i \text{ nor } j \text{ are sampled })$$

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

labeled star sampling

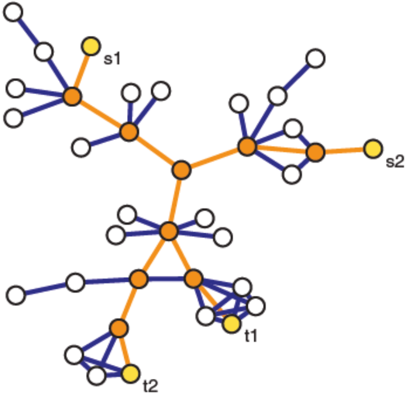
$$\pi_i = \sum_{L \subseteq \mathcal{N}_i^+} (-1)^{|L|+1} \mathbb{P}(L)$$

$$\pi_{\{i,j\}} = 1 - \mathbb{P}(\text{ neither } i \text{ nor } j \text{ are sampled })$$

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

Link Tracing Sampling

Link Tracing Subgraph Sampling



$$\pi_i \approx 1 - (1 - \rho_s - \rho_t)\exp(-\rho_s \rho_t b_i)$$
$$\pi_{\{i,j\}} \approx 1 - \exp(-\rho_s \rho_t b_{i,j})$$

Figure : Link Tracing Subgraph Sampling

Background on Statistical Sampling Theory

Horvitz-Thompson Estimation for Totals

- Poputation $\mathscr{U}$ of size N_u
- A value y_i associated with each unit $i \in \mathscr{U}$
- Sample S of size n, each unit $i \in \mathscr{U}$ has probability π_i of being included in S
- Population total: $\tau = \sum_{i \in \mathscr{U}} y_i$
- Horvitz-Thompson estimation of τ: $\hat{\tau}_\pi = \sum_{i \in S} y_i/\pi_i$
- $\hat{\tau}_\pi$ is an unbiased estimate of τ

Estimation of Group Size

- Group size N_u is typically needed to compute π_i's, in many cases N_u is unknown
- Capture-recapture estimator: first sample S_1 of size n_1 is taken, and all of the units in S_1 are marked. All of the units in S_1 are then returned to the population. Next, a sample of size n_2 is taken. Then $\hat{N}_u^{(c/r)} = n_1/(m/n_2) = n_1 n_2/m$, where m is the number of marked units observed in the second sample.

Estimation of Totals in Network Graphs

Estimation of Totals in Network Graphs

- Population $\mathscr{U} = \{1, \ldots, N_u\}$
- Unit Values y_i for $i \in \mathscr{U}$.
- Total $\tau = \sum_i y_i$ and average $\mu = \tau / N_u$.

With appropriate choice of a population of units $\mathscr{U}$ and unit values y, various graph summary characteristics $\eta(G)$ can be written in a form that involves a total $\tau = \sum_i y_i$.

Vertex Totals

- Let $\mathscr{U} = V$ and $y_i = d_i$. The average degree of a graph G is obtained by scaling the total $\sum_{i \in V} d_i$ by N_v.
- Let $\mathscr{U} = V$ and y_i be a binary variable indicating that a vertex has a given characteristic. τ counts the number of vertices with that characteristic, and τ / N_v, the proportion.

Given a sample of vertices $V^* \subseteq V$, the Horvitz-Thompson estimator for vertex totals $\tau = \sum_{i \in V} y_i$ takes the form

$$\hat{\tau}_\pi = \sum_{i \in V^*} \frac{y_i}{\pi_i}$$

where the π_i are the vertex inclusion probabilities corresponding to the underlying network sampling design.

Totals on Vertex Pairs

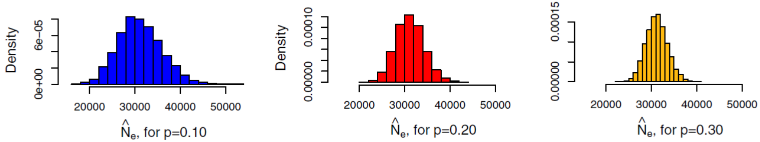
- Let $\mathscr{U} = V^{(2)}$ and $y_{(i,j)} = I_{(i,j)\in E}$ be the indicator of the event that there is an edge between i and j. The number of edges N_e is given by the total $\sum_{(i,j)\in V^{(2)}} I_{(i,j)\in E}$.
- Let $\mathscr{U} = V^{(2)}$ and $y_{(i,j)} = I_{k\in(i,j)}$ be the indicator of the event that the shortest path between i and j contains node k. In the case of unique shortest paths, the betweenness centrality $c_B(k)$ of a vertex $k \in V$ is given by the total $\sum_{(i,j)\in V^{(2)}} I_{k\in(i,j)}$.

Given a sample of vertices pair ${V^*}^{(2)} \subseteq V^{(2)}$, the Horvitz-Thompson estimator for vertex totals $\tau = \sum_{(i,j)\in V^{(2)}} y_{ij}$ takes the form

$$\hat{\tau}_\pi = \sum_{(i,j)\in {V^*}^{(2)}} \frac{y_{i,j}}{\pi_{i,j}}$$

Totals on Vertex Pairs Example

- A network of interactions among $N_v = 5,151$ proteins in S. cerevisiae with $N_e = 31,201$.
- Induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.10, 0.20, 0.30$.
- Estimator of N_e is $\hat{N}_e = \sum_{(i,j) \in V^{*(2)}} \frac{y_{i,j}}{\pi_{i,j}} = \frac{N_e^\star}{p^2}$
- Histograms of $\hat{N}_e$ based on $10,000$ trials.



Totals of Higher Order

Let $\mathscr{U} = V^{(3)}$ be the set of all triples of distinct vertices (i, j, k)

- $y_{ijk} = A_{ij}A_{jk}A_{ki}$ (A is the adjacency matrix of G), then $\tau_\Delta(G) = \sum_{(i,j,k)\in V^{(3)}} y_{(i,j,k)}$ is the number of triangles in the graph.
- $y_{ijk} = A_{ij}A_{jk}(1 - A_{ki}) + A_{ij}(i - A_{jk})A_{ki} + (1 - A_{ij})A_{jk}A_{ki}$, then $\tau_3^\star(G) = \sum_{(i,j,k)\in V^{(3)}} y_{(i,j,k)}$ is the number of vertex triples that are connected by exactly two edges.

Given a sample ${V^*}^{(3)} \subseteq V^{(3)}$, the Horvitz-Thompson estimator for $\tau = \sum_{(i,j,k)\in V^{(3)}} y_{(i,j,k)}$ takes the form

$$\hat{\tau}_\pi = \sum_{(i,j,k)\in {V^*}^{(3)}} \frac{y_{i,j,k}}{\pi_{i,j,k}}$$

Totals of Higher Order

Clustering coefficient cl_T of a graph G:

$$\text{cl}_T(G) = \frac{3\tau_\triangle(G)}{\tau_3(G)} = \frac{3\tau_\triangle(G)}{\tau_3^\star(G) + 3\tau_\triangle(G)}$$

where $\tau_3(G)$ is the number of connected triples, $\tau_\Delta(G)$ the number of triangles in the graph, and $\tau_3^\star(G) = \tau_3(G) - 3\tau_\triangle(G)$, is the number of vertex triples that are connected by exactly two edges.

The value $\text{cl}_T(G)$, called the transitivity of the graph. $\text{cl}_T(G)$ is a function of two different totals of the form

$$\hat{\tau}_\pi = \sum_{(i,j,k)\in V^{(3)}} y_{i,j,k}$$

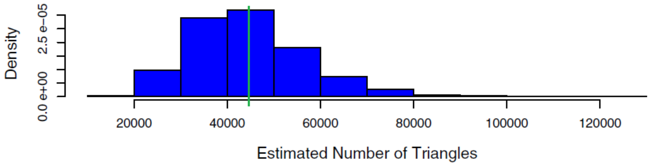
Totals of Higher Order Example

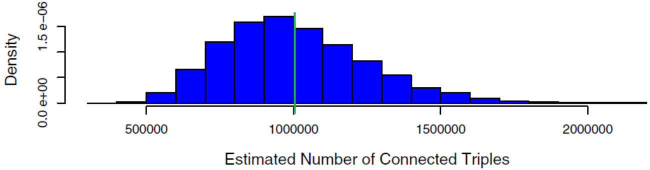
- Protein interactions Network has $\tau_{\triangle}(G) = 44,858$ triangles, $\tau_3^{\star}(G) = 1,006,575$ triples connected by exactly two edges, and a clustering coefficient $\mathrm{cl}\,T(G) = 0.1179$.
- We simulated 10,000 trials of induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.20$.
- Unbiased estimates of the two totals:

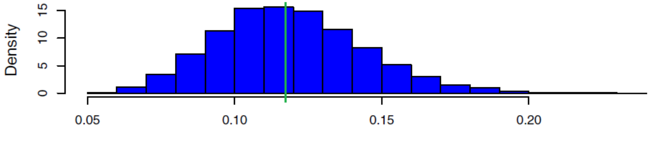
$$\tau_{\triangle}(G) = p^{-3}\tau_{\triangle}(G^*)$$

$$\tau^{\star}(G) = p^{-3}\tau^{\star}(G^*)$$

$$\mathrm{cl}_T(G) = \frac{3\tau_{\triangle}(G)}{\tau_3^{\star}(G) + 3\tau_{\triangle}(G)}$$


Density
0.0 e+00
2.5 e−05
20000
40000
60000
80000
100000
120000
Estimated Number of Triangles


Density
0.0 e+00
1.5 e−06
500000
1000000
1500000
2000000
Estimated Number of Connected Triples


Density
0
5
10
15
0.05
0.10
0.15
0.20

Estimation of Network Group Size

Estimation of Network Group Size

Simple random sampling without replacement or Bernoulli sampling

Doing the sampling twice, after 'marking' the first sample, use capture-recapture estimators,

$$\hat{N}_v = \frac{n_2}{m} n_1.$$

Estimating the Size of a 'Hidden Population'

Snowball Sampling

- $G = (V, E)$ a directed graph.
- $G^\star$ a subgraph of G, with vertices $V^\star = V_0^\star \cup V_1^\star$ obtained through a one-wave snowball sample, $V_0^\star$ selected through Bernoulli sampling with p_0.
- N the size of the initial sample, M_1 the number of arcs among individuals in $V_0^\star$, M_2 the number of arcs pointing from individuals in $V_0^\star$ to individuals in $V_1^\star$.
- Estimator of N_v will be derived using the method-of-moments.

$$\hat{N}_v = n\frac{m_1 + m_2}{m_1}$$

Other Network Graph Estimation Problems

- Estimation of Degree Frequency.
- The estimation of the number of connected components in a graph.
- The estimation of quantities not easily expressed as totals.

Sampling and estimation are also being used as a way of producing computationally efficient 'approximations' to quantities that, if computed for the full network graph, would be prohibitively expensive.

Summary

Formalize the problem of sampling and estimation in network graphs
Describe a handful of common network sampling designs
Develop estimators of a number of quantities of interest.

Thanks for your attention !