Econometrics of Sampled Networks

Arun Chandrasekhar Randall Lewis

(Presented by Khai Xiang Chiong)

March 1, 2012

Motivation

- Applied researchers usually cannot afford to obtain information on the full network, for example, the entire social network of everyone in a big city.
- Instead, they randomly sample a subset of nodes and ask the nodes to name connections and links to other nodes.
- In the previous literature, this *sampled network* is then treated as the true network.
- This sampled network is then used in studies to estimate how network structure affects economic outcomes.
- This paper examines and addresses the econometric problems that arise, i.e. biases in the estimation, when a sampled network is used instead of the true network.

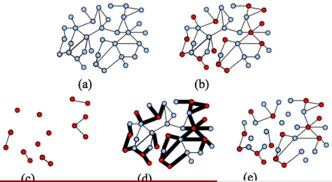
Notation and setup

- A network or a graph is a pair $G = (V, E)$ consisting of a set V of nodes and a set E of edges.
- $w(G)$, graph-level network statistics for the network G:
  - Average path length
  - Average degree
  - Maximum eigenvalue of the adjacency matrix
  - Average clustering
- $w_i(G)$, node-level network statistics for node i and network G:
  - Degree
  - Clustering
  - Eigenvector or betweenness Centrality
  - Path length

Sampling

Typically, there are two types of sampled network data

- Sample a set of m nodes and ask each node about the social connections with the other $m-1$ nodes in that data set. This is called the induced subgraph, as it restricts the network among those who are sampled.
- Sample m nodes from the network and each node can name his or her social connections to anyone in the entire network, the sampled network is called the star subgraph.
- Let ψ be the sampling rate. S be the set of surveyed nodes randomly chosen from V, with $m = |S|$. Then $m = \lfloor \psi n \rfloor$. $G^{|S} = (S, E^{|S})$ is the induced subgraph, whereas $G^S = (V, E^S)$ is the star subgraph.



Econometric Models

Regression of economic outcomes on network characteristics.

$y = \alpha + w(G)\beta_0 + \epsilon$

- Graph-level regression: the observed data is $\{(y_r, w(G_r)) : r = 1, \ldots, R\}$, where $w(G_r)$ is a vector of network statistics, and there are R observations.
- Node-level regression: the data is $\{(y_{ir}, w_{ir}(G_r)) : i = 1, \ldots n, r = 1, \ldots, R\}$, and the regression has nR observations.
- Using sampled networks, $y = \alpha + w(\bar{G})\beta_0 + \epsilon$ is run instead, where $\bar{G}$ is either $G^{|S}$ or G^S.
- Measurement error in $w(G)$ may result in attenuation bias, expansion bias, or even sign switching.

Econometric Models

Regression of economics outcomes on network characteristics.

- $y = (y_1, \ldots, y_n)'$ vector of outcome variables, $x = (x_1, \ldots, x_n)'$ vector of exogenous covariates.
- We want to estimate $y = \alpha\mathbf{1} + \rho_0 w(G)y + \gamma_0 x + \delta_0 w(G)x + \epsilon$, where the economic parameter is $\beta_0 = (\rho_0, \gamma_0, \delta_0)$.
- This captures an economic outcome y_i that depends on exogenous covariates of the individual x_i, as well as the outcome of i's peer group, as captured by $w(G)y$, where $w(G)$ is a (possibly weighted) adjacency matrix that describes how much y_i is affected by others in the network.
- Due to sampling, we mistakenly estimate the model $y = \alpha\mathbf{1} + \rho w(\bar{G})y + \gamma x + \delta w(\bar{G})x + u$

Analytical examples of bias: Average degree

In some cases, we can analytically characterize the bias, and can then correct for the bias.

- The degree of a node, $d_i(G)$ is its number of connections. The average degree of a network G is $d(G) = \frac{\sum_{i=1}^{n} d_i(G)}{n}$.
- The authors proposed the following analytical correction:
  - $\tilde{d}(G^S) = m^{-1} \sum_{i \in S} d_i(G^S)$, i.e. constructing the average degree among the randomly sampled nodes.
  - $\tilde{d}(G^{|S}) = \psi^{-1} d(G^{|S})$, where ψ is the sampling rate.
- Intuitively, the average degree is scaled down as a function of sampling rate, since only a share of social connections are observed.
- Because the regressors are scaled down, the estimated coefficient expands, while dispersion around this expectation induces attenuation.
- They show that using the above correction results in consistency, under some regularity conditions.

Analytical examples of bias: Graph clustering

- Let $\rho(G)$ denote the number of triangles in the graph G, and $\tau(G)$ denotes the number of connected triples. Then the graph clustering is $c(G) = \frac{\rho(G)}{\tau(G)}$.
- Mobius and Szeidl (2006) and Karlan et al. (2009) use a model of trust and social collateral to microfound clustering as a measure of social capital.
- The authors similarly provide the following analytical corrections:
  - $\tilde{c}(G^S) = (\frac{\psi(3-2\psi)}{1+\psi(1-\psi)})^{-1} c(G^S)$
  - $\tilde{c}(G^{|S}) = c(G^{|S})$
  - Under the induced subgraph sampling, to obtain a triangle, we must sample all three nodes. So under random sampling, the ratio $c(G^{|S})$ consistently estimates $c(G)$.

Analytical examples of bias: A model of diffusion

- There are two states: whether or not a household endorses microfinance in a weekly village gathering.
- A non-endorsing household with d_i links choose to endorse with probability $v_0 d_i \sigma_i$, where v_0 is a transmission parameter and σ_i is the fraction of i's neighbors that have decided to endorse.
- An endorsing household may naturally decide not to endorse, with probability δ_0.
- The model is identified up to $\beta_0 = \frac{v_0}{\delta_0}$.
- For a particular network G_r with degree distribution P_r, the equilibrium average endorsement rate of the network G_r is given by $\rho_r = \sum_d \frac{\beta\sigma_r(\beta)d}{1+\beta\sigma_r(\beta)d} P_r(d)$, where $\sigma_r(\beta) = (\mathbb{E}\, d)^{-1} \sum_d \frac{\beta\sigma_r(\beta)d^2}{1+\beta\sigma_r(\beta)d} P_r(d)$

Analytical examples of bias: A model of diffusion

- If we observed the average endorsement rate of R villages $\{y_1, \ldots, y_r, \ldots, y_R\}$ each with network $G_1, \ldots, G_R$.
- Assume that the relationship between y_r and $\rho_r = \sum_d \frac{\beta \sigma_r(\beta) d}{1+\beta \sigma_r(\beta) d} P_r(d)$ is given by $y_r = \rho_r + \epsilon$, where ϵ is an exogenous zero mean shock, then we can estimate β_0 via nonlinear least squares.
- Using sampled network, the parameter estimates exhibit expansion bias: plim $\hat{\beta}(G^S) > \beta_0$, and plim $\hat{\beta}(G^{|S}) > \beta_0$
- Intuitively, sampled network seems as if it has poorer diffusive properties; to generate the same average endorsement rate, the parameter governing the diffusion process must be higher.

Graphical reconstruction estimation

In general, it is difficult to provide analytical correction to many other network statistics (such as betweenness and eigenvector centrality, spectral statistics, etc) the authors proposed a graphical reconstruction method to consistently estimate economic parameter using sampled network.

Random Graphs and Asymptotic Framework

- The idea is to think of the network as a realization of a random network formation process. So G is a random variable and the network characteristic $w(G)$ is a random variable as well.
- Consider a simple but commonly used model: the probability that individuals i and j are connected, conditional on covariate z_{ij}, is given by $P(A_{ij} = 1|z_{ij}, \theta_0) = \Phi(z_{ij}'\theta_0)$
- Why? This allows us to compute the conditional expectation of the regressor $w(G)$ given the observed portion of the network A^{obs}, i.e. $\mathbb{E}[w(G)|A^{obs}; \theta_0]$.
- If $\mathbb{E}[w(G)|A^{obs}; \theta_0]$ consistently estimates $w(G)$ (say we know the true distribution of G), we can use $\mathbb{E}[w(G)|A^{obs}; \theta_0]$ in the regression, which then allows us to consistently estimate β_0.

More generally,

- If we have R networks, then we allow each network to be independently but not identically distributed, so each $\{G_r, r = 1, \ldots, R\}$ is a random draw from a distribution $P_r(G_r; \theta_{0r})$, where θ_{0r} is a parameter governing the distribution.
- In practise, the parameter θ_{0r} is unknown for each network, and we need to estimate $\hat{\theta}_r$ for each network.
- This motivates a two-stage estimation procedure.
- In the first stage, given a collection of sampled network $\{G_r^S : r = 1, \ldots, R\}$, and the variables that predictive in network formation $\{z_r : r = 1, \ldots, R\}$. $\{\hat{\theta}_r : r = 1, \ldots, R\}$ is estimated.
- In the second stage, the conditional expectation of the regressor is computed given the observed data, that is $\mathbb{E}[w_r(G_r)|G_r^S, z_r; \hat{\theta}_r]$, or $\mathbb{E}[w_r(G_r)|G_r^{|S}, z_r; \hat{\theta}_r]$.

First stage of the graphical reconstruction estimation

To illustrate the first stage of the procedure, consider a class of models in which edges are formed independently, given covariates.

- Let Ξ denote the set consisting of all pairs ij, and $s \in \Xi$ is an element of the set. z_s denote a covariate for the pair of nodes i and j. Examples include whether two villages are of the same caste, the distance between their households, etc.
- The probability that an edge forms in graph r is: $P(A_{sr} = 1 | z_{sr}; \theta_{0r}) = \Phi(z'_{sr}\theta_{0r})$
- For each graph r, the log-likelihood function is $|\Xi|^{-1} \sum_{s \in \Xi} q(A_{sr}, z_{sr}; \theta_r)$, where $q(A_{sr}, z_{sr}; \theta_r) = A_{sr} \log \Psi(z'_{sr}\theta_r) + (1 - A_{sr}) \log(1 - \Psi(z'_{sr}\theta_r)$.
- So given the observed part of the network, we can find $\hat{\theta}_r$ that maximizes the log-likelihood above.

First stage of the graphical reconstruction estimation in practice.

1. Use (z_r, A_r^{obs}) to estimate $\hat{\theta}_r$ based on the assumed network formation model.
2. Estimate $\mathcal{E}_r(A_r^{obs}, z_r; \hat{\theta}_r) = \mathbb{E}[w_r(G_r)|A_r^{obs}, z_r; \hat{\theta}_r)]$
   1. Given (z_r, A_r^{obs}), for simulations $s = 1, \ldots S$, draw $A_{r,s}^{miss}$ from $P_{\hat{\theta}_r}(A_r^{miss}|A_r^{obs}, z_r)$.
   2. Construct $w_r(G_{rs}^*)$, where $G_{rs}^* = (A_{rs}^{miss}, A_r^{obs})$.
   3. Estimate $\hat{\mathcal{E}}_r(A_r^{obs}, z_r; \hat{\theta}_r) = \frac{1}{S}\sum_{s=1}^{S} w_r(G_{rs}^*)$.

First stage of the graphical reconstruction estimation

- The authors present the asymptotic distribution of $\hat{\beta}$ under high-level assumptions on $\hat{\theta}_r$.
- We need conditions on n, R and the random graph models such that every network $\{G_r, r = 1, \ldots, R\}$ asymptotically contains enough information to estimate θ_{0r} consistently.
- In particular, they argue that not only do we need $\hat{\theta}_r$ to be consistent, but we also need $\hat{\theta}_r$ to be uniformly consistent, i.e. $\sup_r ||\hat{\theta}_r - \theta_{0r}|| = O_p(a_R^{-1} R^{1/b})$, where a_R is the rate of convergence of $\hat{\theta}_r$.
- For example, under the random graph formation model described above, the high-level assumptions on $\hat{\theta}_r$ roughly translate to the rate requirement that the number of networks R, must grow sufficiently slower than the number of nodes n.

Numerical experiments

Numerical simulations are used to characterize the biases due to sampling, as well as testing the behavior of the analytical and graph reconstruction estimators.

1. Generation of data.
   - Draw R networks from the network formation families.
   - Generate outcome data from a model with β_0 and data-generating process $(y, \epsilon)|G; \beta_0$
   - For each graph G_r, construct sampled graphs $G_r^S, G_r^{|S}$.
2. Estimation of $\hat{\beta}$ using $G_r^S, G_r^{|S}$.
   - Estimate $\hat{\beta}(G^S)$ and $\hat{\beta}(G^{|S})$ directly.
   - If applicable, estimate the analytically corrected estimator $\tilde{\beta}(G^S)$ and $\tilde{\beta}(G^{|S})$.
   - Estimate the graphical reconstruction estimators.
3. Perform (1)-(2) for $\psi \in \{1/4, 1/3, 1/2, 2/3\}$.

Numerical experiments

Overall, sampling the network leads to significant biases.

- Consider 1/3 sampling for the graph and node level.
- At the graph level, the maximum bias is 260% (λ_{max}), the mean is 90.9%, and the minimum is 15 %. (Column 2 of Table 1, page 44)
- At the node level, the maximum bias is 91%, the mean is 63%, and the minimum is 7%. (Column 2 of Table 2, page 45)
- Analytically adjusted estimators perform uniformly better. For example, at 1/3 sampling rate, when comparing to the raw network statistic, the mean reduction in bias percentage is 69 %, with a maximum of 243 %. (Column 7 of Table 1).
- Graphically reconstructed estimators nearly uniformly outperform all the raw estimators. At 1/3 sampling rate, the median bias is 5.7%, the minimum is 0.6%, and the mean reduction in bias is 73 %, and the maximum reduction is 254 %. (Column 12, Table 1)

Application to diffusion of microfinance

- The networks are randomly sampled at around 46%.
- To graphically reconstruct the network, they assume that an edge forms between a pair of households conditionally independently, given a set of covariates such as the Euclidean distance between the two households, the difference in the number of beds, number of rooms, electricity access, and roofing materials.
- The increase of the average eigenvector centrality of the initially informed households by 0.1 corresponds to a 16.3% increase in take-up rate when using the sampled data; graph reconstruction places this estimate as a 24.3% increase in take-up rate. (Column 1 of Table 7)
- Similarly, an increase of 1 on the average path length decreases take-up rate by 5.4% using sampled data, and 9.3% decrease using the graphical reconstruction estimation.
- Thus, sampling causes significant under-estimation of the network effect.