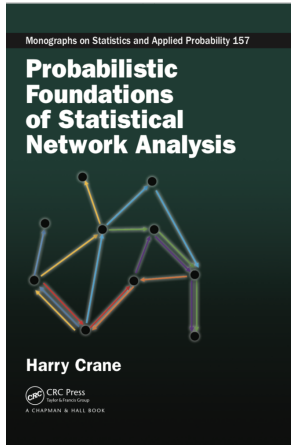
Probabilistic Foundations of Statistical Network Analysis
Chapter 3: Network sampling

Harry Crane

Based on Chapter 3 of *Probabilistic Foundations of Statistical Network Analysis*



Book website: `http://www.harrycrane.com/networks.html`

Table of Contents

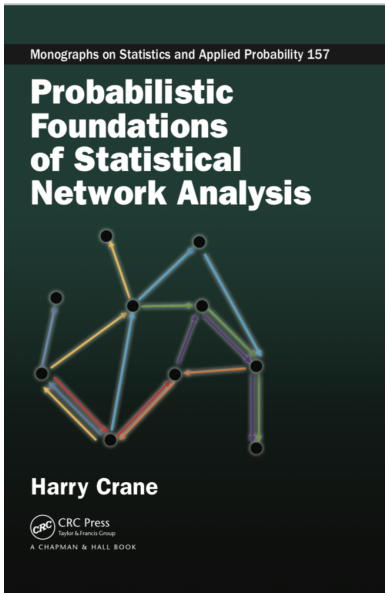


Illustration: the effects of sampling

Let $X_1, X_2, \ldots, X_N$ be i.i.d. from

$$\Pr(X_i = k+1) = \lambda^k e^{-\lambda}/k!, \quad k = 0, 1, \ldots. \tag{1}$$

What is the distribution of X' obtained by:

1. Sampling $\ell = 1, \ldots, N$ uniformly and putting $X' = X_\ell$ and
2. Choosing $\ell = 1, \ldots, N$ according to

$$\Pr(\ell = k \mid X_1, \ldots, X_N) \propto X_k, \quad k = 1, \ldots, N,$$

and putting $X' = X_k$?

Simple observation: Method of sampling affects the distribution of X'. Must be accounted for in inference. Easy for this example. Easier said than done for networks.

1. Under uniform sampling, X' distributed as in (1).
2. Under size-biased sampling, X' distributed as size-biased distribution:

$$\Pr(X' = k+1) \propto (k+1)\lambda^k e^{-\lambda}/k!, \quad k = 0, 1, \ldots.$$

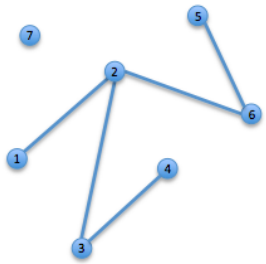
Parameters are not just Greek letters!

Network modeling

Conventional Definition:

A (parameterized) statistical model is a family of probability distributions $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$*, each defined on the sample space.*

- Population or Sample model? And what's the connection?

	Population	Observed network (sample)
		???
Model	$\{P_\theta : \theta \in \Theta\}$	???

- **Guiding Question**: How to draw sound inferences about population model based on sampled network?
- Need to model data in a manner consistent with
  - (i) population model and
  - (ii) sampling mechanism.

Selection sampling

"Selection of $[m]$ from $[n]$":

For example, for $\mathbf{A} = (A_{ij})_{1 \le i,j \le n}$ given by

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2m} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} & \cdots & A_{mn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} & \cdots & A_{nn} \end{pmatrix},$$

the restriction $\mathbf{A}|_{[m]}$, for $m \le n$, is the upper $m \times m$ submatrix given by

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix}.$$

Consistency under selection

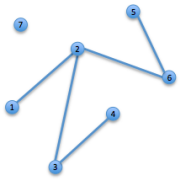
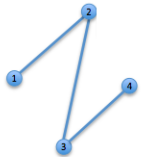
Let $\mathbf{Y}_N$ and $\mathbf{Y}_n$, $n < N$, be random arrays and write $\mathbf{S}_{n,N} : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$ to denote the act of *selecting* $[n]$ from $[N]$.

Definition

The distributions of $\mathbf{Y}_N$ *and* $\mathbf{Y}_n$ *are* consistent under selection *if*

$$\mathbf{Y}_n =_{\mathcal{D}} \mathbf{S}_{n,N}(\mathbf{Y}_N).$$

- **Example**: p_1 model (Why? See Equation (3.10) and Exercise 3.1.)
- ERGMs consistent under selection only if sufficient statistics have 'separable increments' (Shalizi and Rinaldo, 2013).

	Population	Observed network (sample)
	$\mathbf{Y}_N$	$\mathbf{S}_{n,N}(\mathbf{Y}_N)$
		
Distribution	$\mathbf{Y}_N$	$\mathbf{Y}_n$

Significance of sampling consistency

Example:

Suppose $\mathbf{Y}_N$ *follows* p_1 *model with parameters* $(\rho, \theta, \alpha, \beta)$*, for* $\alpha = (\alpha_1, \ldots, \alpha_N)$ *and* $\beta = (\beta_1, \ldots, \beta_N)$.

- Want to estimate reciprocity ρ based on observation $\mathbf{Y}_n = \mathbf{S}_{n,N}\,\mathbf{Y}_N$ for $n < N$.
- By consistency under selection, $\mathbf{Y}_n$ distributed from p_1 model with parameter $(\rho, \theta, \alpha_{[n]}, \beta_{[n]})$ for $\alpha_{[n]} = (\alpha_1, \ldots, \alpha_n)$ and $\beta_{[n]} = (\beta_1, \ldots, \beta_n)$.

  $\Longrightarrow$ If $\mathbf{Y}_N$ from p_1 model and $\mathbf{Y}_n$ obtained from $\mathbf{Y}_N$ by selection sampling, then $\mathbf{Y}_n$ also from p_1 model with same parameters.

  $\Longrightarrow$ ρ, α_i, β_i are the 'same' for $\mathbf{Y}_N$ and $\mathbf{Y}_n$.

  $\Longrightarrow$ estimate $\hat{\rho}_n$ based on $\mathbf{Y}_n$ and use same estimate for $\mathbf{Y}_N$.
- Same logic does not apply to estimating ERGM unless separable increments holds. (See Chapter 2 and Shalizi–Rinaldo (2014).)

Toward a coherent theory for network modeling

- I do **not** suggest that consistency under selection is be-all and end-all.
- It is a useful illustration of the importance of consistency with respect to subsampling.
- But **selection** is just one special kind of subsampling.
- And selection is very unrealistic in almost all networks applications of interest.

Three essential observations:

(i) sampling is an indispensable part of network modeling,

(ii) relationship between observed and unobserved data established by sampling mechanism is critical for statistical inference, and

(iii) nature of this relationship and reason why it is important have not been properly emphasized in the developments of network analysis to date.

Selection from sparse networks

- Suppose $\mathbf{Y}_N = (Y_{ij})_{1 \le i,j \le N}$ is "sparse" (aside: "sparse" a misnomer):

$$\sum_{1 \le i,j \le N} Y_{ij} \approx \varepsilon N \quad \text{for "small" } \varepsilon > 0.$$

- Sample $n \ll N$ vertices uniformly at random and observe the subgraph $\mathbf{Y}_n^*$ induced by $\mathbf{Y}_N$.
- What does $\mathbf{Y}_n^*$ look like?

- Since vertices sampled uniformly, $\mathbf{Y}_n^*$ is exchangeable and

$$\Pr(Y_{12}^* = 1) \approx \varepsilon N / ((N(N-1)) \approx \varepsilon / N \approx 0.$$

- Furthermore, we compute

$$\Pr\left(\bigcup_{1 \le i \ne j \le n} \{Y_{ij}^* = 1\} \right) \le \sum_{1 \le i \ne j \le n} \Pr(Y_{ij}^* = 1) \approx n^2 \varepsilon / N \approx 0.$$

What are the practical implications of this?

Scenario: Ego networks in high school friendships

- Suppose $\mathbf{Y}_N$ modeled by Erdős–Rényi–Gilbert distribution with parameter $\theta \in [0,1]$:

$$\Pr(\mathbf{Y}_N = \mathbf{y}; \theta) = \prod_{1 \le i \neq j \le N} \theta^{y_{ij}} (1-\theta)^{1-y_{ij}}, \quad \mathbf{y} \in \{0,1\}^{N \times N}.$$

- Observe $\mathbf{Y}^*$ by sampling v^* uniformly from $[N]$ and observing $\mathbf{Y}^* = \mathbf{Y}_N|_S$, for

$$S = \{v^*\} \cup \{v : Y_{v^*v} = 1 \text{ or } Y_{vv^*} = 1\}.$$

- What is the distribution of $\mathbf{Y}^*$?

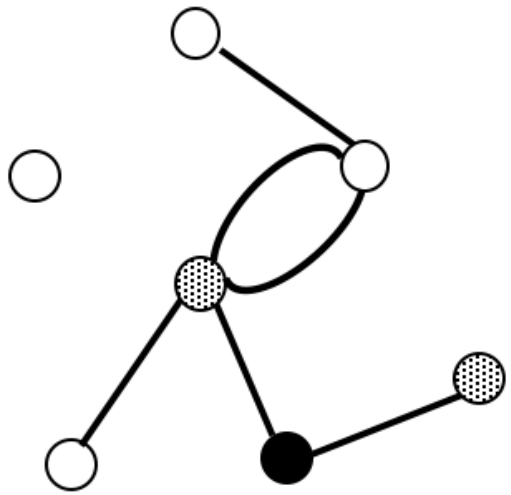


Figure: Depiction of one-step snowball sampling operation in Section 2.4. The solid filled vertex (bottom right) corresponds to the randomly chosen vertex v^* and those partially filled with dots are its one-step neighborhood.

Network sampling schemes

- Vertex sampling: As in Section 2.4 (students in a high school).
- Relational sampling
  - edge sampling: phone calls
  - hyperedge sampling: movie collaborations, co-authorships
  - path sampling: traceroute
- Snowball sampling: As in Section 3.5.

- Sampling scheme affects the units of observation.
- Units of observation affect inference/modeling.

Edge sampling (phone call database)

Table: Database of phone calls. Each row contains information about a single phone call: caller and receiver (identified by phone number), time of call, topic discussed, etc.

Caller	**Receiver**	**Time of Call**	**Topic Discussed**	...
555-7892 (*a*)	555-1243 (*b*)	15:34	Business	...
550-9999 (*c*)	555-7892 (*a*)	15:38	Birthday	...
555-1200 (*d*)	445-1234 (*e*)	16:01	School	...
555-7892 (*a*)	550-9999 (*c*)	15:38	Sports	...
555-1243 (*b*)	555-1200 (*d*)	16:17	Business	...
⋮	⋮	⋮	⋮	⋱

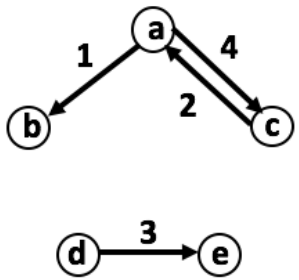


Figure: Network depiction of phone call sequence of caller-receiver pairs (a, b), (c, a), (d, e), (a, c) as in the first four rows of Table 1. Edges are labeled in correspondence with the order in which the corresponding calls were observed.

Traceroute sampling (Path sampling)

- Sample paths in the Internet by sending signals between different IP addresses and tracing the path (traceroute sampling).

```
traceroute to galton.uchicago.edu (128.135.10.17), 64 hops max, 72
byte packets
 1  fios_quantum_gateway (192.653.22.69)  2.557 ms  3.073 ms  3.881
ms
 2  lo0-100.nyp-sec-4513 -319.comcast-xxg .net (158.19.2130)  5.677
ms  15.916 ms  15.397 ms
 3  b3319.  nyp-sec -21.comcast-xxg .net (100.41.209.120)  16.386
ms  10.418 ms  16.390 ms
 4  * * *
 5  0.ae3.br2.nyc4.alter.net (140.222.231.133)  13.368 ms  9.816
ms  13.792 ms
 6  204.255.168.110 (204.255.168.110)  15.426 ms  28.583 ms
10.595 ms
 7  be2061.ccr42.jfk02.atlas.cogentco.com (154.54.3.69)  13.331 ms
15.715 ms  15.677 ms
 8  be2890.ccr22.cle04.atlas.cogentco.com (154.54.82.245)  34.393
ms  30.023 ms  26.264 ms
 9  be2718.ccr42.ord01.atlas.cogentco.com (154.54.7.129)  32.745
ms  29.979 ms  28.970 ms
10  be2522.agr21.ord01.atlas.cogentco.com (154.54.81.62)  38.000
ms  32.219 ms  36.152 ms
11  te0-0-2-0.nr11.b010917-1.ord01.atlas.cogentco.com
(154.24.4.38)  48.702 ms  33.046 ms  29.155 ms
12  38.104.103.10 (38.104.103.10)  28.439 ms  35.973 ms  31.425 ms
13  192.170.192.19 (192.170.192.19)  34.475 ms  31.105 ms  28.312
ms
14  192.170.192.27 (192.170.192.27)  29.062 ms  71.833 ms  28.353
ms
15  * * *
16  galton.uchicago.edu (128.135.10.17)  31.506 ms  30.992 ms
35.041 ms
```

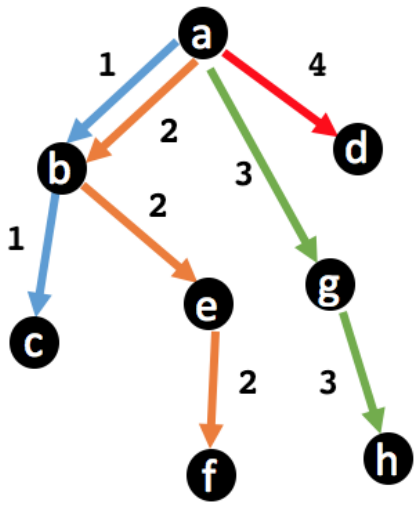


Figure: Path-labeled network constructed from sequence path$(a, c) = (a, b, c)$, path$(a, f) = (a, b, e, f)$, path$(a, h) = (a, g, h)$, and path$(a, d) = (a, d)$. Edges are labeled according to which path they belong. For example, the three edges labeled '2' should be regarded as comprising a single path, namely path$(a, f) = (a, b, e, f)$, and not as three distinct edges $(a, b), (b, e), (e, f)$.

Hyperedge sampling

Actor collaborations:

Movie title	**Starring cast**
Rocky	Sylvester Stallone, Bert Young, Carl Weathers, . . .
Rounders	Matt Damon, Ed Norton, John Malkovich, John Turturro, . . .
Groundhog Day	Bill Murray, Andie McDowell, Chris Elliott, . . .
A Bronx Tale	Robert DeNiro, Chazz Palminteri, Joe Pesci, . . .
Over the Top	Sylvester Stallone, Robert Loggia, . . .
The Room	Tommy Wiseau, Greg Sestero, . . .
⋮	⋮

Scientific coauthorships:

Article title	**Authors**
A nonparametric view of network models . . .	Bickel, Chen
Edge exchangeable models for interaction networks	Crane, Dempsey
Snowball sampling	Goodman
Latent space approaches to social network analysis	Hoff, Raftery, Handcock
⋮	⋮

Units of observation

Statistical units:

- In experimental design literature: the smallest entities to which different treatments can be assigned.
- In network analysis: the basic entities of observation, i.e., the 'atomic elements' from which the network structure is constructed.

Examples:

- Social network by sampling high school students (vertices), vertices are units.
- Network obtained by sampling calls (edges) from database, edges are units.
- Network obtained by sampling emails/articles/movies, hyperedges are units.
- Network obtained by traceroute sampling in the Internet, paths are units.

(Handcock and Gile (2010))

"In most network samples, the unit of sampling is the actor or node. (Handcock and Gile, p. 7)

- Misguided quotation: unit of sampling is rarely the actor/node/vertex in most modern applications. Think about interaction networks (sampling edges, hyperedges, paths, etc.).
- Important to distinguish 'implicit' from 'explicit' units. See Section 3.8.

What is the sample size?

- Age-old question of network science. Still poorly understood:

(Common trope)

An observation of network data is a 'sample of size 1'.

- Misguided: Sample size is not the number of networks. It is the number of **units** observed to construct that network.
- Analogy: What is the sample size of i.i.d. sequence $(X_1, \ldots, X_n)$?
  - Apply same logic above: observe 1 sequence $\rightarrow$ sample size 1.
  - Or: observe n observations from common distribution $\rightarrow$ sample size n.
- Second answer makes more sense for sequences, and also for networks.

The sample size is the number of observed units.

Consistency under subsampling

Inadequacy of selection sampling (and therefore consistency under selection) calls for more general theory for network sampling.

- Selection sampling: $\mathbf{S}_{n,N} : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$ is just *restriction*.
- Define ψ-sampling: for any injection (1-to-1 function) $\psi : [n] \to [N]$ define

$$\mathbf{S}^{\psi}_{m,n} : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$$

$$\mathbf{y} \mapsto \mathbf{S}^{\psi}_{m,n}\,\mathbf{y} = (y_{\psi(i)\psi(j)})_{1\le i,j\le n}.$$

- Let $\Sigma_{n,N}$ be a random sampling scheme chosen from among all ψ-sampling maps $\mathbf{S}^{\psi}_{m,n}$.
- Note: Distribution of $\Sigma_{n,N}$ can depend on the network $\mathbf{Y}_N$ being sampled from. (Degree-biased sampling, snowball, edge sampling, path sampling, etc.)

Definition (Consistency under subsampling)

Call $\mathbf{Y}_N$ *and* $\mathbf{Y}_n$ consistency under sampling from $\Sigma_{n,N}$*, or simply* $\Sigma_{n,N}$-consistent*, if*

$$\Sigma_{n,N}\,\mathbf{Y}_N =_{\mathcal{D}} \mathbf{Y}_n,$$

where the distribution of $\Sigma_{n,N}\,\mathbf{Y}_N$ *is calculated by*

$$Pr(\Sigma_{n,N}\,\mathbf{Y}_N = \mathbf{y}) = \sum_{\psi:[n]\to[N]} \mathbf{1}(\mathbf{S}^{\psi}_{m,n}\,\mathbf{y}^* = \mathbf{y})\, Pr(\Sigma_{n,N} = \mathbf{S}^{\psi}_{m,n} \mid \mathbf{Y}_N = \mathbf{y}^*)\, Pr(\mathbf{Y}_N = \mathbf{y}^*).$$

Consistency under subsampling: Goals

Definition (Consistency under subsampling)

Call $\mathbf{Y}_N$ *and* $\mathbf{Y}_n$ consistency under sampling from $\Sigma_{n,N}$*, or simply* $\Sigma_{n,N}$-consistent*, if*

$$\Sigma_{n,N}\,\mathbf{Y}_N =_{\mathcal{D}} \mathbf{Y}_n,$$

where the distribution of $\Sigma_{n,N}\,\mathbf{Y}_N$ *is calculated by*

$$Pr(\Sigma_{n,N}\,\mathbf{Y}_N = \mathbf{y}) = \sum_{\psi:[n]\to[N]} \mathbf{1}(\mathbf{S}^{\psi}_{m,n}\,\mathbf{y}^* = \mathbf{y})\, Pr(\Sigma_{n,N} = \mathbf{S}^{\psi}_{m,n} \mid \mathbf{Y}_N = \mathbf{y}^*)\, Pr(\mathbf{Y}_N = \mathbf{y}^*).$$

- Short-term goal: build a framework within which to incorporate sampling into network analysis. Proper definition of consistency under subsampling is a start.
- Long-term goal: develop theory of sampling for network analysis.

Coming up:

- **Chapter 4**: Generative models
- **Chapter 5**: Statistical modeling paradigm