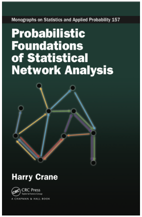
Probabilistic Foundations of Statistical Network Analysis
Chapter 4: Generative models

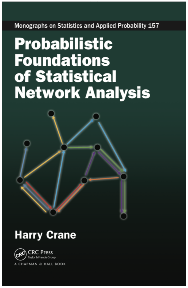
Harry Crane

Based on Chapter 4 of *Probabilistic Foundations of Statistical Network Analysis*



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

Table of Contents



Specification of generative models

- Sampling models (Chapter 3) specified by
  - candidate distributions describing network variation
  - sampling scheme that links the population $\mathbf{Y}_N$ to the sample $\mathbf{Y}_n = \Sigma_{n,N}\mathbf{Y}_N$
- Generative models (Chapter 4) specified by
  - candidate distributions
  - generative scheme to describe network growth
- Describe generative scheme by an *evolution map*.

Definition

For $n \leq N$, call $P : \{0,1\}^{n\times n} \to \{0,1\}^{N\times N}$ an **evolution map** *if*

$$P(\mathbf{y})|_{[n]} = \mathbf{y} \quad \text{for all } \mathbf{y} \in \{0,1\}^{n\times n}.$$

An evolution map is an operation by which $\mathbf{y} \in \{0,1\}^{n\times n}$ 'evolves' into $P(\mathbf{y}) \in \{0,1\}^{N\times N}$ by holding fixed the part of the network that already exists, namely $\mathbf{y}$.

- Let $\mathcal{P}_{n,N}$ be the set of all evolution maps $\{0,1\}^{n\times n} \to \{0,1\}^{N\times N}$.
- A **generating scheme** is a random map $\Pi_{n,N}$ in $\mathcal{P}_{n,N}$. Distribution can depend on $\mathbf{Y}_n$.
- More precisely, $\Pi_{n,N}\mathbf{Y}_n$ is the network with N vertices obtained by first generating $\mathbf{Y}_n$ and, given $\mathbf{Y}_n = \mathbf{y}$, putting $\Pi_{n,N}\mathbf{Y}_n = P(\mathbf{y})$, for $P \in \mathcal{P}_{n,N}$ chosen according to the conditional distribution of $\Pi_{n,N}$ given $\mathbf{Y}_n = \mathbf{y}$.
- The distribution of $\Pi_{n,N}\mathbf{Y}_n$ is computed by

$$\Pr(\Pi_{n,N}\mathbf{Y}_n = \mathbf{y}) = \sum_{P\in\mathcal{P}_{n,N}} \Pr(\Pi_{n,N} = P \mid \mathbf{Y}_n = \mathbf{y}|_{[n]}) \Pr(\mathbf{Y}_n = \mathbf{y}|_{[n]}) \mathbf{1}(P(\mathbf{y}|_{[n]}) = \mathbf{y}), \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Generative consistency

Definition (Generative consistency (Definition 4.1 of PFSNA))

Let $\mathbf{Y}_n$ *and* $\mathbf{Y}_N$ *be random* $\{0,1\}$*-valued arrays and let* $\Pi_{n,N}$ *be a generating scheme. Then* $\mathbf{Y}_n$ *and* $\mathbf{Y}_N$ *are* consistent with respect to $\Pi_{n,N}$ *if*

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

for $\Pi_{n,N}\,\mathbf{Y}_n$ *defined by the distribution in* (1)*.*

Duality between generative consistency and consistency under selection:

For any $\mathbf{Y}_n$ and generating mechanism $\Pi_{n,N}$, define $\mathbf{Y}_N$ by $\mathbf{Y}_N = \Pi_{n,N}\,\mathbf{Y}_n$. Then by the defining property of an evolution map, $\mathbf{Y}_n$ and $\mathbf{Y}_N$ enjoy the relationship

$$\mathbf{S}_{n,N}\,\mathbf{Y}_N = \mathbf{S}_{n,N}\,\Pi_{n,N}\,\mathbf{Y}_n = \mathbf{Y}_n \quad \text{with probability 1;}$$

that is, $\mathbf{Y}_n$ and $\Pi_{n,N}\,\mathbf{Y}_n$ are consistent under selection by default.

Preferential attachment model (Barabási–Albert)

- Dynamics based on Simon's preferential attachment scheme for heavy-tailed distributions.
- Vertices arrive one at a time and attach preferentially to previous vertices based on their degree.

Formal definition:

- Take $m \geq 1$ (integer) and $\delta > -m$ (real number) so that each new vertex attaches randomly to m existing vertices with probability increasing with degree.
- Initiate at a graph $\mathbf{y}_0$ with $n_0 \geq 1$ vertices, which then evolves successively into $\mathbf{y}_1, \mathbf{y}_2, \ldots$ by connecting a new vertex to the existing graph at each step.
- For any $\mathbf{y} = (y_{ij})_{1 \leq i,j \leq n}$ and every $i = 1, \ldots, n$, the *degree* of i in $\mathbf{y}$ is the number of edges incident to i,

$$\deg_{\mathbf{y}}(i) = \sum_{j \neq i} y_{ij}.$$

- At step $n \geq 1$, a new vertex v_n attaches to $m \geq 1$ vertices in $\mathbf{y}_{n-1}$, with each of the m vertices v' chosen independently without replacement with probability proportional to

$$\deg_{\mathbf{y}_{n-1}}(v') + \delta/m.$$

Barabási–Albert model (Generative scheme)

- In keeping with the notation of Section 4.1, let $\Pi_{k,n}^{\delta,m}$, $k \leq n$, denote the generating mechanism for the process parameterized by $m \geq 1$ and $\delta > -m$.
- By letting the parameters $n_0 \geq 1$, $m \geq 1$, and $\delta > -m$ vary over all permissible values and treating the initial conditions $\mathbf{y}_0$ and n_0 as fixed, the above generating mechanism determines a family of distributions for each finite sample size $n \geq 1$, where n is the number of vertices that have been added to $\mathbf{y}_0$.
- For each $n \geq 1$, this process gives a collection of distributions $\mathcal{M}_n$ indexed by (m, δ), and each distribution in $\mathcal{M}_k$ indexed by (m, δ) is related to a distribution in $\mathcal{M}_n$, $n \geq k$, with the same parameters through the preferential attachment scheme $\Pi_{k,n}^{\delta,m}$ associated to the model.
- For any choice of parameter (δ, m), we express the relationship between $\mathbf{Y}_k$ and $\mathbf{Y}_n$, $n \geq k$, by

$$\mathbf{Y}_n =_{\mathcal{D}} \Pi_{k,n}^{\delta,m} \mathbf{Y}_k\,.$$

Barabási–Albert model (Empirical properties)

Sparsity:

- Let $\mathbf{y} = (\mathbf{y}^{(n)})_{n\geq 1}$ be sequence of graphs ($\mathbf{y}^{(n)}$ has n vertices).
- Call $\mathbf{y}$ *sparse* if
$$\lim_{n\to\infty} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} y_{ij}^{(n)} = 0.$$
- Under BA model, $(\mathbf{Y}_n)_{n\geq 1}$ grows by adding one vertex at a time with m new edges, so that
$$\frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} Y_{ij} = \frac{1}{n(n-1)}(mn + n_0) \to 0 \quad \text{as } n \to \infty.$$
- Networks under BA model are sparse with probability 1.

Power law degree distribution:

- For $k \geq 1$, let
$$p_{\mathbf{y}}(k) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(\deg_{\mathbf{y}}(i) = k).$$
- A sequence $\mathbf{y} = (\mathbf{y}^{(n)})_{n\geq 1}$ exhibits power law degree distribution with exponent $\gamma > 1$ if
$$p_{\mathbf{y}^{(n)}}(k) \sim \gamma^{-k} \quad \text{for all large } k \text{ as } n \to \infty,$$
where $a(k) \sim b(k)$ indicates that $a(k)/b(k) \to 1$ as $k \to \infty$.
- BA model with parameter (δ, m) has power law degree distribution with exponent $3 + \delta/m$ with probability 1.

Power law and 'scale-free' networks

- Many real-world networks believed to exhibit power law, or nearly power law, degree distribution (Barabási–Albert, ...).
- Heuristic check: power law degree distribution implies

$$\log p_{\mathbf{y}}(k) \sim -\gamma \log(k), \qquad \text{large } k \geq 1. \tag{2}$$

- Yule–Simon distribution (dotted) vs. line $-3\log(k)$ (solid).

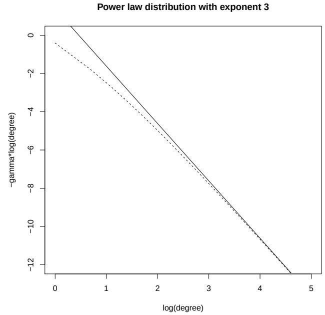


Figure: Dotted line shows log-log plot of the Yule–Simon distribution for $\gamma = 3$. Solid line shows the linear approximation in (2) by approximating $\Gamma(\gamma)/\Gamma(k+\gamma) \sim \gamma^{-k}$, which holds asymptotically for large values of k.

Random walk (RW) models

- Add a new edge at each step (instead of new vertex as in BA model).
- Start with initial graph $\mathbf{y}_0$ and evolve $\mathbf{y}_1, \mathbf{y}_2, \ldots$ as follows.
  - At step $n \geq 1$, choose vertex v_n in $\mathbf{y}_{n-1}$ randomly with distribution F_n (which can depend on $\mathbf{y}_{n-1}$). Then draw a random nonnegative integer L_n from distribution also depending on $\mathbf{y}_{n-1}$.
  - Given v_n and L_n, perform a simple random walk on $\mathbf{y}_{n-1}$ for L_n steps starting at v_n.
  - If after L_n steps the random walk is at $v^* \neq v_n$, then add edge between v^* and v_n; otherwise, add new vertex v^{**} and put edge between v^{**} and v_n.
- Choosing v_n by degree-biased distribution on $\mathbf{y}_{n-1}$ and taking L_n to be large simulates BA model.
- For more details on these models see Bloem-Reddy and Orbanz (`https://arxiv.org/abs/1612.06404`), Bollobas, et al (2003), and related work.

Erdős–Rényi–Gilbert model

- Classical Erdős–Rényi–Gilbert model includes each edge in random graph independently with fixed probability θ.
- Generative description: For any $\theta \in [0,1]$, define $\Pi^{\theta}_{n,N}$ as the generating scheme which acts on $\{0,1\}^{n\times n}$ by

$$\mathbf{y} \quad \mapsto \quad \Pi^{\theta}_{n,N}(\mathbf{y})$$

$$\mathbf{y} \quad \mapsto \quad \begin{pmatrix} & & & B_{1,n+1} & \cdots & B_{1,N} \\ & \mathbf{y} & & \vdots & \ddots & \vdots \\ & & & B_{n,n+1} & \cdots & B_{n,N} \\ B_{n+1,1} & \cdots & B_{n+1,n} & 0 & \cdots & B_{n+1,N} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{N,1} & \cdots & B_{N,n} & B_{N,n+1} & \cdots & 0 \end{pmatrix},$$

which fixes the upper $n \times n$ submatrix to be $\mathbf{y}$ and fills in the rest of the off-diagonal entries with i.i.d. Bernoulli random variables $(B_{ij})_{1\leq i\neq j\leq N}$ with success probability θ.

General sequential construction

- Above examples start with a base case $\mathbf{Y}_0$, from which a family of networks $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ is constructed inductively according to a random scheme.
- A generic way to specify a generative network model is to specify a conditional distribution for $\mathbf{Y}_n$ given $\mathbf{Y}_{n-1}$ such that $\mathbf{Y}_n|_{[n-1]} = \mathbf{Y}_{n-1}$ with probability 1.
- Conditional distribution $\Pr(\mathbf{Y}_n = \cdot \mid \mathbf{Y}_{n-1})$ determines the distribution of a random generating mechanism $\Pi_{n-1,n}$ in $\mathcal{P}_{n-1,n}$

  $\Longrightarrow$ $\mathbf{Y}_n$ can be expressed as $\mathbf{Y}_n = \Pi_{n-1,n}\mathbf{Y}_{n-1}$ for every $n \geq 1$.
- Composing these actions for successive values of n determines the generating mechanism $\Pi_{n,N}$, $n \leq N$, by the law of iterated conditioning:

  $\Longrightarrow$ Given $\mathbf{Y}_n$, construct $\mathbf{Y}_N = \Pi_{n,N}\mathbf{Y}_n$ by

$$\mathbf{Y}_N = \Pi_{N-1,N}(\Pi_{N-2,N-1}(\cdots(\Pi_{n,n+1}\mathbf{Y}_n))).$$

- The conditional distribution of $\mathbf{Y}_N$ given $\mathbf{Y}_n$ computed by

$$\begin{aligned}
&\Pr(\mathbf{Y}_N = \mathbf{y}^* \mid \mathbf{Y}_n = \mathbf{y}^*|_{[n]}) = \\
&= \Pr(\mathbf{Y}_N = \mathbf{y}^* \mid \mathbf{Y}_{N-1} = \mathbf{y}^*|_{[N-1]}) \times \Pr(\mathbf{Y}_{N-1} = \mathbf{y}^*|_{[N-1]} \mid \mathbf{Y}_n = \mathbf{y}^*|_{[n]}) \\
&= \prod_{i=1}^{N-n} \Pr(\Pi_{N-i,N-i+1}(\mathbf{y}^*|_{[N-i]}) = \mathbf{y}^*|_{[N-i+1]} \mid \mathbf{Y}_{N-i} = \mathbf{y}^*|_{[N-i]}).
\end{aligned}$$

Network modeling paradigm (Chapter 5) gives framework to handle sampling models (Chapter 3) and generative models (Chapter 4).

See Chapters 3–5 of *Probabilistic Foundations of Statistical Network Analysis*