Models for Network Graphs

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

April 2, 2020
Random graph models

Small-world models

Network-growth models

Exponential random graph models

Case study: Modeling collaboration among lawyers
Why statistical graph modeling?

- Statistical inference typically conducted in the context of a model
  ⇒ Models key to transition from descriptive to inferential tasks

- In practice, graph models are used for a variety of reasons:

  1) Mechanisms explaining properties observed on real-world networks
     *Ex:* small-world effects, power-law degree distributions

  2) Testing for ‘significance’ of a characteristic $\eta(G)$ in a network graph
     *Ex:* is the observed average degree unusual or anomalous?

  3) Alternative to the design-based framework for estimating $\eta(G)$
     *Ex:* model-based, e.g., maximum likelihood estimation
So far the focus has been on network analysis methods to:

- Collect relational data and construct network graphs
- Characterize and summarize their structural properties
- Obtain sample-based estimates of partially-observed structure

Emphasis now on construction and use of models for network data

**Def:** A model for a network graph is a collection

\[ \{ P_\theta(G), G \in \mathcal{G} : \theta \in \Theta \} \]

- \( \mathcal{G} \) is an ensemble of possible graphs
- \( P_\theta(\cdot) \) is a probability distribution on \( \mathcal{G} \) (often write \( P(\cdot) \))
- Parameters \( \theta \) ranging over values in parameter space \( \Theta \)
Richness of models derives from how we specify $P(\cdot)$. Methods range from the simple to the complex.

1) Let $P(\cdot)$ be uniform on $\mathcal{G}$, add structural constraints to $\mathcal{G}$
   \[\text{Ex: Erdős-Renyi random graphs, generalized random graph models}\]

2) Induce $P(\cdot)$ via application of simple generative mechanisms
   \[\text{Ex: small world, preferential attachment, copying models}\]

3) Model structural features and their effect on $G$’s topology
   \[\text{Ex: exponential random graph models}\]

Computational cost of associated inference algorithms relevant.
Classical random graph models

- Assign equal probability on all undirected graphs of given order and size
  - Specify collection $\mathcal{G}_{N_v, N_e}$ of graphs $G(V, E)$ with $|V| = N_v$, $|E| = N_e$
  - Assign $P(G) = \left(\frac{N}{N_e}\right)^{-1}$ to each $G \in \mathcal{G}_{N_v, N_e}$, where $N = |V(2)| = \binom{N_v}{2}$

- Most common variant is the Erdös-Renyi random graph model $G_{n,p}$
  - Undirected graph on $N_v = n$ vertices
  - Edge $(u, v)$ present w.p. $p$, independent of other edges

- **Simulation**: simply draw $N = \binom{N_v}{2} \approx N_v^2/2$ i.i.d. Ber($p$) RVs
  - Inefficient when $p \sim N_v^{-1}$ ⇒ sparse graph, most draws are 0
  - Skip non-edges drawing Geo($p$) i.i.d. RVs, runs in $O(N_v + N_e)$ time
Properties of $G_{n,p}$

- $G_{n,p}$ is well-studied and tractable. **Noteworthy properties:**

**P1) Degree distribution** $P(d)$ is binomial with parameters $(n - 1, p)$
- Large graphs have concentrated $P(d)$ with exponentially-decaying tails

**P2) Phase transition on the emergence of a giant component**
- If $np > 1$, $G_{n,p}$ has a giant component of size $O(n)$ w.h.p.
- If $np < 1$, $G_{n,p}$ has components of size only $O(\log n)$ w.h.p.

P3) **Small clustering coefficient** $O(n^{-1})$ and short diameter $O(\log n)$ w.h.p.
Recipe for generalization of Erdös-Renyi models

- Specify $\mathcal{G}$ of fixed order $N_v$, possessing a desired characteristic
- Assign equal probability to each graph $G \in \mathcal{G}$

Configuration model: fixed degree sequence $\{d(1), \ldots, d(N_v)\}$

- Size fixed under this model, since $N_e = \bar{d} N_v / 2 \Rightarrow \mathcal{G} \subset \mathcal{G}_{N_v, N_e}$
- Equivalent to specifying model via conditional distribution on $\mathcal{G}_{N_v, N_e}$

Configuration models useful as reference, i.e., ‘null’ models

- Ex: compare observed $G$ with $G' \in \mathcal{G}$ having power law $P(d)$
- Ex: expected group-wise edge counts in modularity measure
Results on the configuration model

P1) Phase transition on the emergence of a giant component
   ▶ Condition depends on first two moments of given $P(d)$
   ▶ Giant component has size $O(N_v)$ as in $G_{N_v,p}$
   ▶ M. Molloy and B. Reed, “A critical point for random graphs with a given degree sequence,” *Random Struct. and Alg.*, vol. 6, pp. 161-180, 1995

P2) Clustering coefficient vanishes slower than in $G_{N_v,p}$

P3) Special case of given power-law degree distribution $P(d) \sim Cd^{-\alpha}$
   ▶ For $\alpha \in (2, 3)$, short diameter $O(\log N_v)$ as in $G_{N_v,p}$
Simulating generalized random graphs

▶ Matching algorithm

Given: nodes with spokes
Randomly match mini-nodes
Sample graph

▶ Switching algorithm

Initialize
Randomly switch a pair of edges
Repeat ~100N_e times
Sample graph
Consider a sample $G^*$ of a population graph $G(V, E)$

⇒ Suppose a given characteristic $\eta(G)$ is of interest
⇒ Q: Useful estimate $\hat{\eta} = \hat{\eta}(G^*)$ of $\eta(G)$?

Statistical inference in sampling theory via design-based methods

⇒ Only source of randomness is due to the sampling design

Augment this perspective to include a model-based component

⇒ Assume $G$ drawn uniformly from the collection $\mathcal{G}$, prior to sampling

⇒ Inference on $\eta(G)$ should incorporate both randomness due to

⇒ Selection of $G$ from $\mathcal{G}$ and sampling $G^*$ from $G$
Example: size of a “hidden population”

- Directed graph $G(V, E)$, $V$ the members of the hidden population
  - Graph describing willingness to identify other members
  - Arc $(i, j)$ when ask individual $i$, mentions $j$ as a member

- For given $V$, model $G$ as drawn from a collection $\mathcal{G}$ of random graphs
  - Independently add arcs between vertex pairs w.p. $p_G$

- Graph $G^*$ obtained via one-wave snowball sampling, i.e., $V^* = V_0^* \cup V_1^*$
  - Initial sample $V_0^*$ obtained via BS from $V$ with probability $p_0$

- Consider the following RVs of interest
  - $N = |V_0^*$|: size of the initial sample
  - $M_1$: number of arcs among individuals in $V_0^*$
  - $M_2$: number of arcs from individuals in $V_0^*$ to individuals in $V_1^*$

- Snowball sampling yields measurements $n, m_1, \text{ and } m_2$ of these RVs
Method of moments estimator

- **Method of moments:** now $A_{ij} = \mathbb{I}\{(i, j) \in E\}$ also a RV

\[
\mathbb{E}[N] = \mathbb{E}\left[ \sum_i \mathbb{I}\{i \in V_0^*\} \right] = N_v p_0 = n
\]

\[
\mathbb{E}[M_1] = \mathbb{E}\left[ \sum_j \sum_{i \neq j} \mathbb{I}\{i \in V_0^*\} \mathbb{I}\{j \in V_0^*\} A_{ij} \right] = N_v (N_v - 1) p_0^2 p_G = m_1
\]

\[
\mathbb{E}[M_2] = \mathbb{E}\left[ \sum_j \sum_{i \neq j} \mathbb{I}\{i \in V_0^*\} \mathbb{I}\{j \notin V_0^*\} A_{ij} \right] = N_v (N_v - 1) p_0 (1 - p_0) p_G = m_2
\]

- Expectation w.r.t. randomness in selecting $G$ and sample $V_0^*$. Solution:

\[
\hat{p}_0 = \frac{m_1}{m_1 + m_2}, \quad \hat{p}_G = \frac{m_1 (m_1 + m_2)}{n [(n - 1) m_1 + nm_2]}, \quad \text{and} \quad \hat{N}_v = n \left( \frac{m_1 + m_2}{m_1} \right)
\]

\Rightarrow \text{Same estimates for } p_0 \text{ and } N_v \text{ as in the design-based approach}
Directly modeling $\eta(G)$

- So far considered modeling $G$ for model-based estimation of $\eta(G)$
  \[\Rightarrow\] Classical random graphs typical in social networks research

- Alternatively, one may specify a model for $\eta(G)$ directly

Example

- Estimate the power-law exponent $\eta(G) = \alpha$ from degree counts

- A power law implies the linear model $\log P(d) = C - \alpha \log d + \epsilon$
  \[\Rightarrow\] Could use a model-based estimator such as least squares

- Better form the MLE for the model $f(d; \alpha) = \frac{\alpha-1}{d_{\text{min}}} \left( \frac{d}{d_{\text{min}}} \right)^{-\alpha}$

  Hill estimator \[\Rightarrow\] $\hat{\alpha} = 1 + \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \log \left( \frac{d_i}{d_{\text{min}}} \right) \right]^{-1}$
Task 2: Assessing significance in network graphs

Consider a graph $G^{obs}$ derived from observations.

Q: Is a structural characteristic $\eta(G^{obs})$ significant, i.e., unusual?

⇒ Assessing significance requires a frame of reference, or null model.
⇒ Random graph models often used in setting up such comparisons.

Define collection $\mathcal{G}$, and compare $\eta(G^{obs})$ with values $\{\eta(G) : G \in \mathcal{G}\}$.

⇒ Formally, construct the reference distribution

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

⇒ If $\eta(G^{obs})$ found to be sufficiently unlikely under $P_{\eta, \mathcal{G}}(t)$
⇒ Evidence against the null $H_0: G^{obs}$ is a uniform draw from $\mathcal{G}$.
Example: Zachary’s karate club

- Zachary’s karate club has clustering coefficient \( \text{cl}(G^{obs}) = 0.2257 \)
  \( \Rightarrow \) Random graph models to assess whether the value is unusual

- Construct two ‘comparable’ abstract frames of reference
  1) Collection \( G_1 \) of random graphs with same \( N_v = 34 \) and \( N_e = 78 \)
  2) Add the constraint that \( G_2 \) has the same degree distribution as \( G^{obs} \)

- \( |G_1| \approx 8.4 \times 10^{96} \) and \( |G_2| \) much smaller, but still large
  \( \Rightarrow \) Enumerating \( G_1 \) intractable to obtain \( P_{\eta,G_1}(t) \) exactly

- Instead use simulations to approximate both distributions
  \( \Rightarrow \) Draw 10,000 uniform samples \( G \) from each \( G_1 \) and \( G_2 \)
  \( \Rightarrow \) Calculate \( \eta(G) = \text{cl}(G) \) for each sample, plot histograms
Example: Zachary’s karate club (cont.)

- Plot histograms to approximate the distributions

![Histograms of clustering coefficients](Fig. 6.1)

*Fig. 6.1* Histograms of clustering coefficients $c_l(G)$ for random graphs of order $N_v = 34$, generated uniformly from those with the same number of edges $N_e$ (top, in red) and the same degree distribution (bottom, in blue) as in the karate club network.

- Unlikely to see a value $c_l(G_{obs}) = 0.2257$ under both graph models
  - Ex: only 3 out of 10,000 samples from $G_1$ had $c_l(G) > 0.2257$

- Strong evidence to reject $G_{obs}$ obtained as sample from $G_1$ or $G_2$
Task 3: Detecting network motifs

- Related use of random graph models is for detecting network motifs
  ⇒ Find the simple ‘building blocks’ of a large complex network

- **Def:** Network motifs are small subgraphs occurring far more frequently in a given network than in comparable random graphs

- **Ex:** there are $L_3 = 13$ different connected 3-vertex subdigraphs

- Let $N_i$ be the count in $G$ of the $i$-th type $k$-vertex subgraph, $i = 1, \ldots, L_k$
  ⇒ Each value $N_i$ can be compared to a suitable reference $P_{N_i,G}$
  ⇒ Subgraphs for which $N_i$ is extreme are declared as network motifs
Example: AIDS blog network

- **AIDS blog network** $G^{obs}$ with $N_v = 146$ bloggers and $N_e = 183$ links
  - Examining evidence for motifs of size $k = 3$ and $4$ vertices

- Simulated 10,000 digraphs using a switching algorithm
  - Fixed in- and out-degree sequences, mutual edges as in $G^{obs}$
  - Constructed approximate reference distributions $P_{N_i,G(t)}$

- Ex: two bloggers with a mutual edge and a common ‘authority’
Challenges in detecting motifs

- Individual motifs frequently overlap with other copies of itself
  ⇒ May require them to be frequent and mostly disjoint subgraphs

- With large graphs come significant computational challenges
  ⇒ Number of different potential motifs $L_k$ grows fast with $k$
  
  Ex: Connected subdigraphs $L_3 = 13$, $L_4 = 199$, $L_5 = 9364$

- May sample subgraphs $H$ along with the HT estimation framework

$$\hat{N}_i = \sum_{H \text{ of type } i} \pi_H^{−1}$$
Small-world models

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Case study: Modeling collaboration among lawyers
Arguably the most important innovation in modern graph modeling is the transition from traditional random graph models to models mimicking observed "real-world" properties.
A “small” world?

- **Six degrees of separation** popularized by a play [Guare’90]
  - Short paths between us and everyone else on the planet
  - Term relatively new, the concept has a long history

- Traced back to F. Karinthy in the 1920s
  - ‘Shrinking’ modern world due to increased human connectedness
  - **Challenge**: find someone whose distance from you is > 5
  - Inspired by G. Marconi’s Nobel prize speech in 1909

- First mathematical treatment [Kochen-Pool’50]
  - Formally modeled the mechanics of social networks
  - But left ‘degrees of separation’ question unanswered

- Chain of events led to a groundbreaking experiment [Milgram’67]
Milgram’s experiment

- **Q1:** What is the typical geodesic distance between two people?
  - ⇒ Experiment on the global friendship (social) network
  - ⇒ Cannot measure in full, so need to probe explicitly

- **S. Milgram’s ingenious small-world experiment in 1967**
  - 296 letters sent to people in Wichita, KS and Omaha, NE
  - Letters indicated a (unique) contact person in Boston, MA
  - Asked them to forward the letter to the contact, following rules

- **Def:** friend is someone known on a first-name basis
  - **Rule 1:** If contact is a friend then send her the letter; else
  - **Rule 2:** Relay to friend most-likely to be a contact’s friend

- **Q2:** How many letters arrived? How long did they take?
Milgram’s experimental results

- 64 of 296 letter reached the destination, average path length $\bar{\ell} = 6.2$
  ⇒ Inspiring Guare’s ‘6 degrees of separation’

- **Conclusion**: short paths connect arbitrary pairs of people

Moment to reflect

- Milgram demonstrated that short paths are in abundance

- Q: Is the small-world theory reasonable? Sure, e.g., assumes:
  - We have 100 friends, each of them has 100 other friends, . . .
  - After 5 degrees we get $10^{10}$ friends $>\text{twice the Earth’s population}$

- Not a realistic model of social networks exhibiting:
  - Homophily [Lazarfeld’54]
  - Triadic closure [Rapoport’53]

- Q: How can networks be highly-structured locally and globally small?
Structure and randomness as extremes

- **One-dimensional regular lattice** $G_r$ on $N_v$ vertices
  - Each node is connected to its $2r$ closest neighbors ($r$ to each side)
  
  Structure yields high clustering and high diameter
  
  $$\text{cl}(G_r) = \frac{3r - 3}{4r - 2} \quad \text{and} \quad \text{diam}(G_r) = \frac{N_v}{2r}$$

- Other extreme is a $G_{N_v,p}$ random graph with $p = O(N_v^{-1})$
  
  Randomness yields low clustering and low diameter
  
  $$\text{cl}(G_{N_v,p}) = O(N_v^{-1}) \quad \text{and} \quad \text{diam}(G_{N_v,p}) = O(\log N_v)$$
The Watts-Strogatz model

- **Small-world model**: blend of structure with little randomness

  **S1**: Start with regular lattice that has desired clustering
  
  **S2**: Introduce randomness to generate shortcuts in the graph

  ⇒ Each edge is randomly rewired with (small) probability \( p \)

  ![Diagram](https://via.placeholder.com/150)

  - Rewiring interpolates between the **regular** and **random** extremes
Numerical results

- Simulate Watts-Strogatz model with $N_v = 1,000$ and $r = 6$
  - Rewiring probability $p$ varied from 0 (lattice $G_r$) to 1 (random $G_{N_v,p}$)
  - Normalized $\text{cl}(G)$ and $\text{diam}(G)$ to maximum values ($p = 0$)

- Broad range of $p \in [10^{-3}, 10^{-1}]$ yields small $\text{diam}(G)$ and high $\text{cl}(G)$

![Graph showing clustering coefficient ($\text{cl}(G)$) and average distance ($\text{diam}(G)$) as functions of $\log_{10}(p)$]
Closing remarks

- Structural properties of Watts-Strogatz model [Barrat-Weigt’00]
  **P1:** Large $N_v$ analysis of clustering coefficient

  \[ \text{cl}(G) \approx \frac{3r - 3}{4r - 2}(1 - p^3) = \text{cl}(G_r)(1 - p^3) \]

  **P2:** Degree distribution concentrated around $2r$

- Small-world graph models of interest across disciplines

- Particularly relevant to ‘communication’ in a broad sense
  - Spread of news, gossip, rumors
  - Spread of natural diseases and epidemics
  - Search of content in peer-to-peer networks
Network-growth models

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Case study: Modeling collaboration among lawyers
Many networks grow or otherwise evolve in time

*Ex: Web, scientific citations, Twitter, genome . . .*

General approach to model construction mimicking network growth

- Specify simple mechanisms for network dynamics
- Study emergent structural characteristics as time $t \to \infty$

Q: Do these properties match observed ones in real-world networks?

Two fundamental and popular classes of growth processes

- Preferential attachment models
- Copying models

Tenable mechanisms for popularity and gene duplication, respectively
Preferential attachment model

- Simple model for the creation of e.g., links among Web pages
- Vertices are created one at a time, denoted 1, . . . , $N_v$
- When node $j$ is created, it makes a single arc to $i$, $1 \leq i < j$
- Creation of $(j, i)$ governed by a probabilistic rule:
  - With probability $p$, $j$ links to $i$ chosen uniformly at random
  - With probability $1 - p$, $j$ links to $i$ with probability $\propto d_i^{in}$
- The resulting graph is directed, each vertex has $d_v^{out} = 1$
- Preferential attachment model leads to “rich-gets-richer” dynamics
  - $\Rightarrow$ Arcs formed preferentially to (currently) most popular nodes
  - $\Rightarrow$ Prob. that $i$ increases its popularity $\propto i$’s current popularity
The preferential attachment model gives rise to a power-law in-degree distribution with exponent $\alpha = 1 + \frac{1}{1-p}$, i.e.,

$$P(d_{in} = d) \propto d^{-(1+\frac{1}{1-p})}$$

- **Key:** "j links to i with probability $\propto d_{in}^i$" equivalent to copying, i.e., "j chooses k uniformly at random, and links to i if (k, i) $\in E$"

- **Reflect:** Copy other’s decision vs. independent decisions in $G_{n,p}$

- **As $p \to 0$ $\Rightarrow$ Copying more frequent $\Rightarrow$ Smaller $\alpha \to 2$
  - **Intuitive:** more likely to see extremely popular pages (heavier tail)
Barabási-Albert (BA) model is for undirected graphs

Initial graph $G_{BA}(0)$ of $N_v(0)$ vertices and $N_e(0)$ edges ($t = 0$)

For $t = 1, 2, \ldots$ current graph $G_{BA}(t-1)$ grows to $G_{BA}(t)$ by:

- Adding a new vertex $u$ of degree $d_u(t) = m \geq 1$
- The $m$ new edges are incident to $m$ different vertices in $G_{BA}(t-1)$
- New vertex $u$ is connected to $v \in V(t-1)$ w.p.

$$P ((u, v) \in E(t)) = \frac{d_v(t-1)}{\sum_{v' d_{v'}(t-1)}$$

Vertices connected to $u$ preferentially towards higher degrees

$\Rightarrow G_{BA}(t)$ has $N_v(t) = N_v(0) + t$ and $N_e(t) = N_e(0) + tm$

BA model ambiguous in how to select \( m \) vertices \( \propto \) to their degree

\[ \Rightarrow \text{Joint distribution not specified by marginal on each vertex} \]

Linearized chord diagram (LCD) model removes ambiguities

For \( m = 1 \), start with \( G_{\text{LCD}}(0) \) consisting of a vertex with a self-loop

For \( t = 1, 2, \ldots \) current graph \( G_{\text{LCD}}(t - 1) \) grows to \( G_{\text{LCD}}(t) \) by:

- Adding a new vertex \( v_t \) with an edge to \( v_s \in V(t) \)
- Vertex \( v_s, 1 \leq s \leq t \) is chosen w.p.

\[
P(s = j) = \begin{cases} 
\frac{d_{vj}(t-1)}{2(t-1)}, & \text{if } 1 \leq j \leq t - 1, \\
\frac{1}{2(t-1)}, & \text{if } j = t
\end{cases}
\]

For \( m > 1 \) simply run the above process \( m \) times for each \( t \)
- Collapse all created vertices into a single one, retaining edges

Properties of the LCD model

P1) The LCD model allows for loops and multi-edges, occurring rarely

P2) $G_{LCD}(t)$ has power-law degree distribution with $\alpha = 3$, as $t \to \infty$

P3) The BA model yields connected graphs if $G_{BA}(0)$ connected

$\Rightarrow$ Not true for the LCD model, but $G_{LCD}(t)$ connected w.h.p.

P4) Small-world behavior

$$\text{diam}(G_{LCD}(t)) = \begin{cases} 
O(\log N_v(t)), & m = 1 \\
O\left(\frac{\log N_v(t)}{\log \log N_v(t)}\right), & m > 1
\end{cases}$$

P5) Unsatisfactory clustering, since small for $m > 1$

$$\mathbb{E}[\text{cl}(G_{LCD}(t))] \approx \frac{m - 1}{8} \frac{(\log N_v(t))^2}{N_v(t)}$$

$\Rightarrow$ Marginally better than $O(N_v^{-1})$ in classical random graphs
Copying is another mechanism of fundamental interest

Ex: gene duplication to re-use information in organism’s evolution

Different from preferential attachment, but still results in power laws

Initialize with a graph $G_C(0)$ ($t = 0$)

For $t = 1, 2, \ldots$ current graph $G_C(t - 1)$ grows to $G_C(t)$ by:

- Adding a new vertex $u$
- Choosing vertex $v \in V(t - 1)$ with uniform probability $\frac{1}{N_v(t-1)}$
- Joining vertex $u$ with $v$’s neighbors independently w.p. $p$

Case $p = 1$ leads to full duplication of edges from an existing node

F. Chung et al, “Duplication models for biological networks,”
Asymptotic degree distribution

- Degree distribution tends to a power law w.h.p. [Chung et al’03]
  - Exponent $\alpha$ is the plotted solution to the equation
    \[
    p(\alpha - 1) = 1 - p^{\alpha-1}
    \]

- Full duplication does not lead to power-law behavior; but does if
  - Partial duplication performed a fraction $q \in (0, 1)$ of times
Fitting network growth models

- Most common practical usage of network growth models is predictive

  **Goal:** compare characteristics of $G^{obs}$ and $G(t)$ from the models

- Little attempt to date to fit network growth models to data
  - Expected due to simplicity of such models
  - Still useful to estimate e.g., the duplication probability $p$

- To fit a model ideally would like to observe a sequence $\{G^{obs}(\tau)\}_{\tau=1}^t$
  - Unfortunately, such dynamic network data is still fairly elusive

- **Q:** Can we fit a network growth model to a single snap-shot $G^{obs}$?
- **A:** Yes, if we leverage the Markovianity of the growth process
Similar to all network growth models described so far, suppose:

**As1:** A single vertex is added to $G(t - 1)$ to create $G(t)$; and
**As2:** The manner in which it is added depends only on $G(t - 1)$

In other words, we assume $\{G(t)\}_{t=0}^{\infty}$ is a Markov chain

Let graph $\delta(G(t), v)$ be obtained by deleting $v$ and its edges from $G(t)$

**Def:** vertex $v$ is removable if $G(t)$ can be obtained from $\delta(G(t), v)$ via copying. If $G(t)$ has no removable vertices, we call it irreducible

The class of duplication-attachment (DA) models satisfies:

(i) The initial graph $G(0)$ is irreducible; and
(ii) $P_{\theta}(G(t) \mid G(t - 1)) > 0 \iff G(t)$ obtained by copying a vertex in $G(t - 1)$

Example: reducible graph

- Vertex \( v_A \) is removable (likewise \( v_C \) by symmetry)
  - Obtain \( G(t) \) from \( \delta(G(t, v_a)) \) by copying \( v_C \)

- This implies that \( G(t) \) is reducible
  - Notice though that \( v_B \) or \( v_D \) are not removable
MLE for DA model parameters

- Suppose that $G^{obs} = G(t)$ represents the observed network graph.
- The likelihood for the parameter $\theta$ is recursively given by

$$\mathcal{L}(\theta; G(t)) = \frac{1}{t} \sum_{v \in \mathcal{R}_{G(t)}} P_{\theta}(G(t) \mid \delta(G(t), v)) \mathcal{L}(\theta; \delta(G(t), v))$$

$\Rightarrow \mathcal{R}_{G(t)}$ is the set of all removable nodes in $G(t)$.
- The MLE for $\theta$ is thus defined as

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta; G(t))$$

$\Rightarrow$ Computing $\mathcal{L}(\theta; G(t))$ non-trivial, even for modest-size graphs.
- Monte Carlo methods to approximate $\mathcal{L}(\theta; G(t))$ [Wiupf et al’06]

$\Rightarrow$ Open issues: vector $\theta$, other growth models, scalability.
Exponential random graph models

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Case study: Modeling collaboration among lawyers
Good statistical network graph models should be [Robbins-Morris’07]:
⇒ Estimable from and reasonably representative of the data
⇒ Theoretically plausible about the underlying network effects
⇒ Discriminative among competing effects to best explain the data

Network-based versions of canonical statistical models
⇒ Regression models - Exponential random graph models (ERGMs)
⇒ Latent variable models - Latent network models
⇒ Mixture models - Stochastic block models

Focus here on ERGMs, also known as $p^*$ models

Exponential family

- **Def:** discrete random vector \( Z \in \mathcal{Z} \) belongs to an exponential family if

\[
P_{\theta}(Z = z) = \exp \left\{ \theta^\top g(z) - \psi(\theta) \right\}
\]

- \( \theta \in \mathbb{R}^p \) is a vector of parameters and \( g : \mathcal{Z} \mapsto \mathbb{R}^p \) is a function
- \( \psi(\theta) \) is a normalization term, ensuring \( \sum_{z \in \mathcal{Z}} P_{\theta}(z) = 1 \)
- **Ex:** Bernoulli, binomial, Poisson, geometric distributions

- For continuous exponential families, the pdf has an analogous form
  - **Ex:** Gaussian, Pareto, chi-square distributions

- Exponential families share useful algebraic and geometric properties
  - Mathematically convenient for inference and simulation
Let $G(V, E)$ be a random undirected graph, with $Y_{ij} := \mathbb{I}\{(i, j) \in E\}$.

- Matrix $Y = [Y_{ij}]$ is the random adjacency matrix, $y = [y_{ij}]$ a realization.
- An ERGM specifies in exponential family form the distribution of $Y$, i.e.,

$$
P_\theta(Y = y) = \left( \frac{1}{\kappa(\theta)} \right) \exp \left\{ \sum_H \theta_H g_H(y) \right\},$$

where

(i) each $H$ is a configuration, meaning a set of possible edges in $G$;
(ii) $g_H(y)$ is the network statistic corresponding to configuration $H$

$$g_H(y) = \prod_{y_{ij} \in H} y_{ij} = \mathbb{I}\{H \text{ occurs in } y\}$$

(iii) $\theta_H \neq 0$ only if all edges in $H$ are conditionally dependent; and
(iv) $\kappa(\theta)$ is a normalization constant ensuring $\sum_y P_\theta(y) = 1$
Graph order $N_v$ is fixed and given, only edges are random

⇒ Assumed unweighted, undirected edges. Extensions possible

ERGMs describe random graphs ‘built-on’ localized patterns

▶ These configurations are the structural characteristics of interest
▶ Ex: Are there reciprocity effects? Add mutual arcs as configurations
▶ Ex: Are there transitivity effects? Consider triangles

(IN)dependence is conditional on all other variables (edges) in $G$

⇒ Control configurations relevant (i.e., $\theta_H \neq 0$) to the model

Well-specified dependence assumptions imply particular model classes
In positing an ERGM for a network, one implicitly follows five steps

⇒ Explicit choices connecting hypothesized theory to data analysis

Step 1: Each edge (relational tie) is regarded as a random variable

Step 2: A dependence hypothesis is proposed

Step 3: Dependence hypothesis implies a particular form to the model

Step 4: Simplification of parameters through e.g., homogeneity

Step 5: Estimate and interpret model parameters
Example: Bernoulli random graphs

- Assume edges present independently of all other edges (e.g., in $G_{n,p}$)
  - Simplest possible (and unrealistic) dependence assumption

- For each $(i, j)$, we assume $Y_{ij}$ independent of $Y_{uv}$, for all $(u, v) \neq (i, j)$
  - $\theta_H = 0$ for all $H$ involving two or more edges

- Edge configurations i.e., $g_H(y) = y_{ij}$ relevant, and the ERGM becomes

  $$P_\theta(Y = y) = \left(\frac{1}{\kappa(\theta)}\right) \exp\left\{\sum_{i,j} \theta_{ij} y_{ij}\right\}$$

- Specifies that edge $(i, j)$ present independently, with probability

  $$p_{ij} = \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}$$
Constraints on parameters: homogeneity

- Too many parameters makes estimation infeasible from single $y$
  $\Rightarrow$ Under independence have $N_v^2$ parameters $\{\theta_{ij}\}$. Reduction?

- Homogeneity across all $G$, i.e., $\theta_{ij} = \theta$ for all $(i, j)$ yields

$$P_\theta(Y = y) = \left( \frac{1}{\kappa(\theta)} \right) \exp \{ \theta L(y) \}$$

- Relevant statistic is the number of edges observed $L(y) = \sum_{i,j} y_{ij}$
- ERGM identical to $G_{n,p}$, where $p = \frac{\exp \theta}{1 + \exp \theta}$

**Ex:** suppose we know a priori that vertices fall in two sets

- Can impose homogeneity on edges within and between sets, i.e.,

$$P_\theta(Y = y) = \left( \frac{1}{\kappa(\theta)} \right) \exp \{ \theta_1 L_1(y) + \theta_{12} L_{12}(y) + \theta_2 L_2(y) \}$$
Example: Markov random graphs

- Markov dependence notion for network graphs [Frank-Strauss'86]
  - Assumes two ties are dependent if they share a common node
  - Edge status $Y_{ij}$ dependent on any other edge involving $i$ or $j$

Theorem

Under homogeneity, $G$ is a Markov random graph if and only if

$$P_\theta(Y = y) = \left( \frac{1}{\kappa(\theta)} \right) \exp \left\{ \sum_{k=1}^{N_v-1} \theta_k S_k(y) + \theta_T T(y) \right\}, \text{ where}$$

$S_k(y)$ is the number of $k$-stars, and $T(y)$ the number of triangles

\begin{align*}
1\text{-star}=\text{edge} & \quad 2\text{-star} & \quad 3\text{-star} & \quad \text{Triangle}
\end{align*}
Alternative statistics

- Including many higher-order terms challenges estimation
  ⇒ High-order star effects often omitted, e.g., $\theta_k = 0, k \geq 4$
  ⇒ But these models tend to fit real data poorly. Dilemma?

- Idea: Impose parametric form $\theta_k \propto (-1)^k \lambda^{2-k}$ [Snijders et al’06]

- Combine $S_k(y), k \geq 2$ into a single alternating $k$-star statistic, i.e.,

$$AKS_\lambda(y) = \sum_{k=2}^{N_v-1} (-1)^k \frac{S_k(y)}{\lambda^{k-2}}, \quad \lambda > 1$$

- Can show $AKS_\lambda(y) \propto$ the geometrically-weighted degree count

$$GWD_\gamma(y) = \sum_{d=0}^{N_v-1} e^{-\gamma d} N_d(y), \quad \gamma > 0$$

⇒ $N_d(y)$ is the number of vertices with degree $d$
Incorporating vertex attributes

- Straightforward to incorporate vertex attributes to ERGMs
  Ex: gender, seniority in organization, protein function

- Consider a realization \( x \) of a random vector \( X \in \mathbb{R}^{N_v} \) defined on \( V \)

- Specify an exponential family form for the conditional distribution

\[
P_{\theta}(Y = y \mid X = x)
\]

⇒ Will include additional statistics \( g(\cdot) \) of \( y \) and \( x \)

- Ex: configurations for Markov, binary vertex attributes
MLE for the parameter vector $\theta$ in an ERGM is

$$\hat{\theta} = \arg \max_{\theta} \left\{ \theta^\top g(y) - \psi(\theta) \right\}, \quad \text{where } \psi(\theta) := \log \kappa(\theta)$$

Optimality condition yields

$$g(y) = \nabla \psi(\theta) \big|_{\theta = \hat{\theta}}$$

Using also that $\mathbb{E}_\theta[g(Y)] = \nabla \psi(\theta)$, the MLE solves

$$\mathbb{E}_{\hat{\theta}}[g(Y)] = g(y)$$

Unfortunately $\psi(\theta)$ cannot be computed except for small graphs

$\Rightarrow$ Involves a summation over $2^{\binom{N_v}{2}}$ values of $y$ for each $\theta$

$\Rightarrow$ Numerical methods needed to obtain approximate values of $\hat{\theta}$
Proof of $\mathbb{E}[g(Y)] = \nabla \psi(\theta)$

- The pmf of $Y$ is $P_\theta(Y = y) = \exp\left\{ \theta^\top g(y) - \psi(\theta) \right\}$, hence

  $$
  \mathbb{E}_\theta[g(Y)] = \sum_y g(y) P_\theta(Y = y)
  = \sum_y g(y) \exp \left\{ \theta^\top g(y) - \psi(\theta) \right\}
  $$

- Recall $\psi(\theta) = \log \sum_y \exp \left\{ \theta^\top g(y) \right\}$ and use the chain rule

  $$
  \nabla \psi(\theta) = \frac{\sum_y g(y) \exp \left\{ \theta^\top g(y) \right\}}{\sum_y \exp \left\{ \theta^\top g(y) \right\}} = \frac{\sum_y g(y) \exp \left\{ \theta^\top g(y) \right\}}{\exp \psi(\theta)}
  = \sum_y g(y) \exp \left\{ \theta^\top g(y) - \psi(\theta) \right\}
  $$

- The red and blue sums are identical $\Rightarrow \mathbb{E}_\theta[g(Y)] = \nabla \psi(\theta)$ follows
Markov chain Monte Carlo MLE

- Idea: for fixed $\theta_0$, maximize instead the log-likelihood ratio

$$r(\theta, \theta_0) = \ell(\theta) - \ell(\theta_0) = (\theta - \theta_0) \top g(y) - [\psi(\theta) - \psi(\theta_0)]$$

- Key identity: will show that

$$\exp \{\psi(\theta) - \psi(\theta_0)\} = \mathbb{E}_{\theta_0} \left[ \exp \{(\theta - \theta_0) \top g(Y)\} \right]$$

- Markov chain Monte Carlo MLE algorithm to search over $\theta$

  **Step 1:** draw samples $Y_1, \ldots, Y_n$ from the ERGM under $\theta_0$
  **Step 2:** approximate the above $\mathbb{E}_{\theta_0} [\cdot]$ via sample averaging
  **Step 3:** the logarithm of the result approximates $\psi(\theta) - \psi(\theta_0)$
  **Step 4:** evaluate an $\approx$ log-likelihood ratio $r(\theta, \theta_0)$

- For large $n$, the maximum value found approximates the MLE $\hat{\theta}$
Derivation of key identity

- Recall \( \exp \psi(\mathbf{\theta}) = \sum_y \exp \left\{ \mathbf{\theta}^\top \mathbf{g}(y) \right\} \) to write

\[
\exp \left\{ \psi(\mathbf{\theta}) - \psi(\mathbf{\theta}_0) \right\} = \frac{\sum_y \exp \left\{ \mathbf{\theta}^\top \mathbf{g}(y) \right\}}{\exp \psi(\mathbf{\theta}_0)}
\]

- Multiplying and dividing by \( \exp \left\{ \mathbf{\theta}_0^\top \mathbf{g}(y) \right\} > 0 \) yields

\[
\exp \left\{ \psi(\mathbf{\theta}) - \psi(\mathbf{\theta}_0) \right\} = \sum_y \exp \left\{ (\mathbf{\theta} - \mathbf{\theta}_0)^\top \mathbf{g}(y) \right\} \times \frac{\exp \left\{ \mathbf{\theta}_0^\top \mathbf{g}(y) \right\}}{\exp \psi(\mathbf{\theta}_0)}
\]

\[
= \sum_y \exp \left\{ (\mathbf{\theta} - \mathbf{\theta}_0)^\top \mathbf{g}(y) \right\} P_{\mathbf{\theta}_0}(Y = y)
\]

\[
= \mathbb{E}_{\mathbf{\theta}_0} \left[ \exp \left\{ (\mathbf{\theta} - \mathbf{\theta}_0)^\top \mathbf{g}(Y) \right\} \right]
\]

- Used \( \exp \left\{ \mathbf{\theta}_0^\top \mathbf{g}(y) - \psi(\mathbf{\theta}_0) \right\} \) is the exponential family pmf \( P_{\mathbf{\theta}_0}(Y = y) \)
Model goodness-of-fit

- **Best fit** chosen from a given class of models ... may not be a **good fit** to the data if **model class not rich enough**

- Assessing goodness-of-fit for ERGMs
  
  **Step 1**: simulate numerous random graphs from the fitted model
  
  **Step 2**: compare high-level characteristics with those of $G^{obs}$
  
  **Ex**: distributions of degree, centrality, diameter

- If significant differences found in $G^{obs}$, conclude
  
  $\Rightarrow$ Systematic gap between specified model class and data
  
  $\Rightarrow$ Lack of goodness-of-fit

- **Take home**: model specification for ERGMs highly nontrivial
  
  $\Rightarrow$ Goodness-of-fit diagnostics can play key facilitating role
Case study

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Case study: Modeling collaboration among lawyers
Lawyer collaboration network

- Network $G^{obs}$ of working relationships among lawyers [Lazega’01]
  - Nodes are $N_v = 36$ partners, edges indicate partners worked together

- Data includes various node-level attributes:
  - Seniority (node labels indicate rank ordering)
  - Office location (triangle, square or pentagon)
  - Type of practice, i.e., litigation (red) and corporate (cyan)
  - Gender (three partners are female labeled 27, 29 and 34)

- Goal: study cooperation among social actors in an organization
Modeling lawyer collaborations

▶ Assess network effects $S_1(y) = N_e$ and alternating $k$-triangles statistic

$$AKT_\lambda(y) = 3T_1(y) + \sum_{k=2}^{N_v-2} (-1)^{k+1} \frac{T_k(y)}{\lambda^{k-1}}$$

$\Rightarrow T_k(y)$ counts sets of $k$ individual triangles sharing a common base

▶ Test the following set of exogenous effects:

$h^{(1)}(x_i, x_j) = \text{seniority}_i + \text{seniority}_j$, $h^{(2)}(x_i, x_j) = \text{practice}_i + \text{practice}_j$

$h^{(3)}(x_i, x_j) = \mathbb{I}\{\text{practice}_i = \text{practice}_j\}$, $h^{(4)}(x_i, x_j) = \mathbb{I}\{\text{gender}_i = \text{gender}_j\}$

$h^{(5)}(x_i, x_j) = \mathbb{I}\{\text{office}_i = \text{office}_j\}$, $h(x_i, x_j) := [h^{(1)}(x_i, x_j), \ldots, h^{(5)}(x_i, x_j)]^T$

▶ Resulting ERGM

$$\mathbb{P}_{\theta, \beta}(Y = y|X = x) = \frac{1}{\kappa(\theta, \beta)} \exp \left\{ \theta_1 S_1(y) + \theta_2 AKT_\lambda(y) + \beta^T g(y, x) \right\}$$

$g(y, x) = \sum_{i,j} y_{ij} h(x_i, x_j)$
Fitting results using the MCMC MLE approach

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>‘Standard Error’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density ($\theta_1$)</td>
<td>-6.2073</td>
<td>0.5697</td>
</tr>
<tr>
<td>Alternating k-triangles ($\theta_2$)</td>
<td>0.5909</td>
<td>0.0882</td>
</tr>
<tr>
<td>Seniority Main Effect ($\beta_1$)</td>
<td>0.0245</td>
<td>0.0064</td>
</tr>
<tr>
<td>Practice Main Effect ($\beta_2$)</td>
<td>0.3945</td>
<td>0.1103</td>
</tr>
<tr>
<td>Same Practice ($\beta_3$)</td>
<td>0.7721</td>
<td>0.1973</td>
</tr>
<tr>
<td>Same Gender ($\beta_4$)</td>
<td>0.7302</td>
<td>0.2495</td>
</tr>
<tr>
<td>Same Office ($\beta_5$)</td>
<td>1.1614</td>
<td>0.1952</td>
</tr>
</tbody>
</table>

⇒ Standard errors heuristically obtained via asymptotic theory

⇒ Identified factors that may increase odds of cooperation
  Ex: same practice, gender and office location double odds

⇒ Strong evidence for transitivity effects since $\hat{\theta}_2 \gg se(\hat{\theta}_2)$
  ⇒ Something beyond basic homophily explaining such effects
Assessing goodness-of-fit

- Assess goodness-of-fit to $G^{obs}$
  - Sample from fitted ERGM

- Compared distributions of
  - Degree
  - Edge-wise shared partners
  - Geodesic distance

- Plots show good fit overall

Fig. 6.8 Goodness-of-fit plots comparing original Lazega lawyer network and 100 realizations from the model in (6.43), with the parameters in Table 6.1. Comparisons are made based on the distribution of degree, edge-wise shared partners, and geodesic distance over the 100 realizations, represented by box-plots and curves showing 10-th and 90-th quantiles – both in green. Values for the Lazega network itself are shown with solid blue lines. In the distribution of geodesic distances between pairs, the rightmost box-plot is separate and corresponds to the proportion of nonreachable pairs.
Glossary

- Network graph model
- Random graph models
- Configuration model
- Matching algorithm
- Switching algorithm
- Model-based estimation
- Assessing significance
- Reference distribution
- Network motif
- Small-world network
- Decentralized search
- Watts-Strogatz model
- Time-evolving network
- Network-growth models
- Preferential attachment
- Barabási-Albert model
- Copying models
- Exponential family
- Exponential random graph models
- Configurations
- Network statistic
- Homogeneity
- Markov random graphs