Tractable and Consistent Random Graph Models

Arun G. Chandrasekhar‡ Matthew O. Jackson*

‡Stanford, Dept. of Economics

*Stanford, Dept. of Economics, Santa Fe Institute, CIFAR

September 27, 2014
Network formation models:
- Risk-sharing relationships typically function when embedded in grouped relationships?
- Do cross race/ethnicity/caste friendships typically occur in private?
- Social learning, etc.

Core motivations for models:
1. Can test economic hypotheses from theory
2. Counterfactuals/policy
• Estimable network formation models

• Link dependencies
  • More likely to link to friends of friends?
  • Do relationships need to be embedded in others to function? (e.g., need for support, cliques, etc.?)
  • Link to others who are well-connected?

• Good statistical properties (i.e., consistency)

• Random utility foundations
Typical dataset:
- Observe one draw $g$ and covariates $x = (x_1, \ldots, x_n)$

Extremes:
- Extreme 1: Full independence
  Each decision is independent $\implies \binom{n}{2}$ independent obs.
- Extreme 2: Full dependence
  e.g., Nodes 1, 2 linked $\iff$ all nodes linked.
• Triangular array of random graphs: \( \{g_n : n \in \mathbb{N} \} \) drawn

\[
g \sim P^n_{\beta_n}(\cdot)
\]

• Want to consistently estimate \( \beta^n \) as \( n \to \infty \).

• Without any restrictions on preferences and action space

\[
u_i(g, X; \beta^n)
\]

impossible to proceed.
Most models used are not consistent with one draw $g$

- ERGM literature (see Shalizi and Rinaldo ‘13); Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)
$$P_\beta(g) = \frac{\exp(\beta \cdot S(g))}{\sum_{g'} \exp(\beta \cdot S(g'))}$$

$S(g)$ a vector of sufficient statistics, usually subgraph counts
- e.g., $\#$ links, $\#$ triangles, $\#$ $k$-stars, ...
Most models used are not consistent with one draw $g$
- ERGM literature (see Shalizi and Rinaldo ‘13); Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Infeasible or difficult estimation
- ERGM literature; Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)
Most models used are not consistent with one draw $g$
- ERGM literature (see Shalizi and Rinaldo ‘13); Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Infeasible or difficult estimation
- ERGM literature; Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Techniques that apply for dense graphs
- Bickel and Chen (‘09)
- Chatterjee, Diaconis and Sly (‘11); Graham (‘14)
Unknown $\beta := (\beta_1, \ldots, \beta_n)$.

$$p_{ij} = \frac{\exp(\beta_i + \beta_j)}{1 + \exp(\beta_i + \beta_j)}.$$ 

Show with probability $1 - O(n^{-2})$,

$$\max_i |\hat{\beta}_i - \beta_i| \leq O\left(\sqrt{\log n \over n}\right).$$
Most models used are not consistent with one draw \( g \)
- ERGM literature (see Shalizi and Rinaldo ‘13); Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Infeasible or difficult estimation
- ERGM literature; Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Techniques that apply for dense graphs
- Bickel and Chen (‘09)
- Chatterjee, Diaconis and Sly (‘11); Graham (‘14)
Most models used are not consistent with one draw \( g \)
- ERGM literature (see Shalizi and Rinaldo ‘13); Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Infeasible or difficult estimation
- ERGM literature; Christakis, Fowler, Imbens and Kalyanaraman (‘10); Mele (‘13); Koenig (‘13); Badev (‘14); Sheng (‘14)

Techniques that apply for dense graphs
- Bickel and Chen (‘09)
- Chatterjee, Diaconis and Sly (‘11); Graham (‘14)

Techniques that apply for near block-diagonal graphs
Random Geometric Graph Approach
Random Geometric Graph Approach
Random Geometric Graph Approach
Random Geometric Graph Approach
Random Geometric Graph Approach

Representation of $n \times n$ adjacency matrix
Develop new models that:

- extend the workhorse model in the literature
  - emphasize subgraphs as objects of interest; resulting network is a projection

- under certain assumptions, have estimators that are consistent and asymptotically normally distributed

- are feasible to estimate

- generate graphs consistent with data

- have microfoundations
Summary of Approach
1. Dynamic (myopic) revision

2. Mutual consent

3. Strategic search
A. **Payoffs:** Additively separable in subgraphs

\[ u_i(g, X; \beta) = \sum_{g_\ell, i \in g_\ell} v(g_\ell, X_\ell; \beta_\ell). \]
A. Payoffs: Additively separable in subgraphs

\[ u_i(g, X; \beta) = \sum_{g_\ell, i \in g_\ell} v(g_\ell, X_\ell; \beta_\ell). \]

B. Stability concept: Pairwise stability with transfers

- \( ij \in g \) implies that \( u_i(g) + u_j(g) \geq u_i(g - ij) + u_j(g - ij) \), and
- \( ij \notin g \) implies that \( u_i(g) + u_j(g) \geq u_i(g + ij) + u_j(g + ij) \).
A. **Payoffs:** Additively separable in subgraphs

\[ u_i(g, X; \beta) = \sum_{g_\ell, i \in g_\ell} v(g_\ell, X_\ell; \beta_\ell). \]

B. **Stability concept:** Pairwise stability with transfers

- \( ij \in g \) implies that \( u_i(g) + u_j(g) \geq u_i(g - ij) + u_j(g - ij) \), and
- \( ij \not\in g \) implies that \( u_i(g) + u_j(g) \geq u_i(g + ij) + u_j(g + ij) \).

**Remark:** \( A+B \implies \) existence of a potential function: \( f(\cdot) \).
**Microfoundation Examples**

**Dynamic Revision**

---

**A. Payoffs:** Additively separable in subgraphs

\[ u_i(g, X; \beta) = \sum_{g_\ell, i \in g_\ell} v(g_\ell, X_\ell; \beta_\ell). \]

**B. Stability concept:** Pairwise stability with transfers

- \( ij \in g \) implies that \( u_i(g) + u_j(g) \geq u_i(g - ij) + u_j(g - ij) \), and
- \( ij \notin g \) implies that \( u_i(g) + u_j(g) \geq u_i(g + ij) + u_j(g + ij) \).

*Remark:* A+B \( \implies \) existence of a potential function: \( f(\cdot) \).

**C. Meeting process and shocks:** \( \forall t \), at most one pair meets; every \( ij \) have some positive probability of being selected.
A. Payoffs: Additively separable in subgraphs

\[ u_i(g, X; \beta) = \sum_{g_\ell, i \in g_\ell} v(g_\ell, X_\ell; \beta_\ell). \]

B. Stability concept: Pairwise stability with transfers

- \( ij \in g \) implies that \( u_i(g) + u_j(g) \geq u_i(g - ij) + u_j(g - ij) \), and
- \( ij \notin g \) implies that \( u_i(g) + u_j(g) \geq u_i(g + ij) + u_j(g + ij) \).

Remark: A+B \implies existence of a potential function: \( f(\cdot) \).

C. Meeting process and shocks: \( \forall t \), at most one pair meets; every \( ij \) have some positive probability of being selected.

Result [Butts ‘09, Mele ‘13, This paper]: Invariant distribution is

\[ g \sim P_\beta(g) := \frac{\exp(f(g; \beta))}{\sum_{g'} \exp(f(g'; \beta))} \]
A. **Payoffs:** Payoffs by subgraph

\[ u_i(g_\ell, X_\ell; \beta_\ell) = \beta_\ell h_i(g_\ell, X_\ell) - \epsilon_{i,\ell} \]
A. Payoffs: Payoffs by subgraph

\[ u_i(g_\ell, X_\ell; \beta_\ell) = \beta_\ell h_i(g_\ell, X_\ell) - \epsilon_{i,\ell} \]

B. Stability concept: Mutual consent

\[ g_\ell \iff u_i(g_\ell, X_\ell; \beta) \geq^* 0, \forall i \in g_\ell. \]
A. **Payoffs:** Payoffs by subgraph

\[ u_i(g_\ell, X_\ell; \beta_\ell) = \beta_\ell h_i(g_\ell, X_\ell) - \epsilon_{i,\ell} \]

B. **Stability concept:** Mutual consent

\[ g_\ell \iff u_i(g_\ell, X_\ell; \beta_\ell) \geq *0, \ \forall i \in g_\ell. \]

C. **Meeting process and shocks:** Subgroups of size \( m_\ell \) with with probability \( \pi_\ell \) and shocks \( \epsilon_{i,\ell} \sim F_\ell(\cdot) \):

\[ p_\ell(X_\ell; \beta_\ell) = \pi_\ell \cdot \prod_{i \in g_\ell} F_\ell(\beta_\ell h_i). \]
A. Payoffs: Payoffs by subgraph

\[ u_i(g_\ell, X_\ell; \beta_\ell) = \beta_\ell h_i(g_\ell, X_\ell) - \epsilon_{i,\ell} \]

B. Stability concept: Mutual consent

\[ g_\ell \iff u_i(g_\ell, X_\ell; \beta) \geq 0, \forall i \in g_\ell. \]

C. Meeting process and shocks: Subgroups of size \( m_\ell \) with probability \( \pi_\ell \) and shocks \( \epsilon_{i,\ell} \sim F_\ell(\cdot) \):

\[ p_\ell(X_\ell; \beta) = \pi_\ell \cdot \prod_{i \in g_\ell} F_\ell(\beta_\ell h). \]

Result [This paper]: Under some assumptions
- \( \beta \) consistently estimable
- distribution is \( g \sim P_\beta(g) := \frac{\exp(f(g; \beta))}{\sum_{g'} \exp(f(g'; \beta))} \)
Agents put in (costly) search effort to form cliques for a payoff

Whether a given clique forms is stochastic, depending on efforts of all involved

Resulting graph reflects subgraphs (cliques) of various sizes occurring with various frequencies
(I) **Simple example**

(II) Subgraph Generated Models

(III) Illustrations
Workhorse: **Exponential Random Graph Models**

**Example:** studied extensively by Strauss 86, Wasserman and Pattison 96, Park and Newman 04, Butts 06, Chatterjee and Diaconis 11, Mele 14,...

- e.g., probability of a graph depends on
  - how many links present
  - how many triangles present - e.g., Coleman’s closure, support in Jackson et al., etc.
Want probability of graph $g$ to depend on

$$\theta_I \#\text{isolates}(g) + \theta_L \#\text{links}(g) + \theta_T \#\text{triangles}(g).$$
Want probability of graph $g$ to depend on

$$\theta_I \#\text{isolates}(g) + \theta_L \#\text{links}(g) + \theta_T \#\text{triangles}(g).$$

Set

$$P_{\theta}(g) \propto \exp \left( \theta_I \#\text{isolates}(g) + \theta_L \#\text{links}(g) + \theta_T \#\text{triangles}(g) \right)$$
More generally

\[ P_\theta(g) = \frac{\exp(\theta_1 S_1(g) + \ldots + \theta_k S_k(g))}{\sum_{g' \in G^n} \exp(\theta_1 S_1(g') + \ldots + \theta_k S_k(g'))} \]

where:

- \( g \) is a graph
- \( S_l(g) \) is a “statistic” - often the number of copies of some subgraph within the parent graph
More generally

\[
P_\theta(g) = \frac{\exp(\theta_1 S_1(g) + \ldots + \theta_k S_k(g))}{\sum_{g' \in G^n} \exp(\theta_1 S_1(g') + \ldots + \theta_k S_k(g'))}
\]

where:

- \( g \) is a graph
- \( S_l(g) \) is a “statistic” - often the number of copies of some subgraph within the parent graph

MCMC techniques for estimation have led to these becoming standard (Snijders 02, Handcock 03, ...)

ERGMs
Bhamidi, Bresler and Sly (2008), Chatterjee and Diaconis (2011)

For “large classes” (dense enough) ERGMs...

- MCMC: (Gibbs sampling)
  Mixing time is less than exponential \textit{only if} networks have asymptotically approximately independent links (ER-like)

- So when ERGMs are interesting, then MCMC estimation will be exponentially slow

- Simulation evidence of instability even in the sparse case...
Subgroups of nodes sometimes meet
Make structures at random (e.g., cliques)
Here: triangles, links, isolates
Example: Isolates, Links and Triangles

$$P_\theta (g) = \frac{\exp (\theta I S_I (g) + \theta L S_L (g) + \theta T S_T (g))}{\sum_{g'} \exp (\theta I S_I (g') + \theta L S_L (g') + \theta T S_T (g'))}.$$
Example: Isolates, Links and Triangles

$$P_{\theta}(g) = \frac{\exp(\theta_I S_I(g) + \theta_L S_L(g) + \theta_T S_T(g))}{\sum_{g'} \exp(\theta_I S_I(g') + \theta_L S_L(g') + \theta_T S_T(g'))}.$$  

$S_T(g)$: number of triangles.  
$S_L(g)$: number of edges.  
$S_I(g)$: number of isolated nodes.
Example: Isolates, Links and Triangles

\[ P_\theta (g) = \frac{\exp (\theta_I S_I (g) + \theta_L S_L (g) + \theta_T S_T (g))}{\sum_{g'} \exp (\theta_I S_I (g') + \theta_L S_L (g') + \theta_T S_T (g'))}. \]

- \( S_T(g) \): number of triangles.
- \( S_L(g) \): number of edges.
- \( S_I(g) \): number of isolated nodes.

- \( n = 50 \) nodes
Example: Isolates, Links and Triangles

$P_\theta (g) = \frac{\exp (\theta I S_I (g) + \theta L S_L (g) + \theta T S_T (g))}{\sum_{g'} \exp (\theta I S_I (g') + \theta L S_L (g') + \theta T S_T (g'))}.$

$S_T(g)$: number of triangles.
$S_L(g)$: number of edges.
$S_I(g)$: number of isolated nodes.

- $n = 50$ nodes
- 20 isolates, 45 links, 10 triangles on avg.
**Example: Isolates, Links and Triangles**

\[
P_\theta (g) = \frac{\exp (\theta_I S_I (g) + \theta_L S_L (g) + \theta_T S_T (g))}{\sum_{g'} \exp (\theta_I S_I (g') + \theta_L S_L (g') + \theta_T S_T (g'))}.
\]

- \( S_T(g) \): number of triangles.
- \( S_L(g) \): number of edges.
- \( S_I(g) \): number of isolated nodes.

- \( n = 50 \) nodes

- 20 isolates, 45 links, 10 triangles on avg.

- Corresponds to

\[
p_I = 0.33, \ p_L = 0.042, \ p_T = 0.0014
\]
Isolates: $\hat{\theta}_I$
Isolates: usually not awful (sometimes)
Links: usually too dense
ERGM Parameter Estimate
TRIANGLES: TOO MUCH CLOSURE OR EMPTY
Conclusions:

- Even fixed value of $S$ leads to unstable estimates of $\theta$,
- with unreasonable chains of graphs being drawn,
- and unreasonable inference suggested.
Conclusions:

- Even fixed value of $S$ leads to unstable estimates of $\theta$,
- with unreasonable chains of graphs being drawn,
- and unreasonable inference suggested.

But the model was so easy to describe...
Many $g$’s but fewer possible statistics (values for $S$)

Many networks lead to the same statistics
  - Probabilities depend on statistics
  - Networks with same statistics are equally likely

Can collapse all equivalent networks to sufficient statistics
Exponential random graph model

\[ P_{\beta}(g) = \frac{\exp (\beta \cdot S(g))}{\sum_{g' \in G_n} \exp (\beta \cdot S(g'))} \]

with \( N(s) = |\{g : S(g) = s\}| \).
Rewrite denominator as sum over equivalence classes

\[ P_\beta(g) = \frac{\exp(\beta \cdot S(g))}{\sum_{s' \in \mathcal{S}_n} N_S(s') \exp(\beta \cdot s')} \]

with \( N(s) = |\{g : S(g) = s\}|. \)
Rewrite denominator as sum over equivalence classes

\[
P_\beta(g) = \frac{\exp(\beta \cdot S(g))}{\sum_{s' \in S^n} N_{S}(s') \exp(\beta \cdot s')}
\]

with \(N(s) = |\{g : S(g) = s\}|.\)
Sum LHS and numerator over equivalence classes

$$P_\beta(s) = \frac{N_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} N_S(s') \exp(\beta \cdot s')}$$

with $N(s) = |\{g : S(g) = s\}|$. 
Sum LHS and numerator over equivalence classes

\[ P_\beta(s) = \frac{N_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} N_S(s') \exp(\beta \cdot s')} \]

with \( N(s) = |\{g : S(g) = s\}| \).
SERGMs

$P_\beta(s) = \frac{N_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in \mathcal{S}^n} N_S(s') \exp(\beta \cdot s')}$

Nothing special about $N_S(s)$ - just the counting measure.

- But typically impossible to compute...
\[
\begin{align*}
P_\beta(s) &= \frac{N_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} N_S(s') \exp(\beta \cdot s')} \\
\text{Nothing special about } N_S(s) - \text{just the counting measure.} \\
&\quad \bullet \text{ But typically impossible to compute...} \\
\end{align*}
\]

\[
\begin{align*}
P_\beta(s) &= \frac{K_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} K_S(s') \exp(\beta \cdot s')} \\
\text{for other } K_S(\cdot) \neq N_S(\cdot).
\end{align*}
\]
\[ P_\beta(s) = \frac{K_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} K_S(s') \exp(\beta \cdot s')} \]
\[ P_\beta(s) = \frac{K_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in S^n} K_S(s') \exp(\beta \cdot s')} \]

- Natural micro-foundations for other \( K(\cdot) \neq N(\cdot) \).
- \( K(\cdot) \neq N(\cdot) \) may recover desirable statistical properties.
\[ P_\beta(s) = \frac{K_S(s) \cdot \exp(\beta \cdot s)}{\sum_{s' \in A^n} K_S(s') \exp(\beta \cdot s')} \]

\( S \) can encode many things:

- Links, cliques, \( k \)-stars, other subgraphs, friends in common per link, multigraphs, etc.
- Can do preference based models
- Allows for node characteristics
(I) Simple example
(II) Subgraph Generated Models
(III) Illustrations
What do many economic networks look like?

(1) Interdependency between links:

- more than random (conditional on covariates) chance of one’s friends being friends themselves

(II) Mostly empty.
Subgraph Generation Models

- Subnetworks generated, network is a by-product/projection

- People form links, triangles; some folk are anti-social (isolates), etc...
- Nature forms subgraphs of each type w/ some probability
- May intersect, may overlap
- Observe resulting graph, infer probabilities
Triangles Form
Desired Network
subgraphs: $G = (G_1, \ldots, G_k)$
\textbf{subgraphs:} \( G = (G_1, \ldots, G_k) \)

\textbf{def. \textit{nicely ordered}:}

- \( G_\ell \not\subseteq G_{\ell'}, \ell' > \ell \)
- \( g'_\ell \in G'_{\ell'} \iff g'_\ell \not\subseteq g_\ell, g_\ell \in G_\ell, \ell' > \ell \)
**Model**

- **subgraphs:** $G = (G_1, ..., G_k)$

- **def. nicely ordered:**
  - $G_\ell \not\subset G_{\ell'}$, $\ell' > \ell$
  - $g'_\ell \in G'_{\ell} \iff g'_\ell \not\subset g_\ell$, $g_\ell \in G_\ell$, $\ell' > \ell$
  - e.g., $G_1 = \text{triangles}$ and $G_2 = \{\text{links} \not\subset \text{triangles}\}$
- **subgraphs**: $G = (G_1, ..., G_k)$

- **def. nicely ordered**:
  - $G_\ell \not\subset G_{\ell'}$, $\ell' > \ell$
  - $g'_{\ell} \in G'_{\ell} \iff g'_{\ell} \not\subset g_{\ell}$, $g_{\ell} \in G_{\ell}$, $\ell' > \ell$
  - e.g., $G_1 = \text{triangles}$ and $G_2 = \{\text{links} \not\subset \text{triangles}\}$

- **parameters**: $p^n = (p^n_1, ..., p^n_k)$
  - or w/ covariates $p^n(x; \gamma) = (p^n_\ell(x_{\ell}; \gamma_{\ell}))_{\ell=1}^k$
subgraphs: $G = (G_1, \ldots, G_k)$

def. nicely ordered:

- $G_\ell \not\subset G_{\ell'}$, $\ell' > \ell$
- $g'_{\ell'} \in G'_{\ell'} \iff g'_{\ell'} \not\subset g_{\ell}$, $g_{\ell} \in G_{\ell}$, $\ell' > \ell$
- e.g., $G_1 = \text{triangles}$ and $G_2 = \{\text{links} \not\subset \text{triangles}\}$

parameters: $p^n = (p^n_1, \ldots, p^n_k)$

- or w/ covariates $p^n(x; \gamma) = (p^n_\ell(x_\ell; \gamma_\ell))_{\ell=1}^k$

formation:

- Type 1 subgraphs form with probability $p^n_1$.
- Type 2 subgraphs form with probability $p^n_2$.
- ...
- Type $\ell$ subgraphs form with probability $p^n_\ell$.
- ...
Problem: Incidental Generation

- **def. generating class:** for every $G_\ell$ there are classes $J$ of subgraphs that can generate $g_\ell \in G_\ell$.
  - e.g., \{Triangles, Links\}. Triangles have 4 generating classes: $(T, T, T)$; $(T, L, L)$; $(T, T, L)$; $(L, L, L)$.

- **def. relatively sparse:** a sequence of models w/ nicely ordered subgraphs and parameters is relatively sparse if the expected share of incidentally generated subgraphs to truly generated subgraph vanishes

\[
\prod_{j \in J} \mathbb{E}_{p^n}[S_{\ell_j}(g)] \\
\frac{n^{M_J} \mathbb{E}_{p^n}[S_{\ell}(g)]}{n^M \mathbb{E}_{p^n}[S_{\ell}(g)]} \rightarrow 0
\]
Relative Sparsity Reasonable?

Example \{\text{Triangles, Links}\}

- \((T,T,T)\):

\[
\frac{n^3 p_T^3}{p_T} = p_T^2 n^3 \to 0 \implies p_T = o(n^{-3/2})
\]

Remark 1: a given node can be in \((\binom{n}{2})p_T = o(\sqrt{n})\) triangles.
Example \{\text{Triangles, Links}\}

- \( (T,T,T)\):
  \[
  \frac{n^3 p_T^3}{p_T} = p_T^2 n^3 \to 0 \implies p_T = o\left(n^{-3/2}\right)
  \]

Remark 1: a given node can be in \( \binom{n}{2} p_T = o(\sqrt{n}) \) triangles.

- \( (L,L,L)\):
  \[
  \frac{p_L^3}{p_T} \to 0 \implies p_L = o\left(n^{-1/2}\right)
  \]

Remark 2: a given node can be in \( (n-1)p_L = o(\sqrt{n}) \) links.
Example: Leider, Mobius, Rosenblatt, Do (‘09)

- Harvard social network: $n \approx 4000$
- Avg number of links per person: 7
- Avg % of one’s friends that are friends themselves: 8%

Observe that $p_T = 3/n^2$, $p_L = 2/n$ yields

- Avg number of links per person: 8
- Avg % of one’s friends that are friends themselves: 10.71%
**Theorem:** Consider a nicely ordered and relatively sparse SUGM with statistics on $m_1, \ldots, m_k$ nodes.

Letting $D_n = \text{diag} \{ p_{\ell \ell} n^{m_{\ell}} \}_{\ell=1}^k$:

$$D_n^{1/2} \left( (\hat{p}_1^n, \ldots, \hat{p}_k^n)' - (p_1^n, \ldots, p_k^n)' \right) \rightsquigarrow \mathcal{N} \left( 0, I_k \sqrt{2} \right).$$
SUGMs behave well

**Theorem:** Consider a nicely ordered and relatively sparse SUGM with statistics on $m_1, \ldots, m_k$ nodes.

Letting $D_n = \text{diag} \{ p_{\ell} n^{m_{\ell}} \}_{\ell=1}^k$:

$$D_n^{1/2} \left( (\hat{p}_1^n, \ldots, \hat{p}_k^n)' - (p_1^n, \ldots, p_k^n)' \right) \rightsquigarrow \mathcal{N} \left( 0, I_k \sqrt{2} \right).$$

**Covariates:**

- Trivial to add discrete covariates.
  - Subgraphs tagged by covariates. [curse of dimensionality]

- Straightforward to add continuous covariates: $p_{\ell}^n(x_{\ell}; \gamma_{\ell}).$
  - cf. Hjort and Pollard (‘93), sub-$\sqrt{n m_{\ell}}$ rates due to sparsity.
Q. Can we view SUGMs as SERGMs?
**Q.** Can we view SUGMs as SERGMs?

**Ans.** Yes, and it motivates specific $K(\cdot)$’s.
Q. Can we view SUGMs as SERGMs?

Ans. Yes, and it motivates specific $K(\cdot)$’s.

**Theorem:** Suppose that the probability that subgraph of type $\ell$ forms is given by $p_{\ell} = \frac{\exp \beta_{\ell}}{1+\exp \beta_{\ell}}$. This form of SUGM can be represented in a SERGM form:

$$P^n_{\beta}(\tilde{S}) = \frac{K^n(\tilde{S}) \exp (\tilde{S} \cdot \beta)}{\sum_{s'} K^n(s') \exp (s' \cdot \beta)},$$

where $K^n_{\ell}(s_{\ell}) = \binom{S_{\ell}}{s_{\ell}}$ and $K^n(s) = \prod_{\ell} K^n_{\ell}(s_{\ell})$. 
**Example:** Researcher has graph \( g \) and covariates \( X_i \) for all \( i \)
Example: Researcher has graph $g$ and covariates $X_i$ for all $i$

1. Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)
Example: Researcher has graph $g$ and covariates $X_i$ for all $i$

(I) Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)

(II) Run a regression on all* 4-tuples of whether it is a 4clique, conditional on covariates

$$P(ijkl \in K_4 | X_{ijkl}) = \Lambda(X'_{ijkl} \gamma_4) \implies \hat{\gamma}_4 \xrightarrow{P} \gamma_4.$$
**Example:** Researcher has graph $g$ and covariates $X_i$ for all $i$

(I) Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)

(II) Run a regression on all* 4-tuples of whether it is a 4clique, conditional on covariates

$$P(ijkl \in K_4 | X_{ijkl}) = \Lambda(X_{ijkl}^\gamma_4) \implies \hat{\gamma}_4 \xrightarrow{P} \gamma_4.$$ 

(III) Remove links used in (ii).

[Note: The asterisk (*) indicates that only a subset of all possible 4-tuples is considered, but the exact condition is not specified.]
**Example:** Researcher has graph $g$ and covariates $X_i$ for all $i$

(i) Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)

(ii) Run a regression on all* 4-tuples of whether it is a 4clique, conditional on covariates

\[ P(ijkl \in K_4 | X_{ijkl}) = \Lambda(X_{ijkl}' \gamma_4) \implies \hat{\gamma}_4 \xrightarrow{P} \gamma_4. \]

(iii) Remove links used in (ii).

(iv) Run a regression on all* triples of whether it is a triangle, conditional on covariates.

\[ P(ijk \in K_3 | X_{ijk}) = \Lambda(X_{ijk}' \gamma_3) \implies \hat{\gamma}_3 \xrightarrow{P} \gamma_3. \]
**Example:** Researcher has graph $g$ and covariates $X_i$ for all $i$

(1) Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)

(2) Run a regression on all* 4-tuples of whether it is a 4clique, conditional on covariates

$$P(ijkl \in K_4 | X_{ijkl}) = \Lambda(X_{ijkl} \gamma_4) \implies \hat{\gamma}_4 \xrightarrow{P} \gamma_4.$$

(III) Remove links used in (ii).

(IV) Run a regression on all* triples of whether it is a triangle, conditional on covariates.

$$P(ijk \in K_3 | X_{ijk}) = \Lambda(X_{ijk} \gamma_3) \implies \hat{\gamma}_3 \xrightarrow{P} \gamma_3.$$

(V) Remove links used in (iv).
Example: Researcher has graph \( g \) and covariates \( X_i \) for all \( i \)

(I) Pick a nicely ordered set of statistics (e.g., 4cliques, triangles, and links)

(II) Run a regression on all* 4-tuples of whether it is a 4clique, conditional on covariates

\[
P(ijkl \in \mathcal{K}_4 | X_{ijkl}) = \Lambda(X'_{ijkl} \gamma_4) \implies \hat{\gamma}_4 \xrightarrow{P} \gamma_4.
\]

(III) Remove links used in (ii).

(IV) Run a regression on all* triples of whether it is a triangle, conditional on covariates.

\[
P(ijk \in \mathcal{K}_3 | X_{ijk}) = \Lambda(X'_{ijk} \gamma_3) \implies \hat{\gamma}_3 \xrightarrow{P} \gamma_3.
\]

(V) Remove links used in (iv).

(VI) Run a regression on all* pairs of whether it is an edge, conditional on covariates.

\[
P(ij \in \mathcal{K}_2 | X_{ij}) = \Lambda(X'_{ij} \gamma_2) \implies \hat{\gamma}_2 \xrightarrow{P} \gamma_2.
\]
(I) Simple example
(II) Subgraph Generated Models
(III) Illustrations
1. Why need SUGMs/SERGMs and not simply dyadic with rich covariates?

2. Social pressure and caste example
NEED FOR SUGMs/SERGMs

- Examine data from 75 Indian villages from Banerjee, Chandrasekhar, Duflo, Jackson ‘13

- Estimate a model and then use it to generate networks

- How well do the model-recreated networks match real networks on non-modeled characteristics
Need for SUGMs/SERGMs

- **Step 1: Estimate models**
  - Dyadic: estimate $p_L(pL)$ (covariates)
    - Geographic distance, caste, amenities (housing material, roofing material, number of beds), etc.
  - SUGM: estimate $p_L^{close}, p_L^{far}, p_T^{close}, p_T^{far}$

- **Step 2: Randomly generate networks**
  - Dyadic: generate links given $p_L(pL)$ (covariates)
  - SUGM: generate subgraphs w/ $p_L^{close}, p_L^{far}, p_T^{close}, p_T^{far}$
## Recreate Networks

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Unsupported Links</td>
<td>160.8</td>
<td>236.2</td>
<td>161.2</td>
<td>161.8</td>
</tr>
<tr>
<td>Number of Triangles</td>
<td>39.2</td>
<td>3.1</td>
<td>39.7</td>
<td>39.5</td>
</tr>
<tr>
<td>Average Degree</td>
<td>2.3243</td>
<td>2.3260</td>
<td>2.5916</td>
<td>2.5219</td>
</tr>
<tr>
<td>Number of Isolates</td>
<td>54.9722</td>
<td>25.7222</td>
<td>31.4444</td>
<td>65.9167</td>
</tr>
<tr>
<td>Average Clustering</td>
<td>0.0895</td>
<td>0.0105</td>
<td>0.1268</td>
<td>0.0829</td>
</tr>
<tr>
<td>Fraction in Giant Component</td>
<td>0.7061</td>
<td>0.8315</td>
<td>0.7982</td>
<td>0.6718</td>
</tr>
<tr>
<td>First Eigenvalue</td>
<td>5.5446</td>
<td>3.8578</td>
<td>4.6762</td>
<td>5.3025</td>
</tr>
<tr>
<td>Spectral Gap</td>
<td>0.9550</td>
<td>0.3354</td>
<td>0.6684</td>
<td>1.0617</td>
</tr>
<tr>
<td>Second Eigenvalue of Stochastized</td>
<td>0.9573</td>
<td>0.9632</td>
<td>0.9559</td>
<td>0.9069</td>
</tr>
<tr>
<td>Average Path Length</td>
<td>4.6921</td>
<td>5.6565</td>
<td>5.1215</td>
<td>4.1180</td>
</tr>
</tbody>
</table>
Cross-caste relationships more likely to occur:
- “in private” with no friends in common?
- or with the same frequency in embedded relationships?

Relatively less likely?
Need more consent to form triad than dyad.

Need to account for these preferences, otherwise mechanically find

\[
\frac{\text{less desired dyad}}{\text{more desired dyad}} > \frac{\text{less desired triad}}{\text{more desired triad}}.
\]
preferences:

\[ u_i(j) = \alpha_L + \beta_L 1\{\text{caste}(i) \neq \text{caste}(j)\} - \epsilon_{ij,L} \]
\[ u_i(jk) = \alpha_T + \beta_T (1 - 1\{\text{caste}(i) = \text{caste}(j) = \text{caste}(k)\}) - \epsilon_{ijk,T} \]

frequencies: \( q_L(x), \ q_T(x) \ x \in \{\text{Same}(1), \text{Diff}(0)\} \)

question:

\[ \frac{q_L(0)}{q_L(1)} > \frac{q_T(0)}{q_T(1)} \]
question: \[ \frac{q_L(0)}{q_L(1)} > \frac{q_T(0)}{q_T(1)} \]

issues:
- heterogenous meeting probabilities: \( \pi_L(x), \pi_T(x) \)
- a link requires 2 consents; a triangle 3 consents

we observe \( g \) and therefore:
\[ p_L(x) = \pi_L(x) \cdot q_L(x)^2 \quad \text{and} \quad p_T(x) = \pi_T(x) \cdot q_T(x)^3. \]
we observe $g$ and therefore:

$$p_L(x) = \pi_L(x) \cdot q_L(x)^2$$

and

$$p_T(x) = \pi_T(x) \cdot q_T(x)^3.$$ 

**assume:** people spend $f > 1/2$ share of time mixing with own caste; $1 - f$ with other caste
we observe \( g \) and therefore:

\[
p_L(x) = \pi_L(x) \cdot q_L(x)^2 \quad \text{and} \quad p_T(x) = \pi_T(x) \cdot q_T(x)^3.
\]

**assume:** people spend \( f > 1/2 \) share of time mixing with own caste; \( 1 - f \) with other caste, implying

\[
\frac{\pi_T(0)}{\pi_T(1)} > \frac{\pi_L(0)}{\pi_L(1)}.
\]
we observe \( g \) and therefore:

\[
p_L(x) = \pi_L(x) \cdot q_L(x)^2 \quad \text{and} \quad p_T(x) = \pi_T(x) \cdot q_T(x)^3.
\]

**assume:** people spend \( f > 1/2 \) share of time mixing with own caste; \( 1 - f \) with other caste, implying

\[
\frac{\pi_T(0)}{\pi_T(1)} > \frac{\pi_L(0)}{\pi_L(1)}.
\]

Then, a sufficient condition for \( \frac{q_L(0)}{q_L(1)} > \frac{q_T(0)}{q_T(1)} \) is

\[
\frac{p_T(0)}{p_T(1)} < \left( \frac{p_L(0)}{p_L(1)} \right)^{3/2}.
\]
\[ \frac{p_T(\text{different})}{p_T(\text{same})} \]

\[ \left( \frac{p_L(\text{different})}{p_L(\text{same})} \right)^{3/2} \]
Below Median
Above Median

\[(p_{L \text{ (different)}/p_{L \text{ (same)}})^{3/2}\]

\[(p_{T \text{ (different)/p_{T \text{ (same)}})}\]
People show significantly stronger preference for cross-caste relationships when links are in isolation.

Suggestive that it is *especially* true with more symmetrically sized castes.
ERGMs face estimation challenges
- Computational difficulties
- Poor asymptotic properties

Work with subgraphs directly to simplify
- Consistency, fast estimation
- Natural (though limited) microfoundations
- Reflects patterns in empirical data

Applications
- Recreate networks, how relationship types are sensitive to context, easy to extend to multigraphs