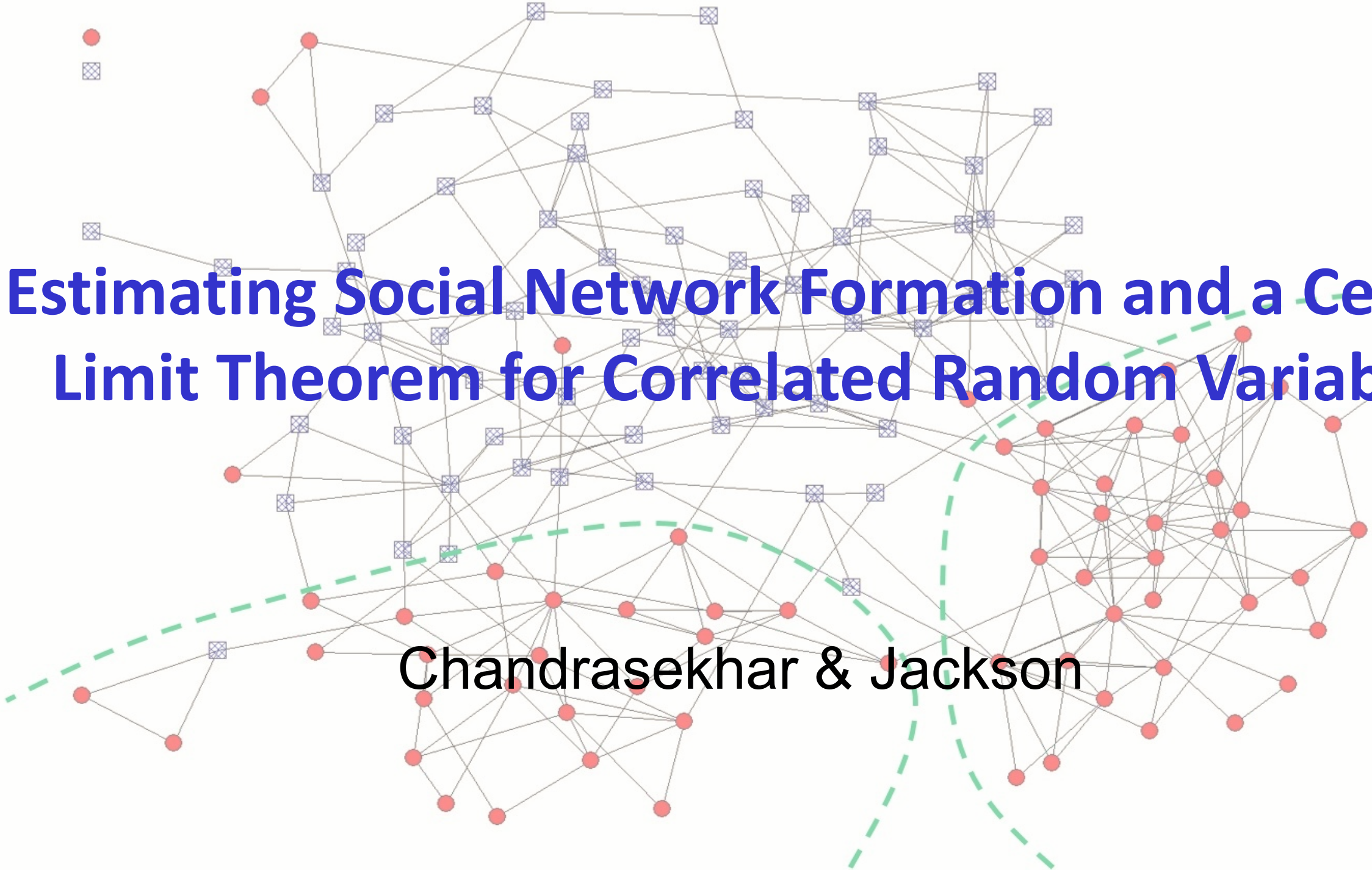
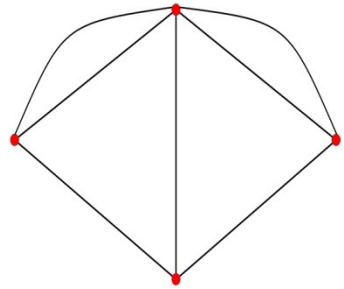


Estimating Social Network Formation and a Central Limit Theorem for Correlated Random Variables

Chandrasekhar & Jackson

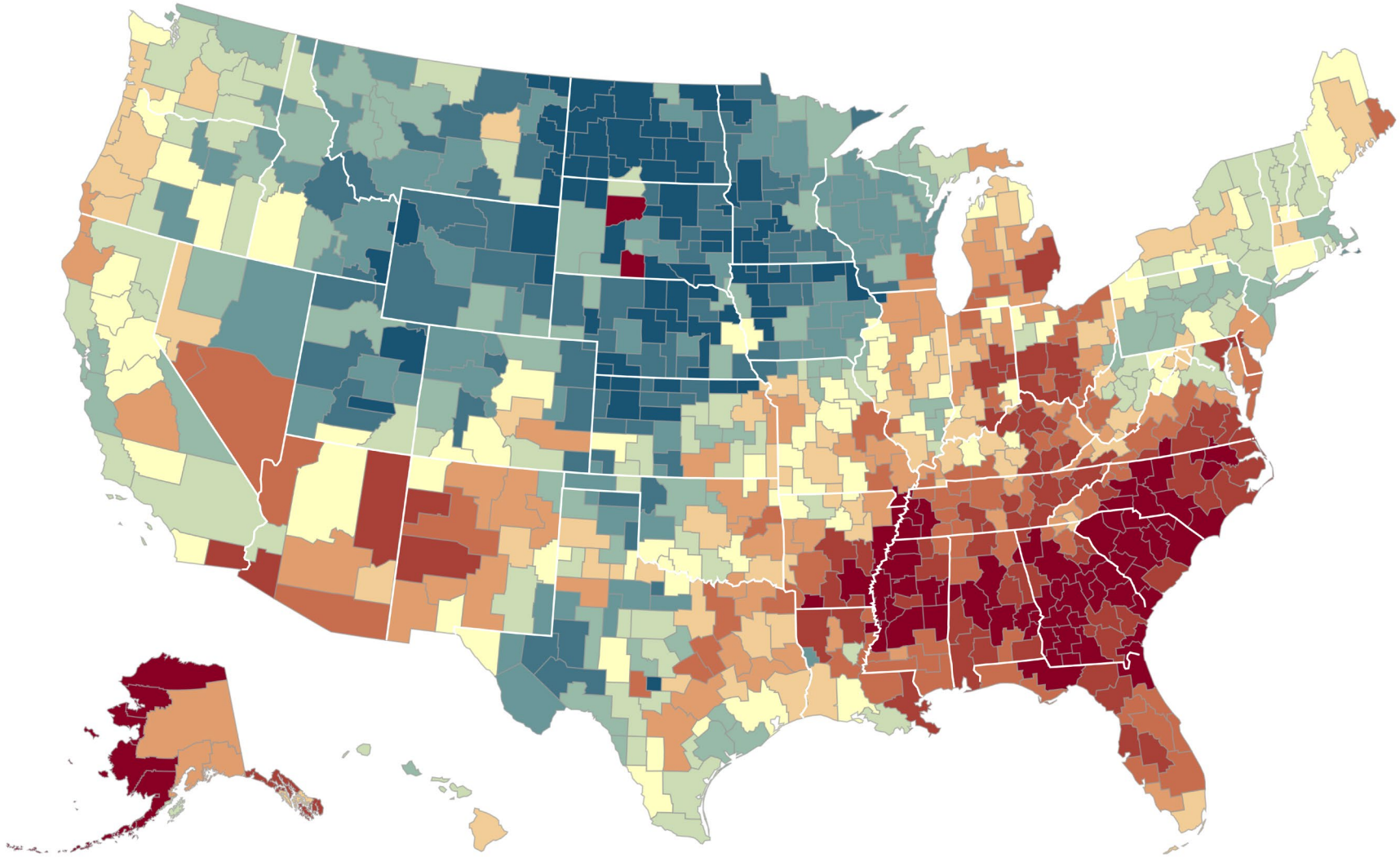
Social Networks in Action:

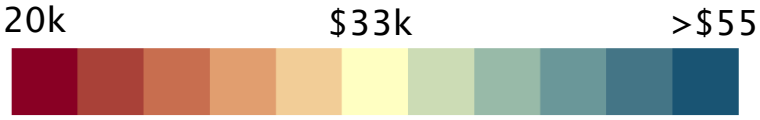


- Labor/Education: job referrals, peer effects, poverty traps
- Development: social learning, diffusion, norms
- Organizations: learning, team efficiency, culture
- Politics: voting, alliances, conflict, polarization
- Trade and Macro: shock propagation, innovation
- Finance: contagion, intermediation, regulation
- Public: corruption, crime

The Geography of Upward Mobility in the United States

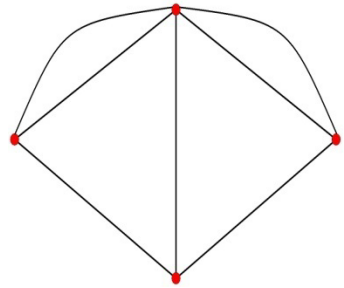
Average Income at Age 35 for Children whose Parents Earned $27,000 (25th percentile)





Note: Blue = More Upward Mobility, Red = Less Upward Mobility
Source: Chetty, Friedman, Hendren, Jones, Porter 2018

Network Explanations?



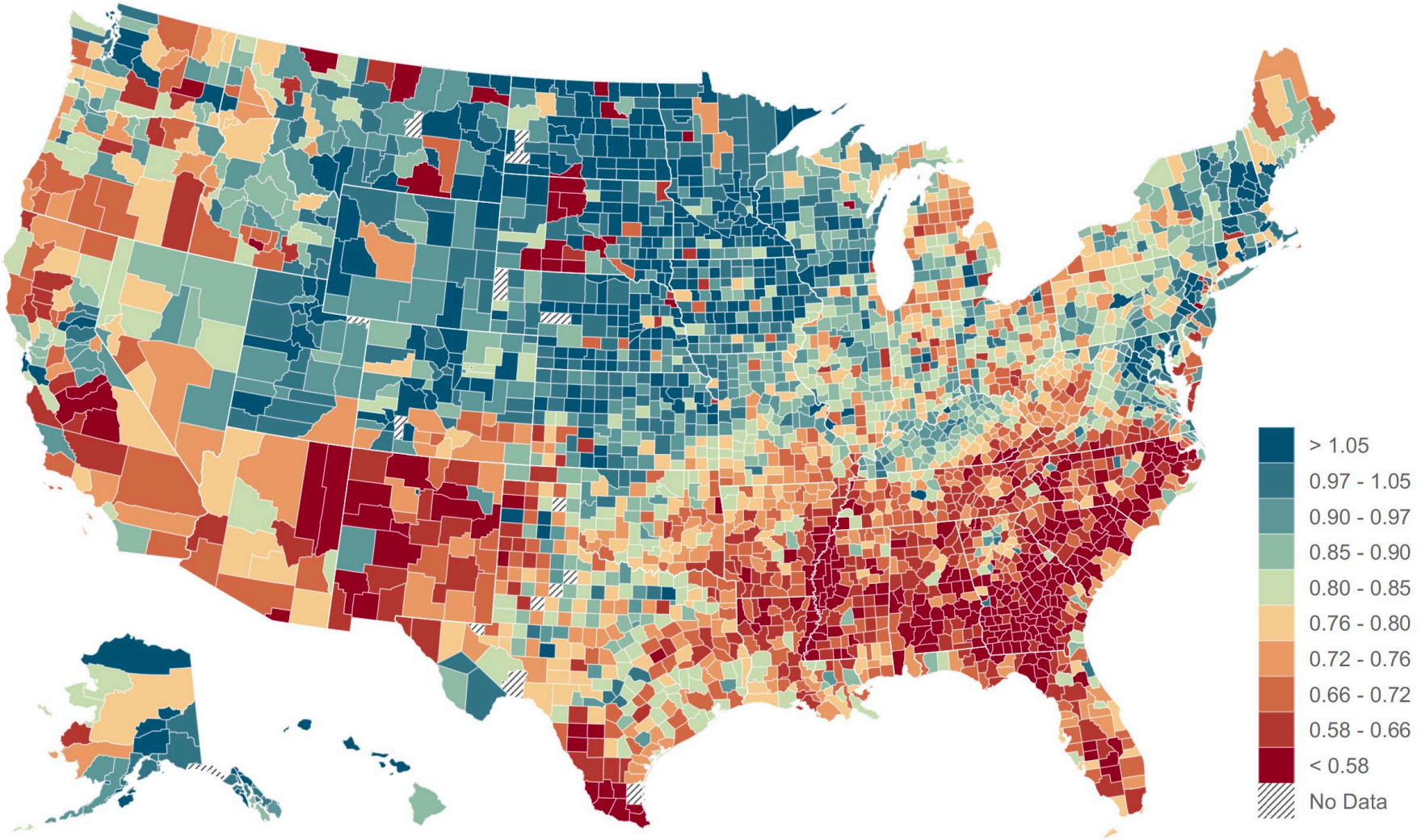
- Examined various measures of social capital

- Economic connectedness - form of economic homophily

  - Examine connections by income:
    - What fraction of below median income person's friends have above median income?
    - Divide by .5, so no homophily would be 1

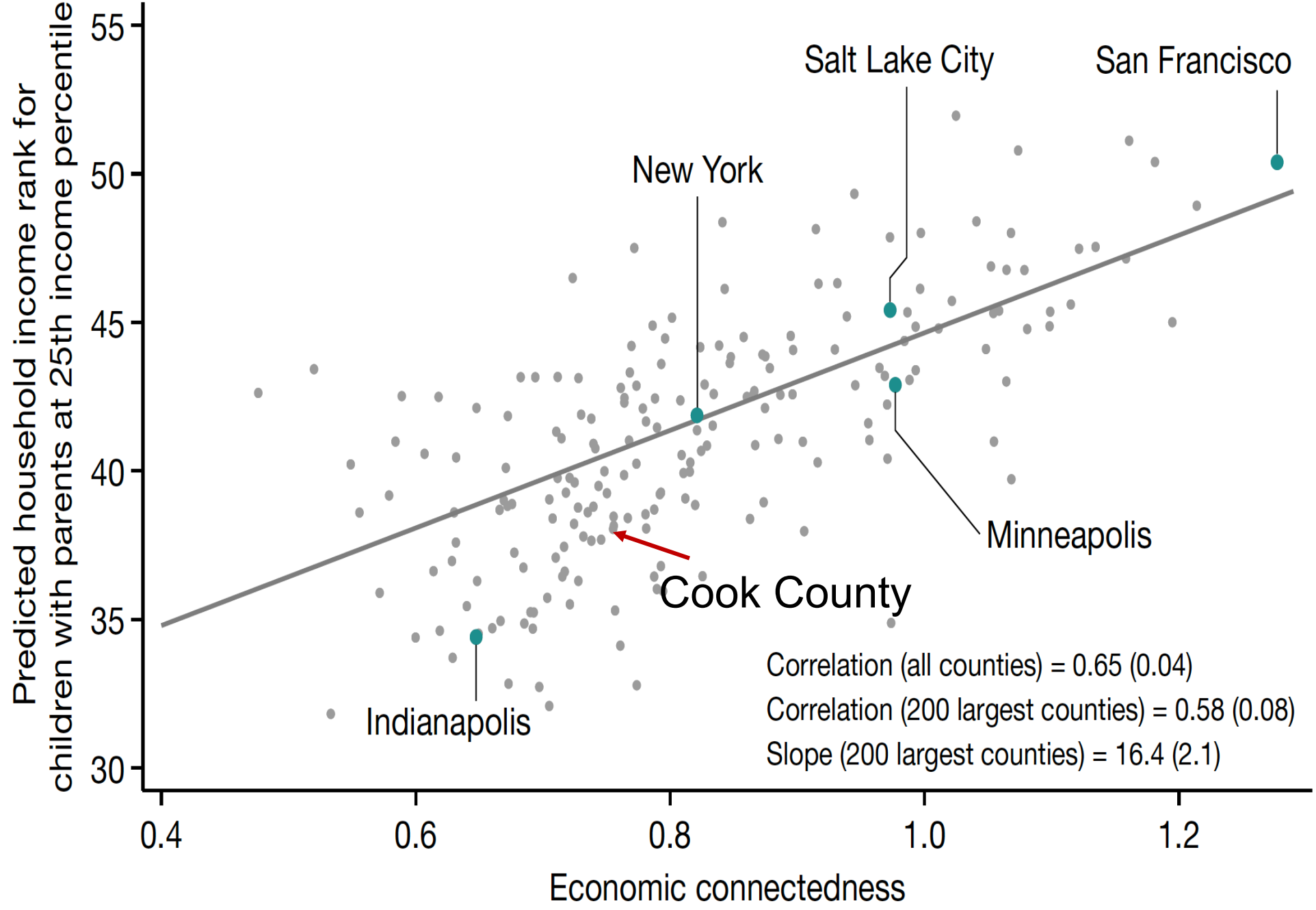
Economic Connectedness of Low-SES Individuals by County

Normalized Share of Above-Median Friends Among Below-Median People

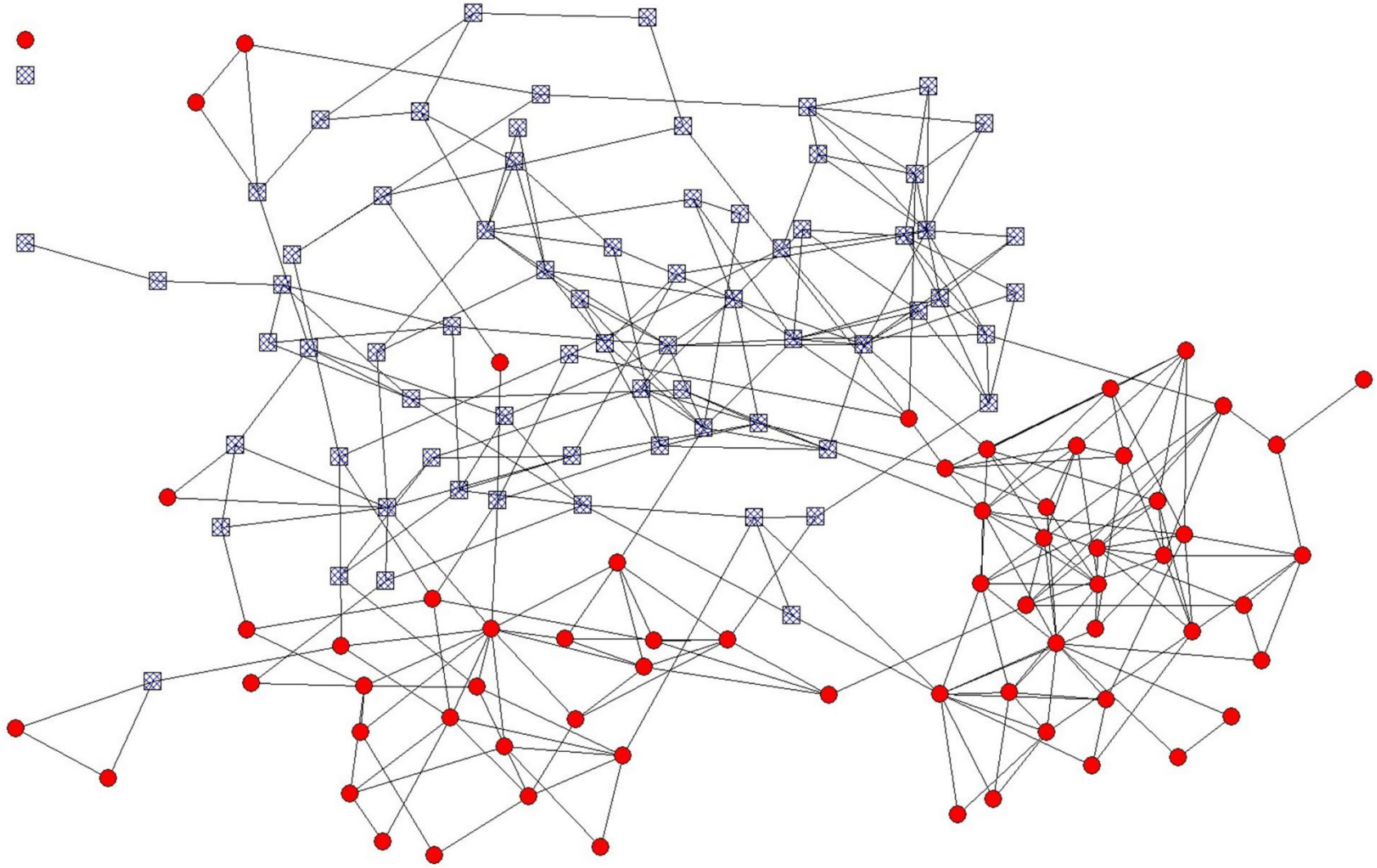


Chetty, Jackson, Kuchler, Stroebel et al 2022 (www.socialcapital.org)

Upward Mobility vs. Economic Connectedness, by County 200 Largest Counties

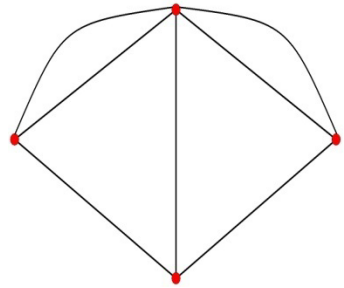


V26 KeroRiceGo Banerjee, Chandresekhar, Duflo, Jackson (2013, 2019)



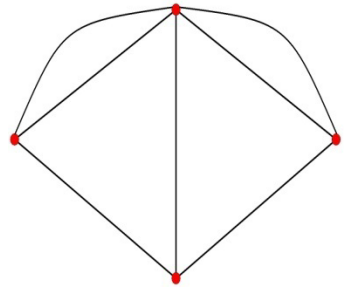
Blue Squares: General Merit and Otherwise Backward Castes
Red Circles: Scheduled Castes and Scheduled Tribes Figure Jackson 2014

Outline



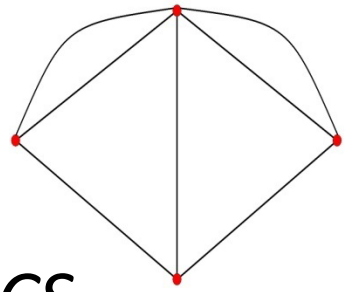
- Background on models of network formation
- SUGMs and Identification
- Estimation and CLT
- Application, Simulations

Outline



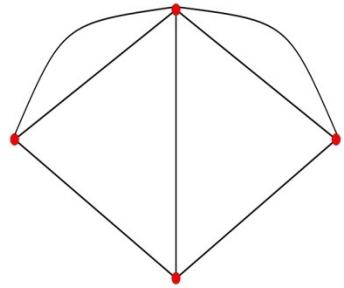
- Background on models of network formation
- SUGMs and Identification
- Estimation and CLT
- Application, Simulations

Network Formation Models



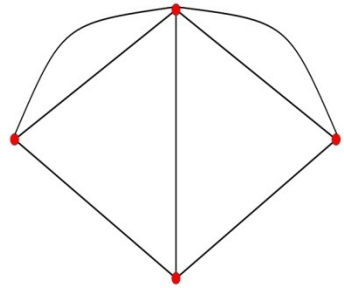
- **Random graphs** – *Math, Sociology, Statistics, Physics, Economics, CS*
  - **Static:** Solomonoff-Rapoport 1951, Rapoport 1957, Erdos-Renyi 1959, 1960, Holland-Laskey-Leinhardt 1983, Molloy-Reed 1995, Watts-Strogatz 1999, Chung-Lu 2002
  - **Dynamic:** Price 1976, Barabasi-Albert 1999, Adamic-Huberman 1999, Leskovec-Kleinberg-Faloutsos 2005, Jackson-Rogers 2007, Snijders-Van de Bunt-Steglich 2010
- **Strategic models** – *Economics, CS, Political Science*
  - **Static:** Jackson-Wolinsky 1996, Dutta-Jackson 2000, Jackson-van den Nouweland 2005, Calvo-Iklic 2004, Jackson-Rogers 2005, Bloch-Jackson 2006, 07, Olaizola-Valenciano 2021
  - **Dynamic:** Aumann-Myerson 1987, Bala-Goyal 2000, Currarini-Morelli 2000, a.Watts 2001, Jackson-Watts 2002ab, Mauleon-Vannetelsoch 2004, Currarini-Jackson-Pin 2009, Christakis-Fowler-Imbens-Kalyanaraman 2020, Jackson-Nei-Snowberg-Yariv 2023

Statistical/Econometric Models

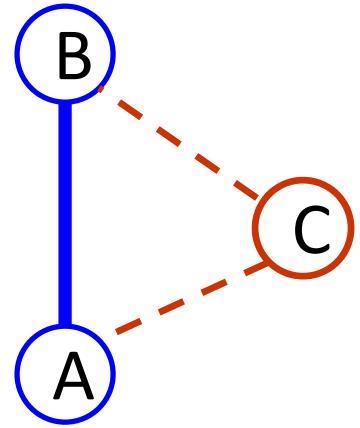


- Stochastic Block Models: Holland-Laskey-Leinhardt (1983) …
- ERGMs (Markov, p*, MRQAP): Holland-Leinhardt (1981), Frank-Strauss (1986), Krackhardt (1988), Butts (2009), Snijders etal (2006), …
- Spatial/Geometric/Latent: Penrose (2003), Hoff etal (2002), Leung (2014), McCormick-Zheng (2015), Boucher Mourifie (2017), Graham (2017)
- Explicit strategic formation models: Mele (2017,2022), Badev (2021), de Paula-RichardsShubik-Tamer (2018), Sheng (2020), Mele-Zhu (2023)

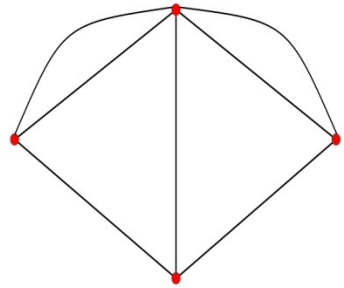
Example: Social Pressure



- Cross caste relationships
  - Are they more likely to occur ``in private'' with no friends in common
  - Or occur with same frequency in embedded relationships?



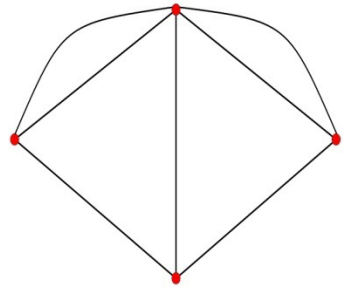
ERGMs



$$\Pr(g) = \frac{\exp[\,\beta_1 s_1(g) + \ldots + \beta_k s_k(g)\,]}{\sum_{g'} \exp[\,\beta_1 s_1(g') + \ldots + \beta_k s_k(g')\,]}$$

MCMC techniques for estimation (Snijders 2002, Handcock 2003,...) have led to these becoming the standard

ERGMs



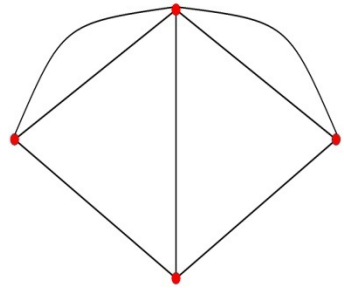
Bhamidi, Bresler, Sly (2008) (see also Chatterjee and Diaconis (2011), Shalizi and Rinaldo (2012)):

For dense enough ERGMs, MCMC (Glauber dynamics - Gibbs sampling) estimates mix less than exponentially ***only if*** networks have approximately independent links.

So, ERGMs that are interesting, cannot be estimated via techniques being used!

Simulations: also problems on sparse ones...

ERGMs

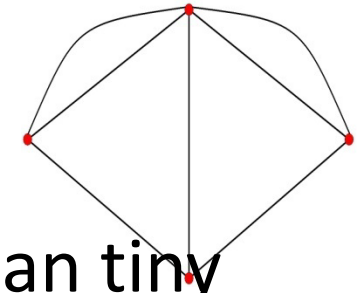


$$\Pr(g) = \frac{\exp[\,\beta_1 s_1(g) + \ldots + \beta_k s_k(g)\,]}{\sum_{g'} \exp[\,\beta_1 s_1(g') + \ldots + \beta_k s_k(g')\,]}$$

n=30 nodes give 2^{435} g's (less than 2^{258} atoms in the universe...)

Sampling g's will not lead to accurate estimates (not just MCMC limitation)

Statistical/Econometric Models

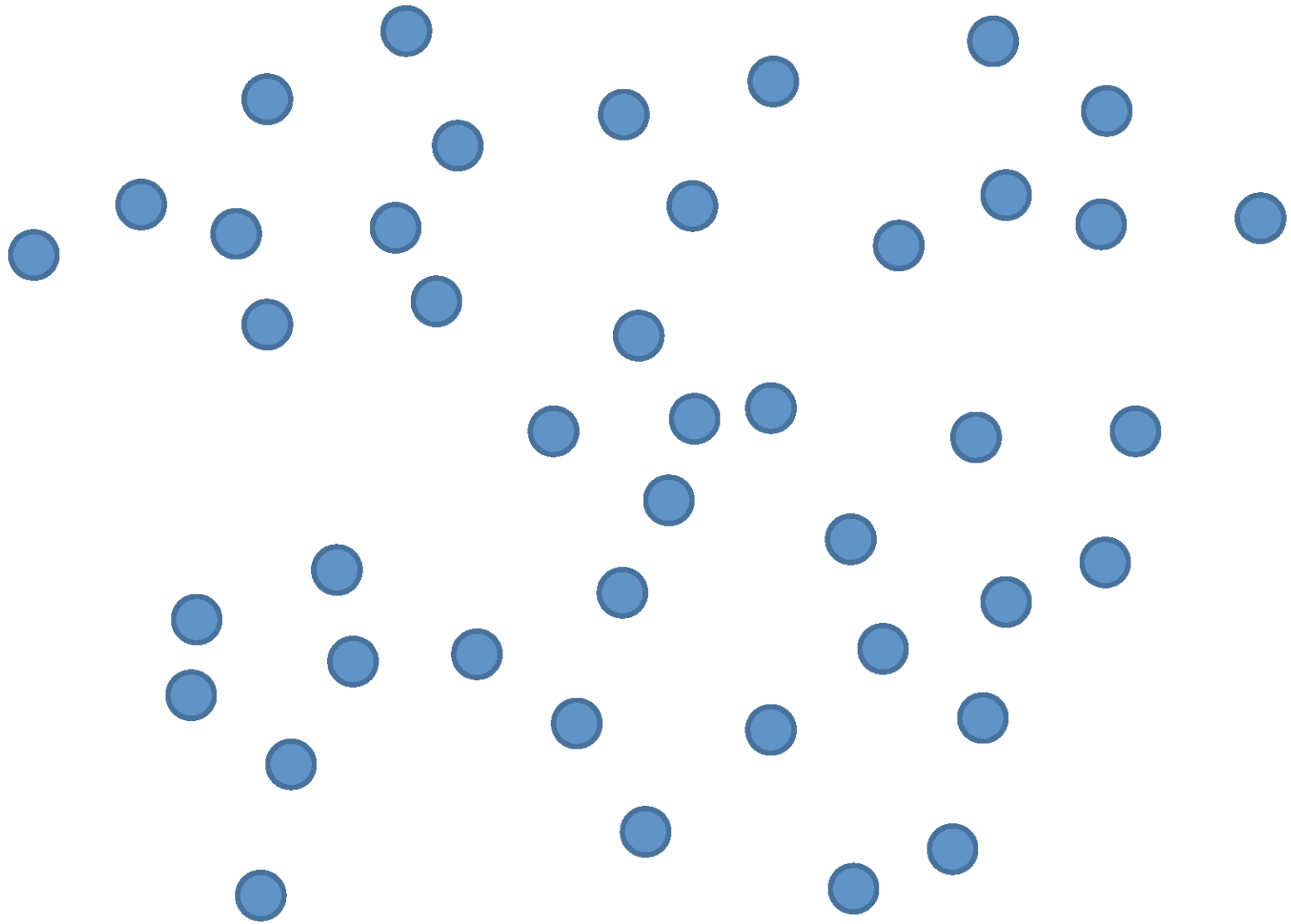


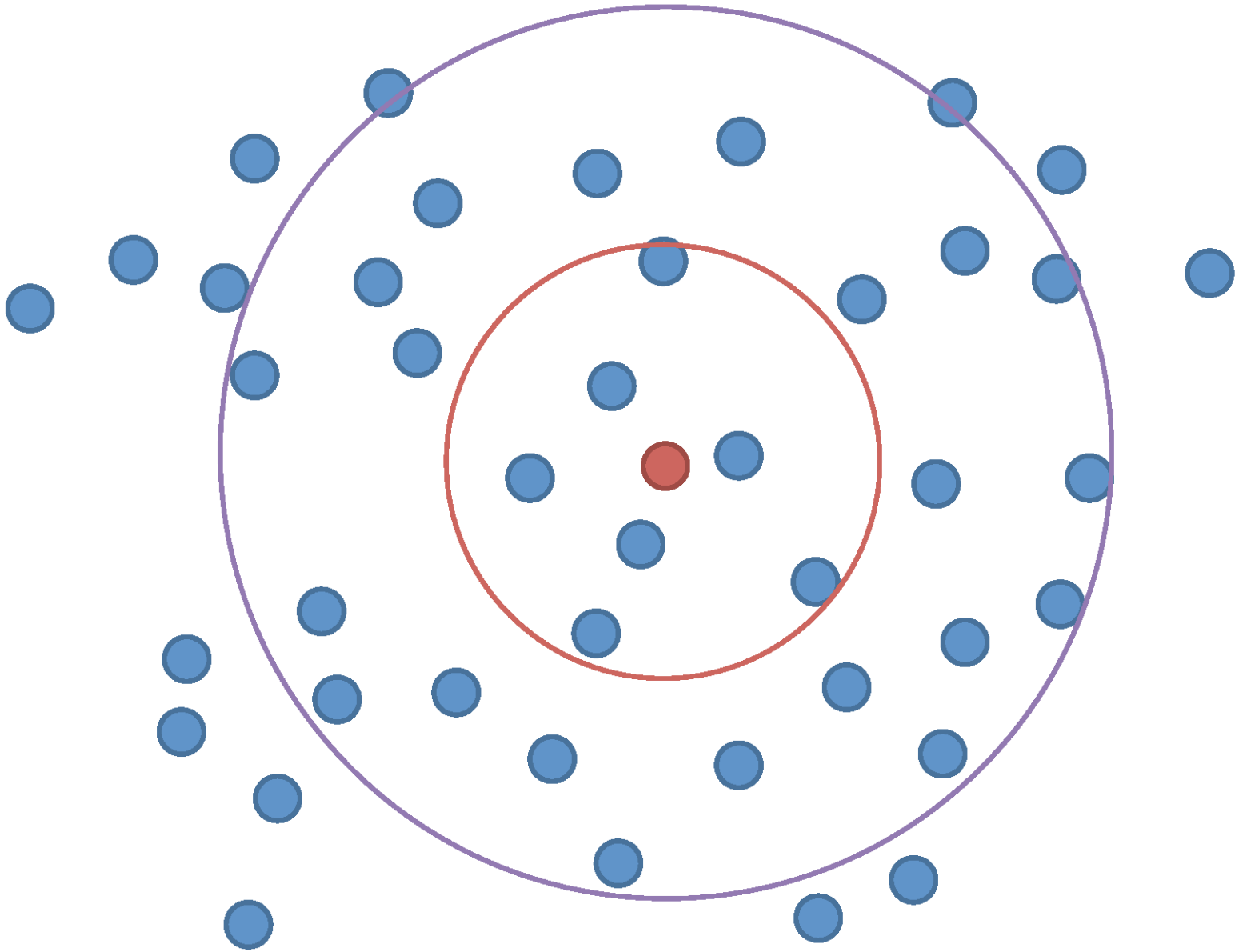
- ERGMs impossible to estimate with rich structures and more than tiny number of nodes

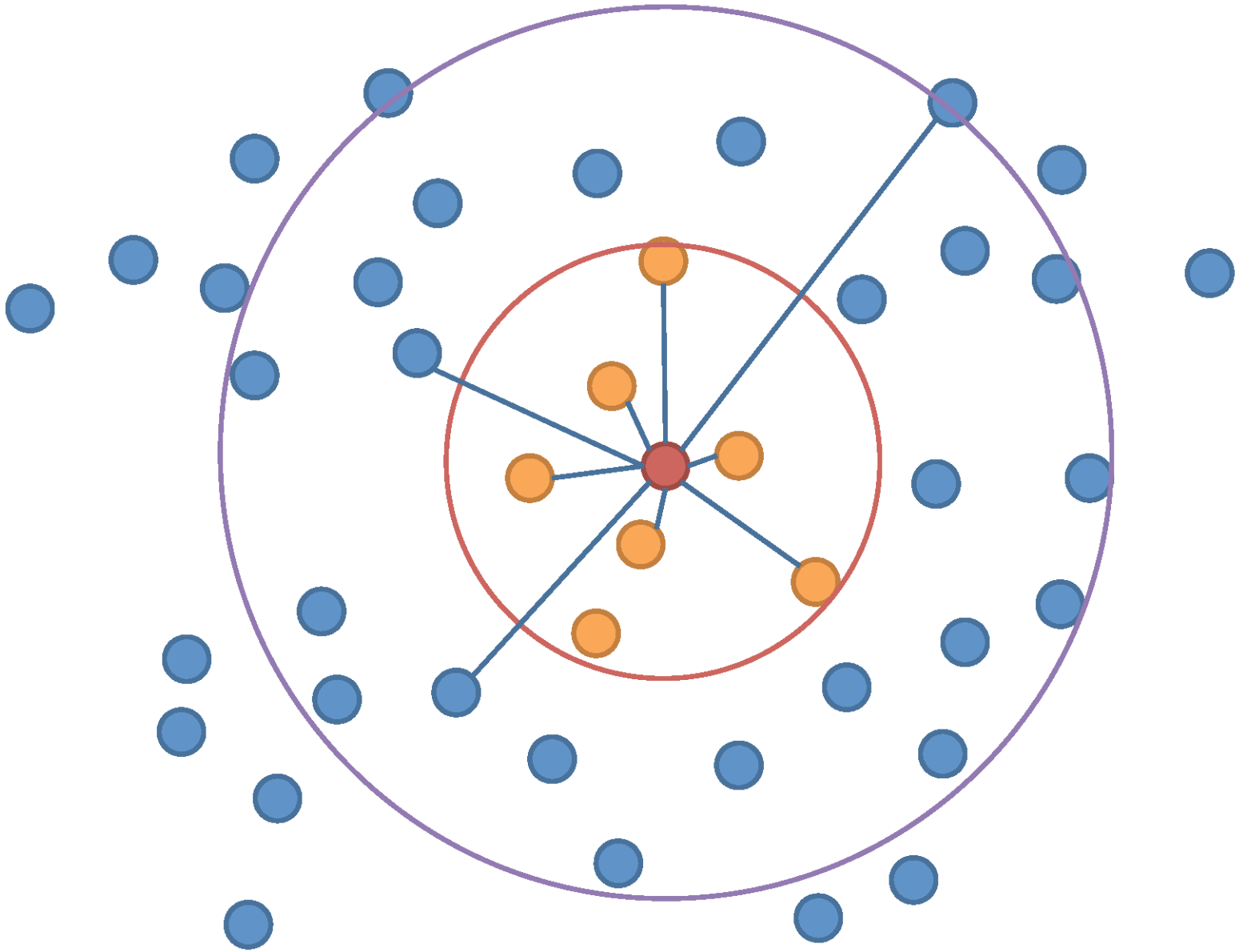
- Spatial/Geometric/Latent:

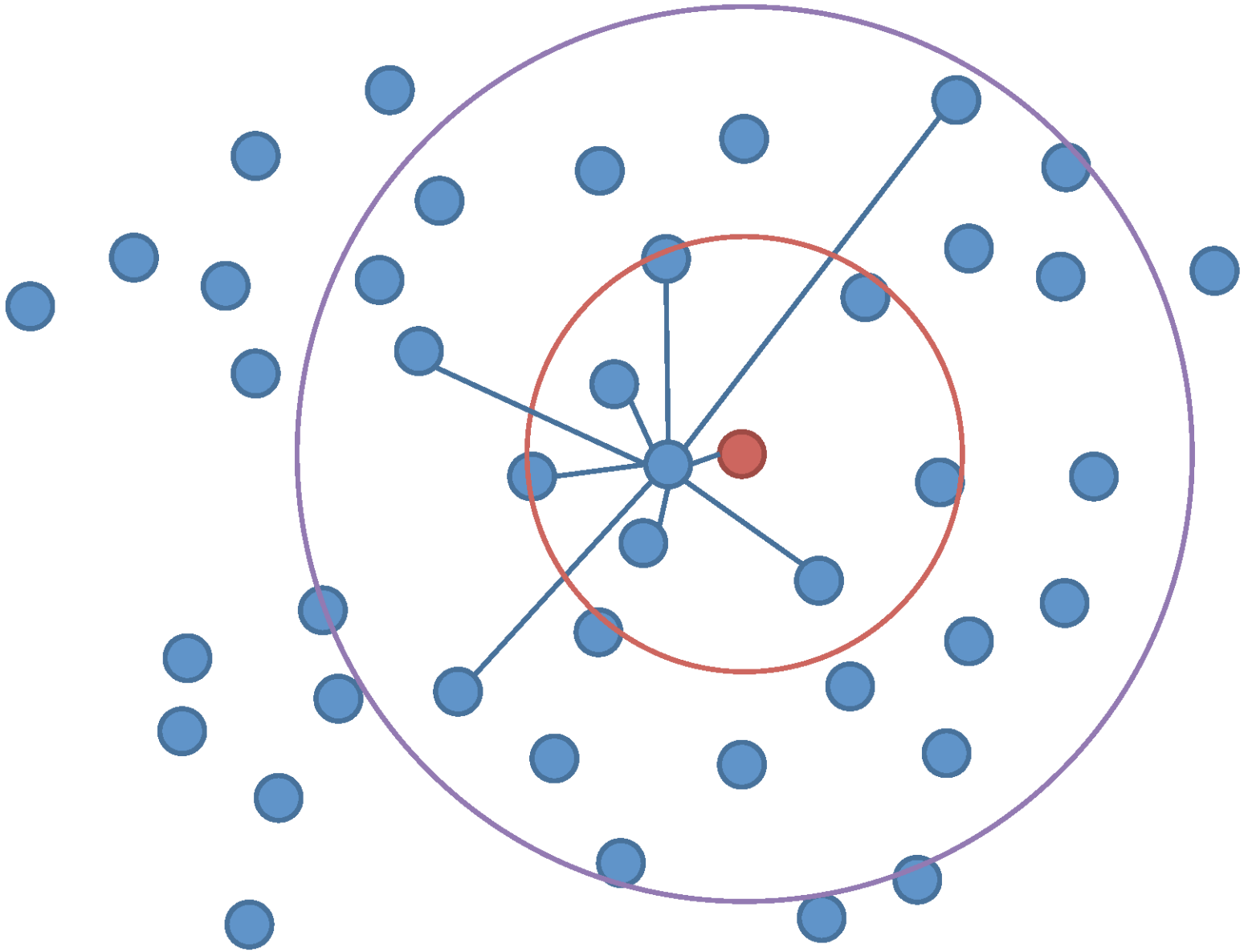
  - Links follow some spatial pattern

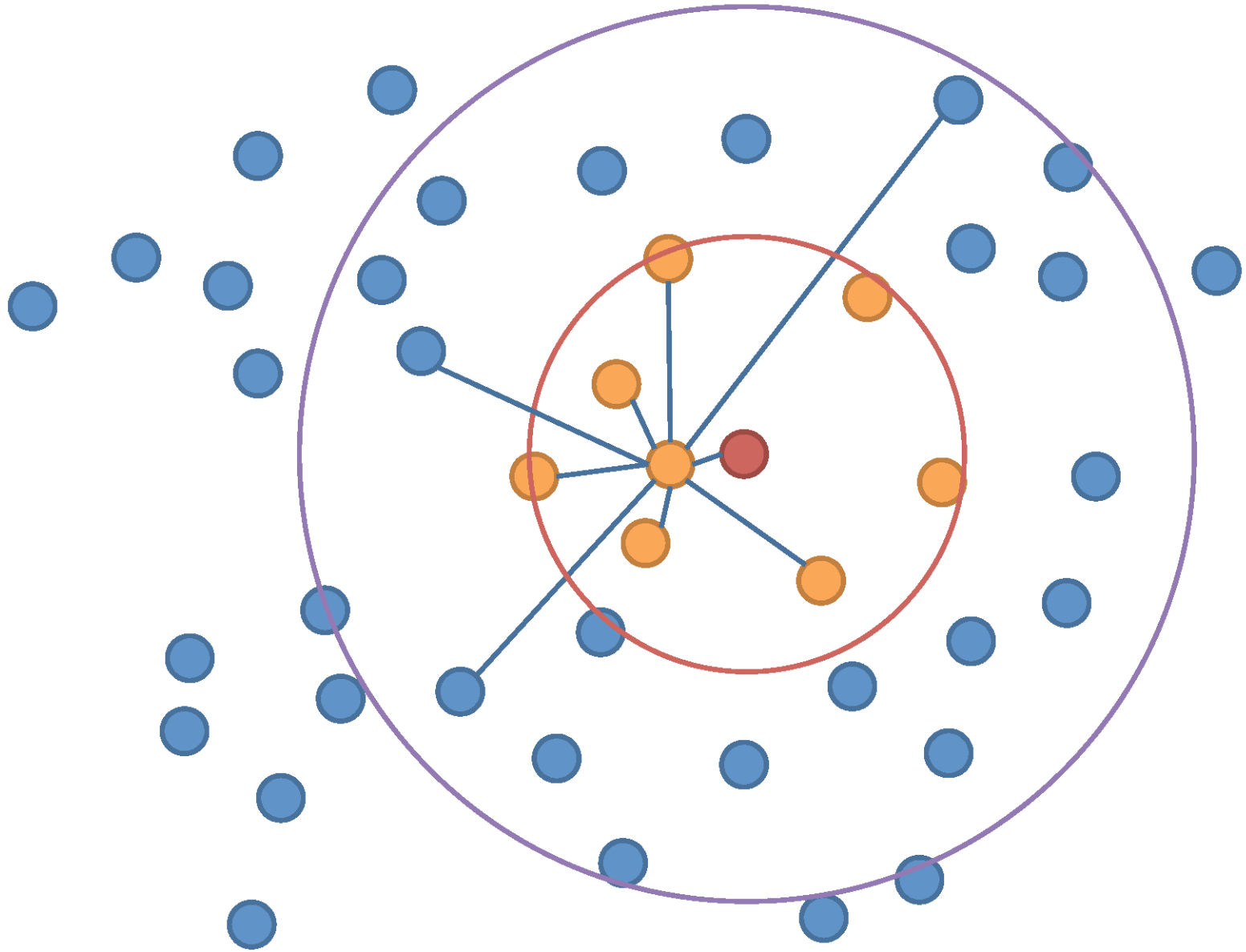
  - But end up with too high density nearby in order to generate triangles or other richer patterns in networks

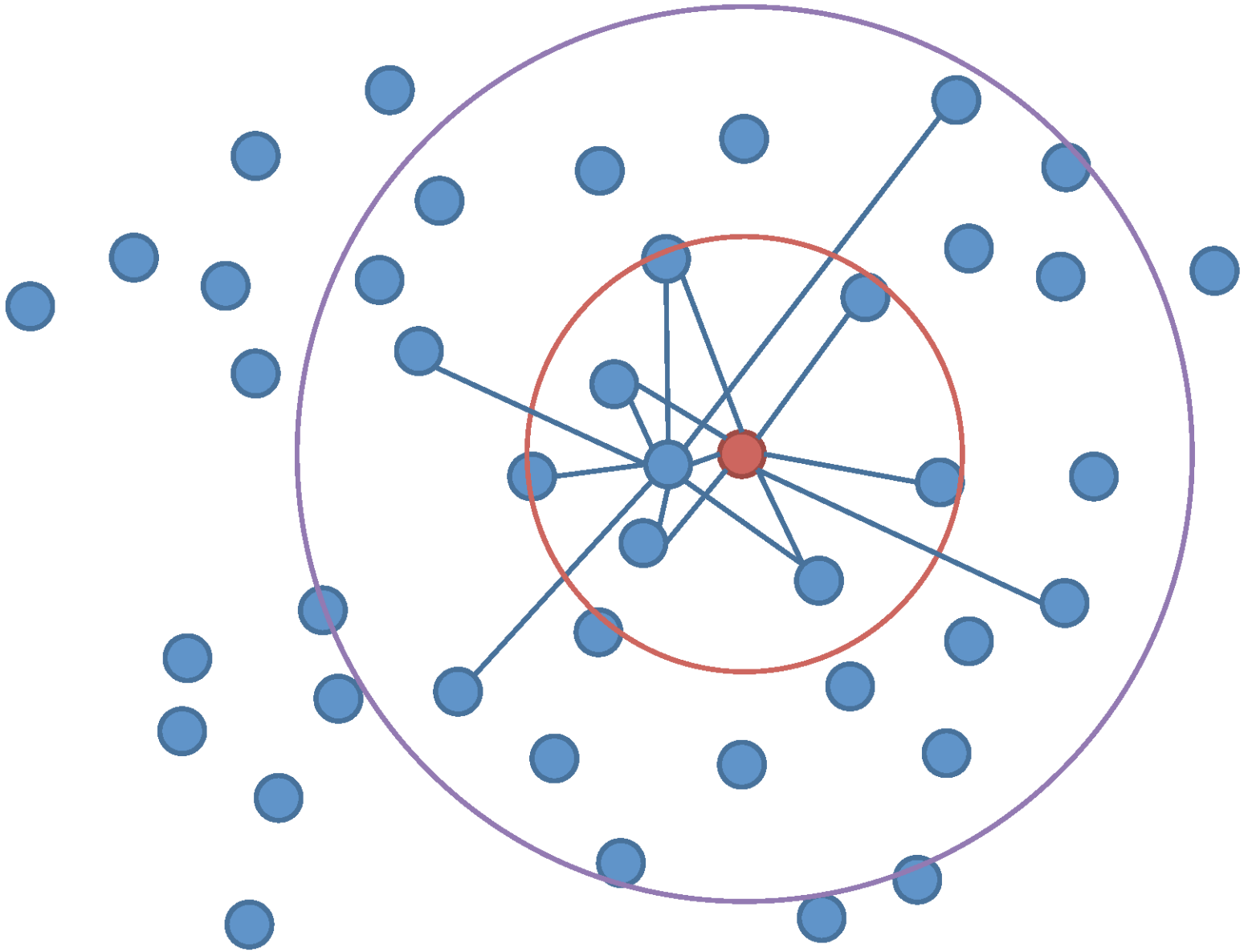


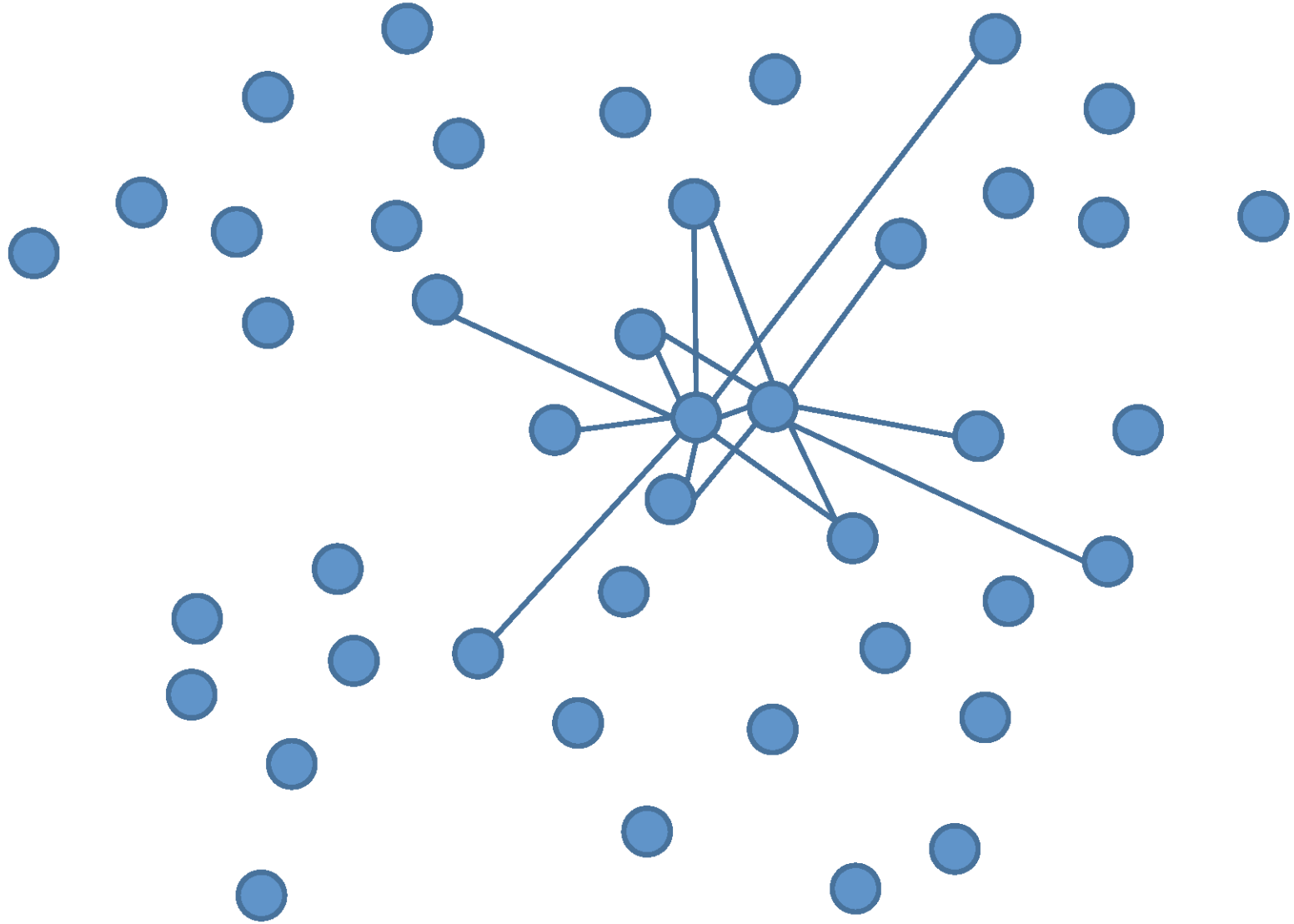




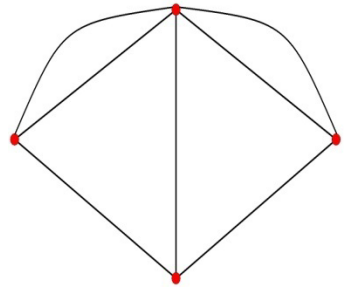






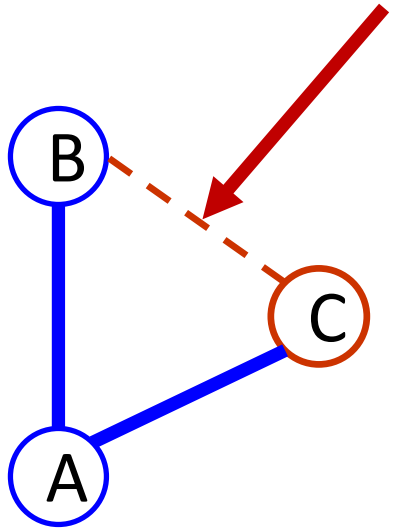


Clustering



- Fraction of pairs of a node's friends that are friends with each other

Fraction of such instances where this link is present



Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8			
Clustering	.29	.05			
Frac. Giant Comp.	.95				
First Eigvalue	10.1				
Homophily	.94				
Avg Path Length	3.5				

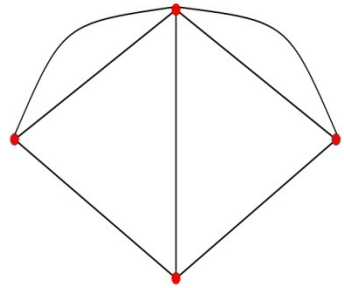
Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8	16.6		
Clustering	.29	.05	.15		
Frac. Giant Comp.	.95				
First Eigvalue	10.1				
Homophily	.94				
Avg Path Length	3.5				

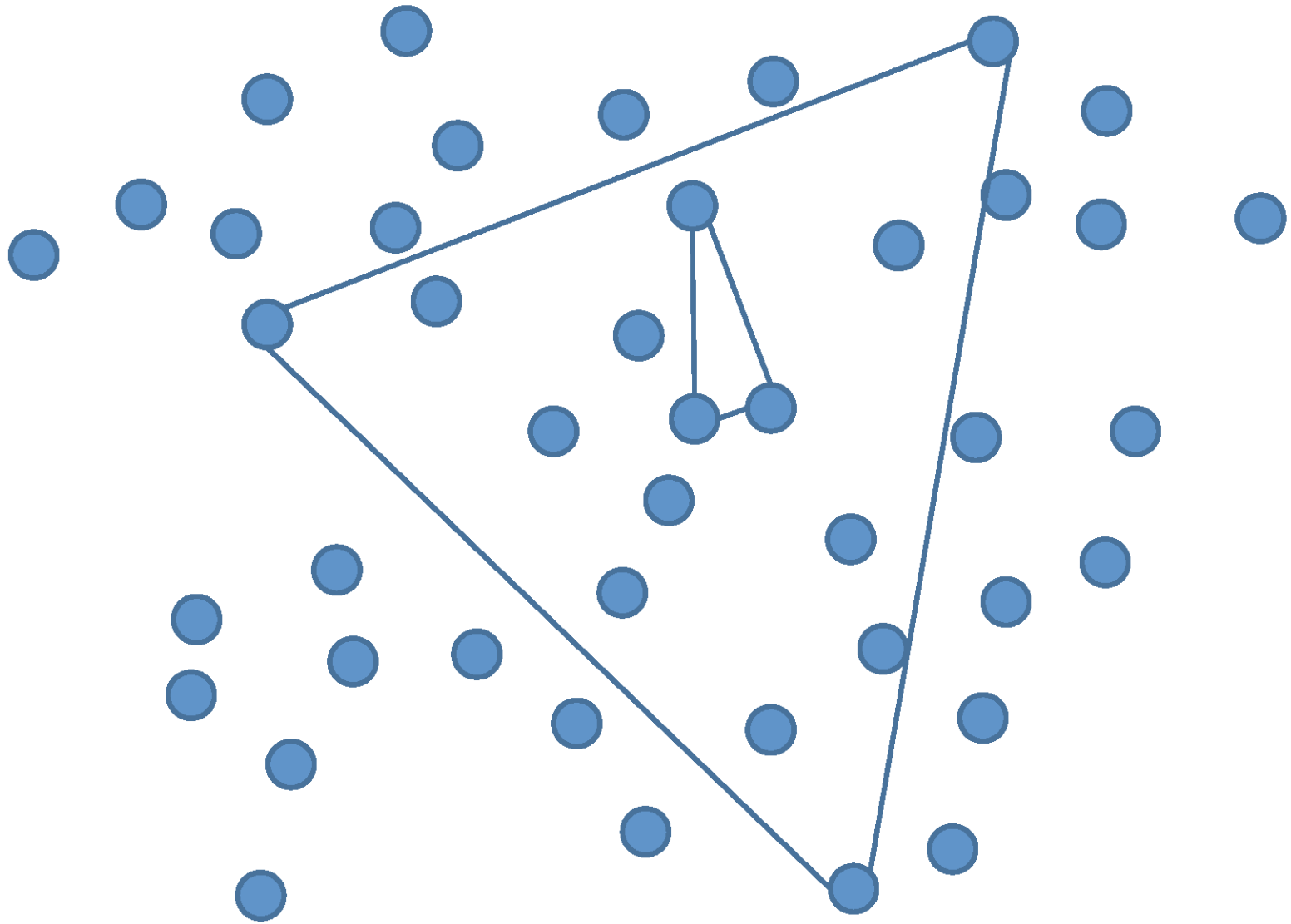
Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8	16.6	13.1	
Clustering	.29	.05	.15	.07	
Frac. Giant Comp.	.95				
First Eigvalue	10.1				
Homophily	.94				
Avg Path Length	3.5				

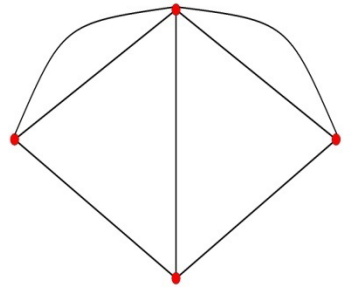
Statistical/Econometric Models



- Spatial/Geometric/Latent:
  - Links follow some spatial pattern
  - But end up with too high density nearby in order to generate clustering
- *Instead: simply generate triangles and other subgraphs directly…*

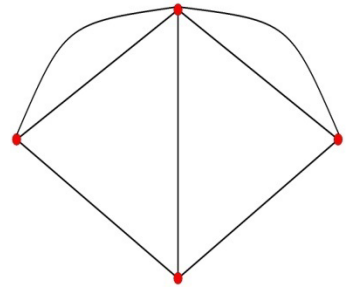


Outline



- Background on models of network formation
- SUGMs and Identification
- Estimation and CLT
- Application, Simulations

SUGMs



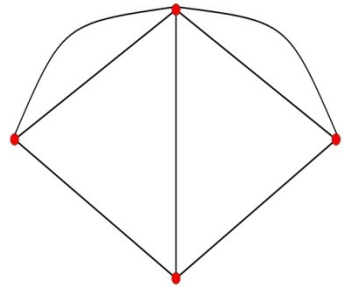
- Subgraph Generation Models

- *Subgraphs* are generated, network is by-product

  people form links, triangles,…

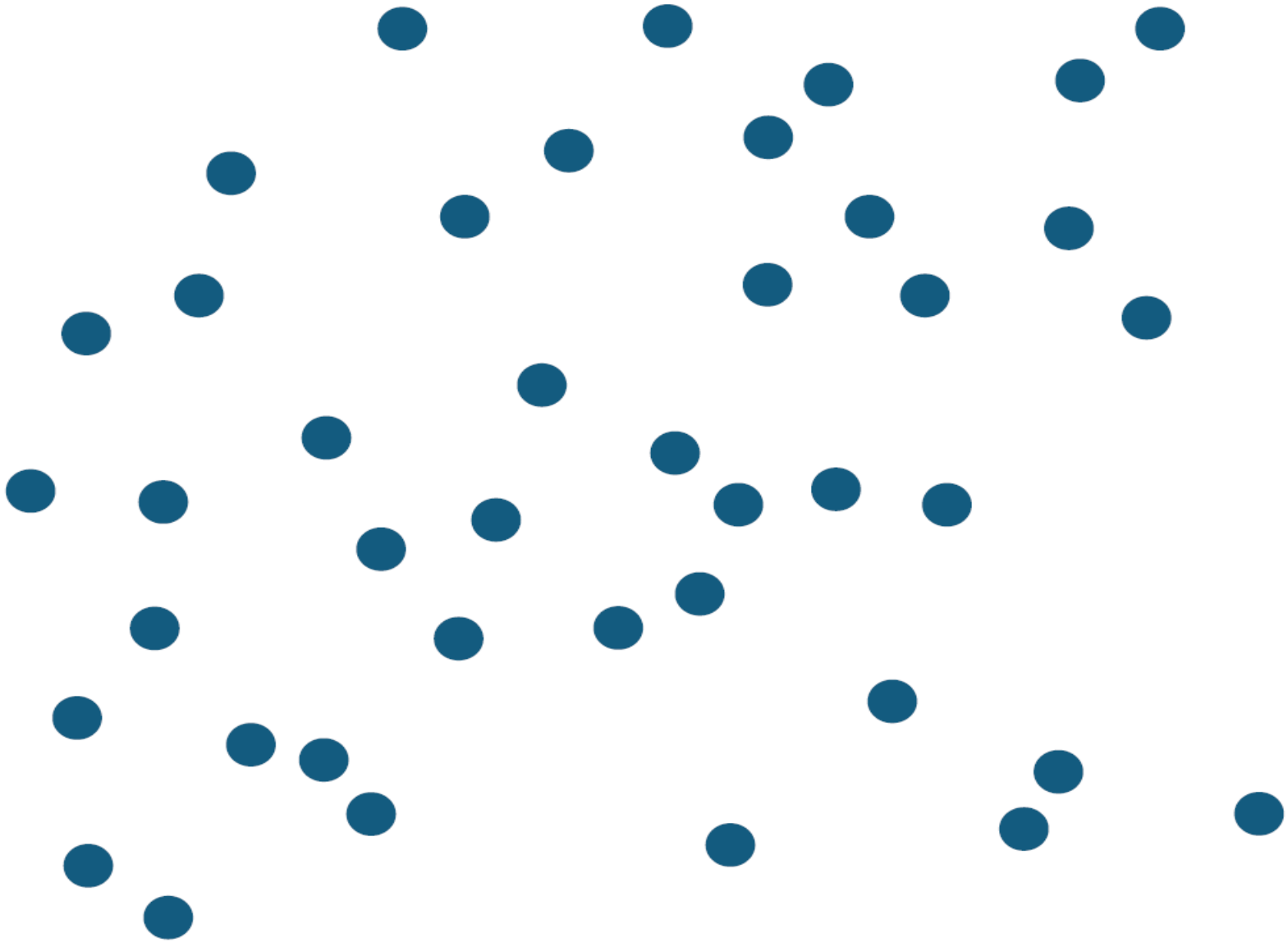
  could depend on X_is of the people involved

SUGMs

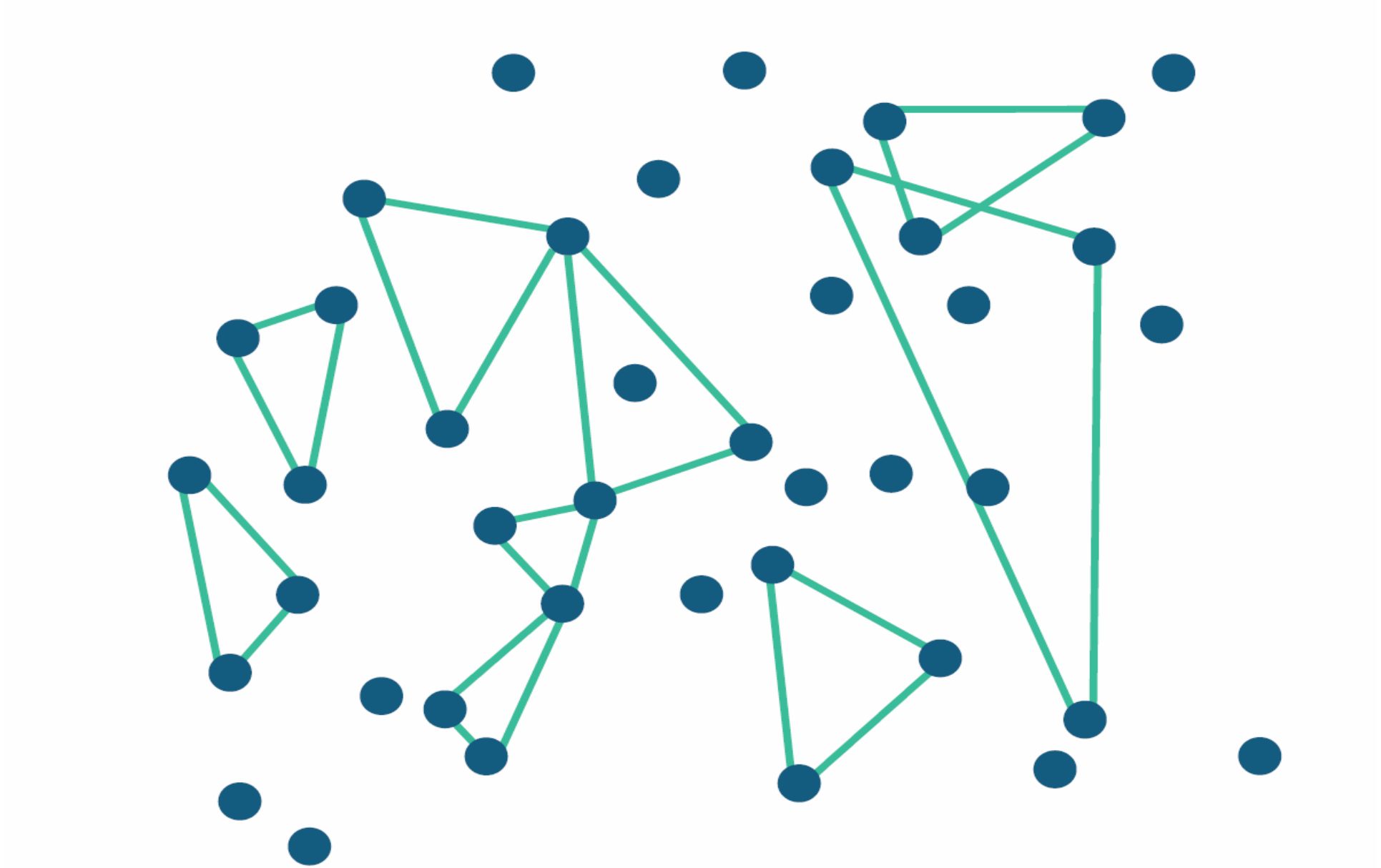


- $N = \{1, \ldots, n\}$ nodes
- G_k set of all subgraphs of type k on n nodes (e.g., triangles that involve more than one caste, triangles of same case)
- $\beta = (\beta_1 \ldots \beta_K)$ probabilities of subgraphs type k forming
- Subgraphs form independently

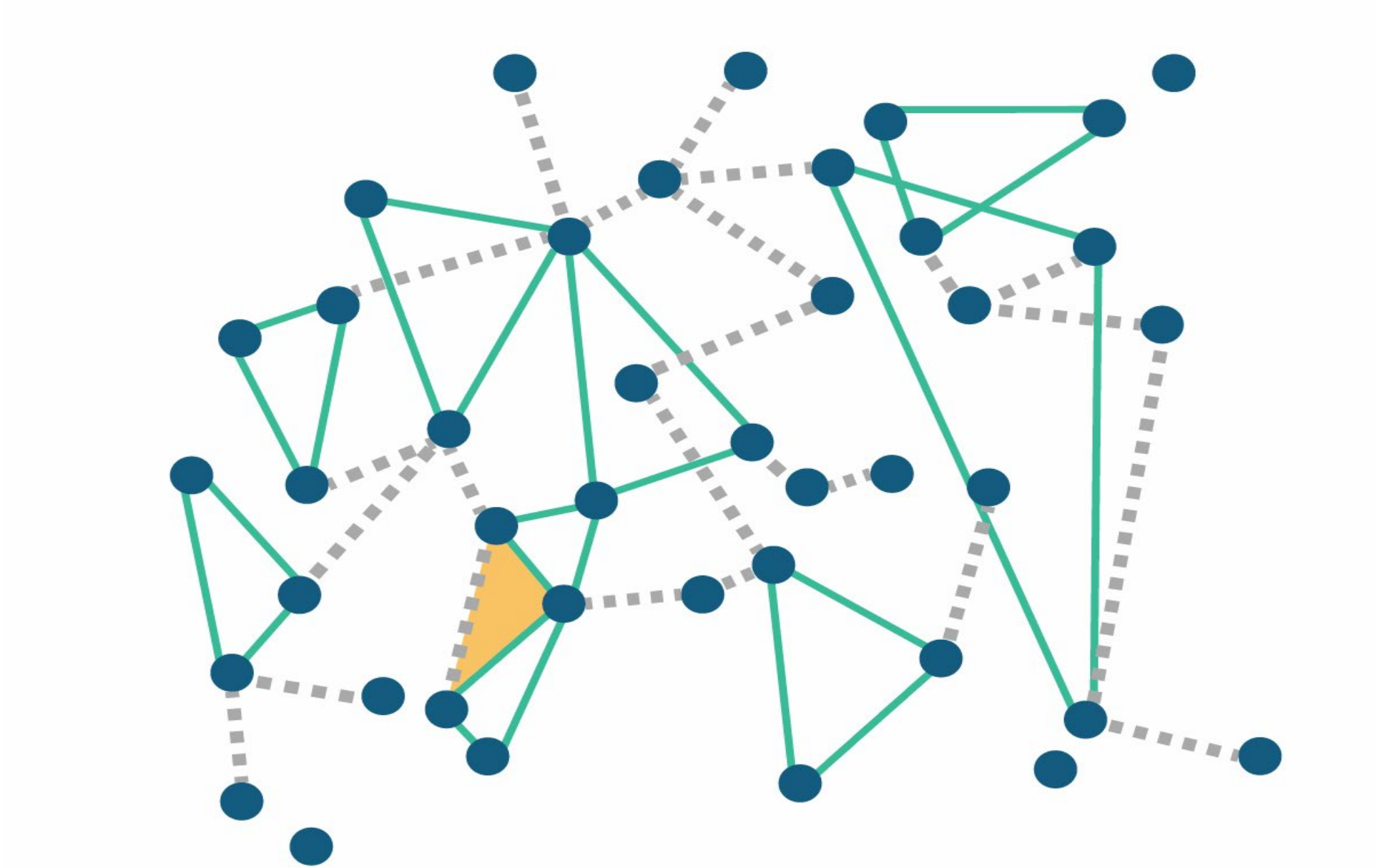
Example:



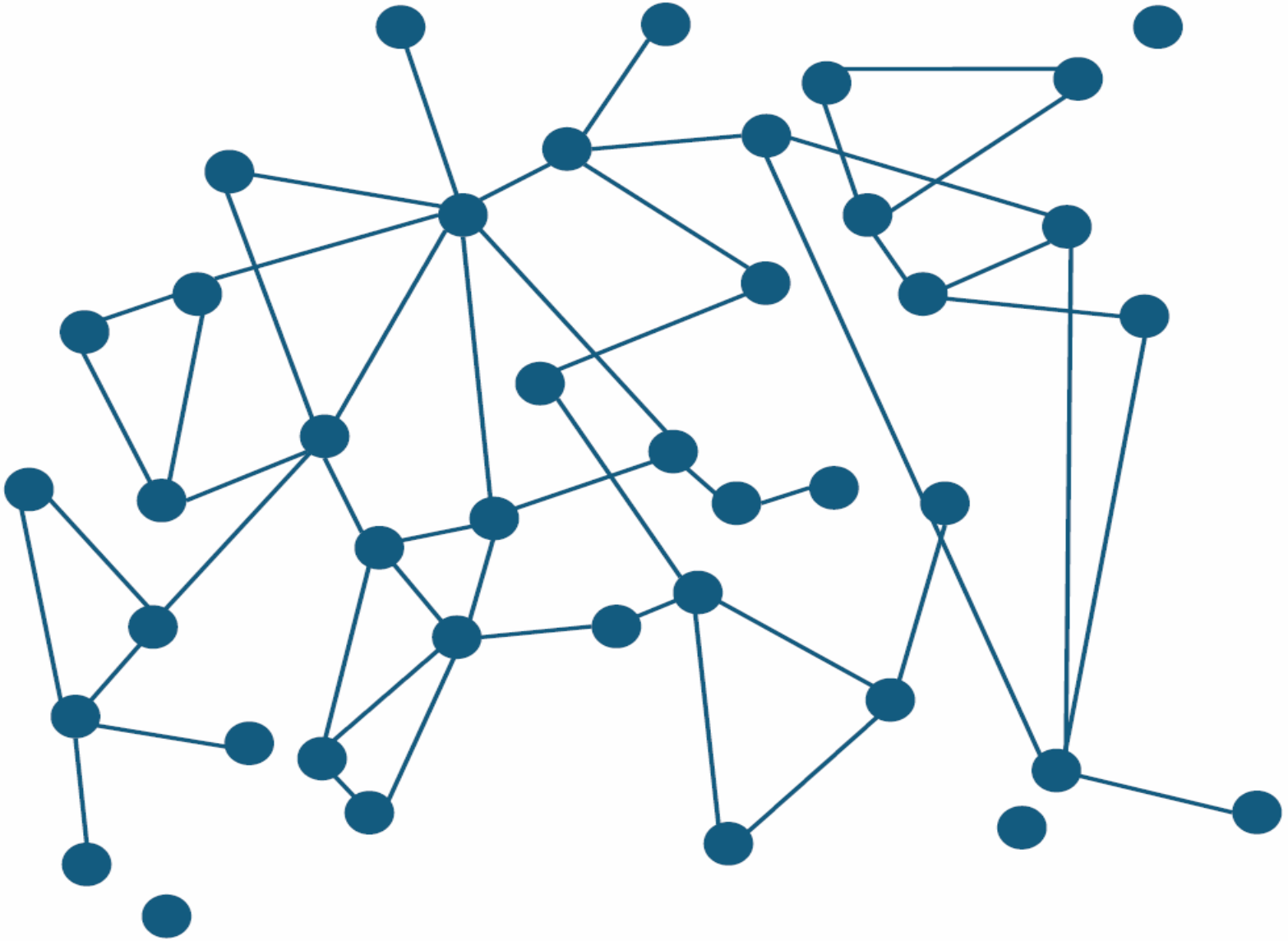
Example:



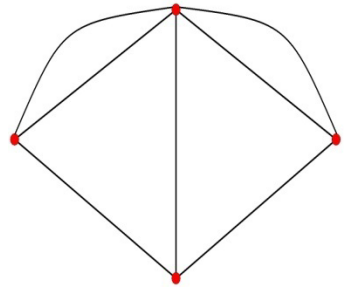
Links Form: Incidental Triangle



We See:



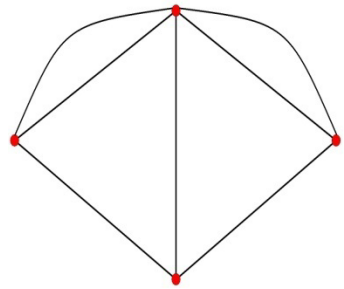
Identification Challenge



Subgraphs are formed directly, but only see the ending network

Was some subgraph generated directly, or incidentally as part of some other links/subgraphs

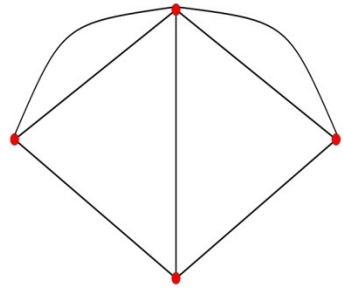
Theorem: Identification



Every SUGM is identified: For any $G = (G_1 \ldots G_K)$

if $\beta = (\beta_1 \ldots \beta_K) \neq \beta' = (\beta'_1 \ldots \beta'_K)$ then $Pr_{\beta}( ) \neq Pr_{\beta'}( )$.

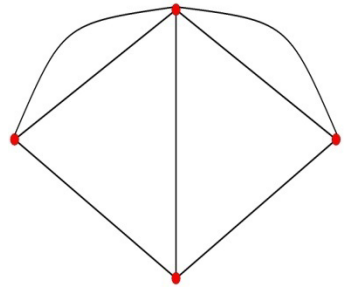
Proof



e.g., links triangles:

- Probability of network with just one link depends only on β_L if those are the same, then
- Probability of some triangle (and nothing else) $\beta_T + (1- \beta_T) \beta_L^3$ increasing in β_T
- General version – find subgraph with fewest links where β differs

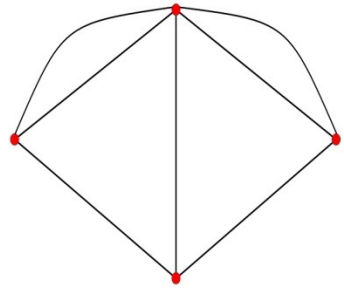
Proposition: Identification



$S_L(g)$ = # links in g, $S_T(g)$ = # triangles in g.

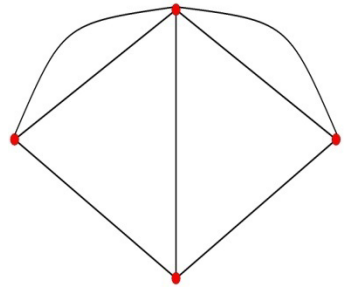
If $(\beta_L, \beta_T) \neq (\beta'_L, \beta'_T)$ then $E_\beta [S_L(g), S_T(g)] \neq E_{\beta'} [S_L(g), S_T(g)]$.

Outline



- Background on models of network formation
- SUGMs and Identification
- Estimation and CLT
- Application, Simulations

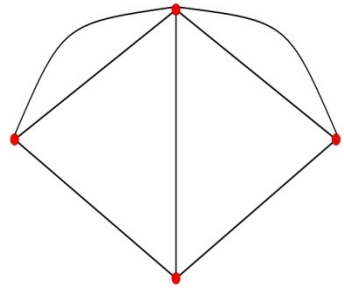
Estimation



- Many networks (``easy'', standard – each network is an independent observation)

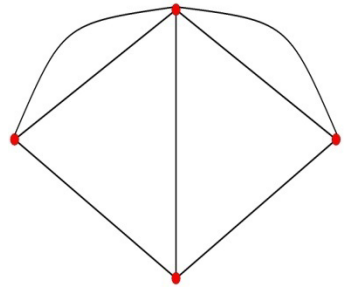
- One network (harder, links potentially all correlated) -- we use two approaches
  - ``Direct count estimator'' - sparse networks, count subgraphs (carefully – larger subgraphs first)
  - Minimum distance estimator - non-sparse case, show how to do it for links, triangles; raw counts

Direct Count Estimator



- Order G_ks from largest to smallest (number of links).
- Count subgraphs in G_1, $S_1(g)$, then eliminate them.
- Iteratively, on remaining network count subgraphs in G_k, $S_k(g)$, then eliminate them.
- $\hat{\beta}_k = S_k(g) / |G_k|$

Sparse Direct Count Estimation

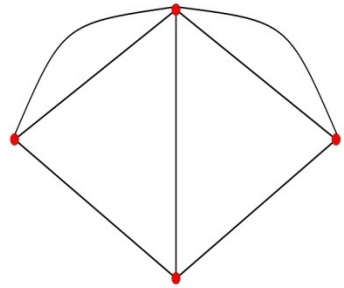


m_k number nodes subgraph k $\beta_k = \frac{b_k}{n^{h_k}}$

$h_k \in (m_k - 1, m_k)$

Subgraphs dense enough to appear in large numbers, but sparse

Sparse Direct Count Estimation

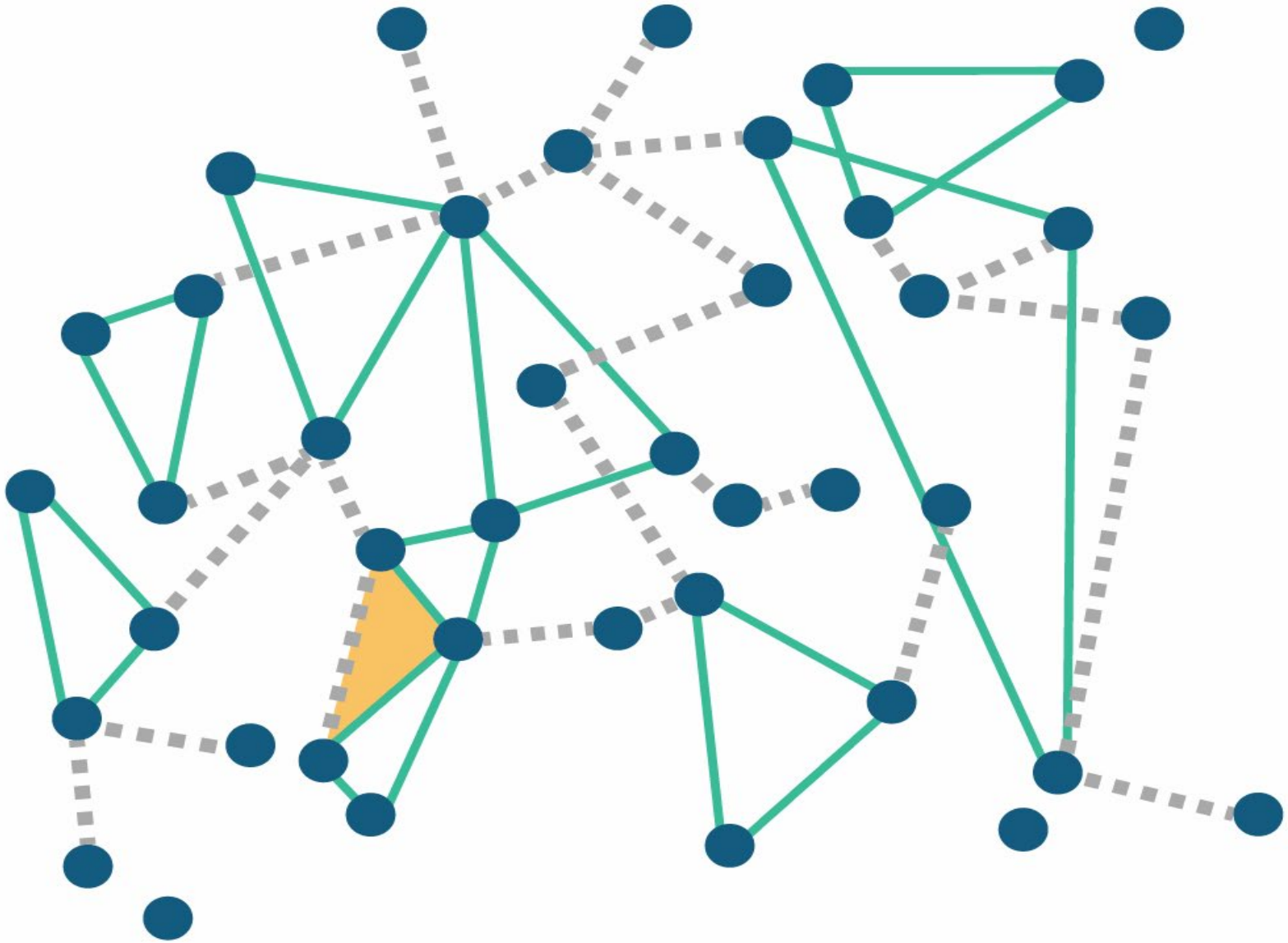


Let $V_n = \text{diag}\left\{ n^{2h_k} \frac{\beta_k^n(1-\beta_k^n)}{|G_k|} \right\}$ and $h_k \in (m_k - 1, m_k)$.

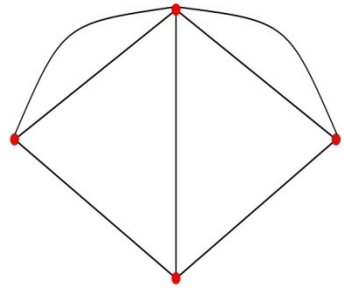
Then $|b^{DC} - b| \to^P 0$ and $V_n^{-1/2}(b^{DC} - b) \rightsquigarrow \mathcal{N}(0, I)$.

(betas converge too, but all going to 0, so state in bs)

Sparse: Few Incidental Triangles



Non-Sparse, Links Triangles



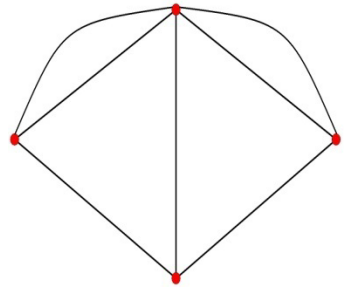
$$R = \text{diag}\{n^{h_L}, n^{h_T}\}$$

$$\beta^{MD} := \text{argmin}\,(S(g) - \mathrm{E}_\beta[S(g)])' R^2 (S(g) - \mathrm{E}_\beta[S(g)])$$

If $h_L \in (2/3,2)$ and $h_T \in [h_L + 1,\ 3h_L]$, with $h_T < 3$

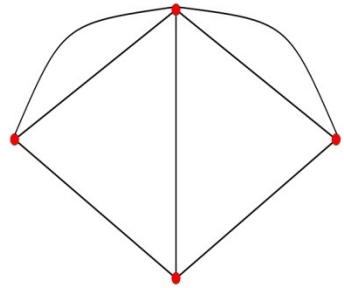
then $|b^{MD} - b| \to^P 0$ and $V_n^{-1/2}(b^{MD} - b) \rightsquigarrow \mathcal{N}(0, I)$.

Proof?



- Show identification from direct statistics and that those vary nontrivially with parameters
- Key missing piece is to get normality
- Subgraph counts are sums of correlated random variables
- Correlation structure can be quite complex (adjacent and non adjacent)

A New Central Limit Theorem

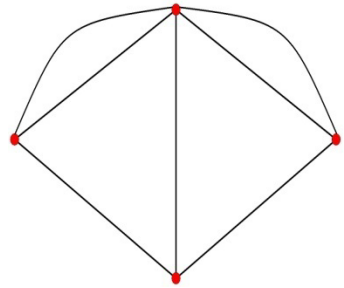


- In order to prove things we need a new Central Limit Theorem that handles correlated random variables

- Existing theorems usually use some time or spatial restrictions on correlation

- We don’t have that: we need a new more flexible Central Limit Theorem that allows for arbitrary correlation patterns as long as total correlation is appropriately bounded

CLT – correlated RVs



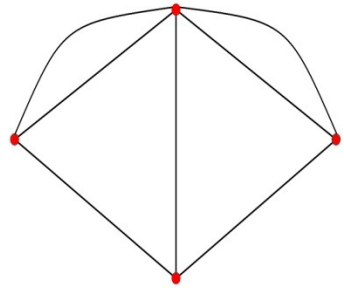
Index set triangular array $\alpha \in \Lambda^N$

Affinity set: $\mathcal{A}(\alpha, N) \subset \Lambda^N$ such that $\alpha \in \mathcal{A}(\alpha, N)$

$$Z_\alpha = X_\alpha - E[X_\alpha]$$

$E[|Z_\alpha^N|^3]/E[(Z_\alpha^N)^2]^{3/2}$ is bounded above

CLT – correlated RVs



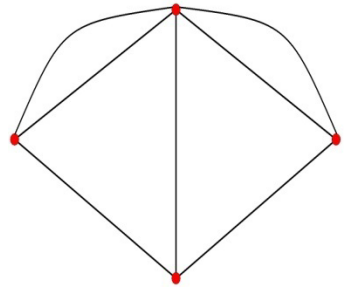
index set triangular array $\alpha \in \Lambda^N$

Affinity set: $\mathcal{A}(\alpha, N) \subset \Lambda^N$ such that $\alpha \in \mathcal{A}(\alpha, N)$

$$Z_\alpha = X_\alpha - E[X_\alpha]$$

$$a_N := \sum_{\alpha;\eta \in \mathcal{A}(\alpha)} cov(Z_\alpha, Z_\eta) \qquad \mathbf{Z}_{-\alpha} := \sum_{\eta \notin \mathcal{A}(\alpha)} Z_\eta$$

Three conditions

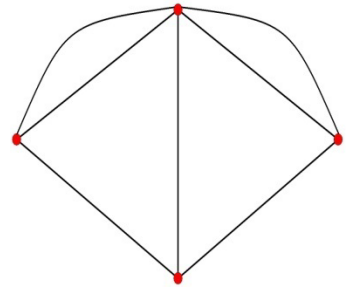


(i) $$\sum_{\alpha;\eta,\gamma\in\mathcal{A}(\alpha)} E\left[|Z_\alpha|Z_\eta Z_\gamma\right] = o\left(a_N^{3/2}\right)$$

(ii) $$\sum_{\alpha,\alpha',\eta\in\mathcal{A}(\alpha),\eta'\in\mathcal{A}(\alpha')} cov\left(Z_\alpha Z_\eta, Z_{\alpha'} Z_{\eta'}\right) = o(a_N^2)$$

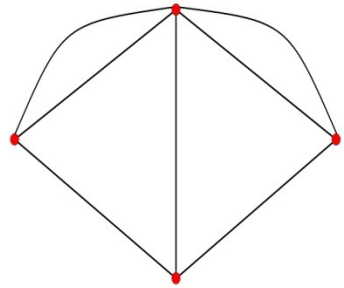
(iii) $$\sum_\alpha E[|E[Z_\alpha \mathbf{Z}_{-\alpha}|\mathbf{Z}_{-\alpha}]|] = \sum_\alpha E[|\mathbf{Z}_{-\alpha} E[Z_\alpha|\mathbf{Z}_{-\alpha}]|] = o(a_N)$$

CLT



Under (i)-(iii) $\sum_{\alpha \in \Lambda^N} Z_\alpha^N / \sqrt{a_N} \rightsquigarrow \mathcal{N}(0,1).$

CLT Corollaries



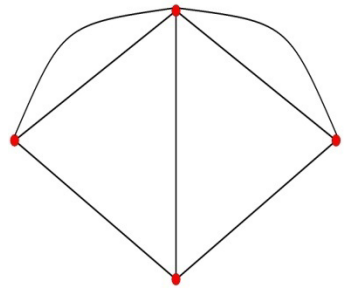
If either $\mathrm{E}[Z_\alpha Z_{-\alpha}|Z_{-\alpha}] \geq 0$ and $\sum_{\alpha,\eta} \mathrm{cov}\left(Z_\alpha^2, Z_\eta^2\right) = o\left(a_N^2\right)$

or X_α are Bernoulli with $\mathrm{E}[X_\alpha] \to 0$ uniformly,

and $\sum_{\alpha \neq \eta} \mathrm{cov}(Z_\alpha, Z_\eta) = o(a_N)$, then

$$\sum_{\alpha \in \Lambda^N} Z_\alpha^N / \sqrt{a_N} \rightsquigarrow \mathcal{N}(0,1).$$

Proof Technique, Extension of Stein's Lemma

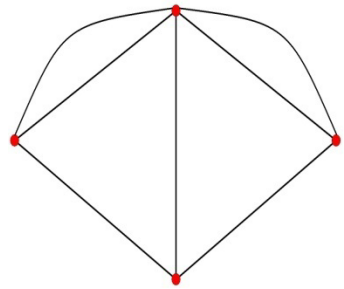


Let $\mathbf{Z}^N := \sum_{\alpha} \mathrm{Z}_{\alpha}^{\mathrm{N}}$ and $\overline{\mathbf{Z}}^N = \mathbf{Z}^N / a_N^{1/2}$.

If $\sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathrm{E}[f'(\overline{\mathrm{Z}}^{\mathrm{N}}) - \overline{\mathrm{Z}}^{\mathrm{N}} f(\overline{\mathrm{Z}}^{\mathrm{N}})]\right| \to 0$

then $\overline{\mathbf{Z}}^N \rightsquigarrow \mathcal{N}(0,1)$

Rest of Proof

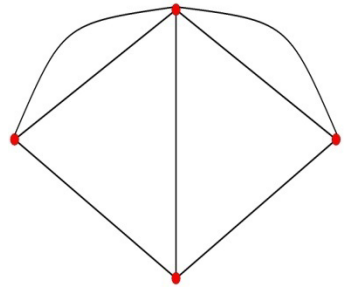


Let $\bar{\mathbf{Z}}_{-\alpha} := \sum_{\eta \notin \mathcal{A}(\alpha)} Z_\eta / a^{1/2}$. A Taylor series expansion around $\bar{\mathbf{Z}}_{-\alpha}$ and some algebra, using (i), gives

$$\left|E\left[f'(\bar{Z}) - \bar{Z}f(\bar{Z})\right]\right| \le \frac{\|f''\|}{2a^{1/2}} \sum_\alpha E\left[|Z_\alpha|\left(\bar{Z} - \bar{Z}_{-\alpha}\right)^2\right] + \left|E\left[f'(\bar{Z})\left(1 - \frac{1}{a^{1/2}}\sum_\alpha Z_\alpha(\bar{Z} - \bar{Z}_{-\alpha})\right)\right]\right| + o(1)$$

Then (ii) and (iii) and more algebra bound the first two expressions.

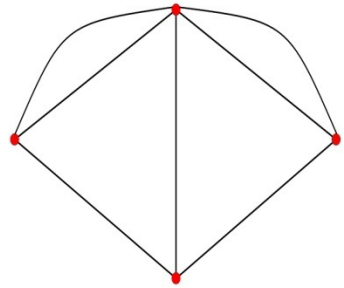
Extensions



Chandrasekhar, Jackson, McCormick, Thiyageswaran

extend theorem to vectors and show nests many other CLTs and show other applications (spillovers, diffusion, …).

Outline



- Background on models of network formation
- SUGMs and Identification
- Estimation and CLT
- Application, Simulations

Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8	16.6	13.1	
Clustering	.29	.05	.15	.07	
Frac. Giant Comp.	.95				
First Eigvalue	10.1				
Homophily	.94				
Avg Path Length	3.5				

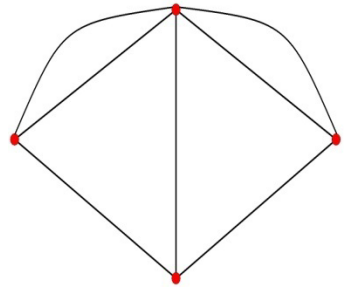
Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8	16.6	13.1	7.2
Clustering	.29	.05	.15	.07	.19
Frac. Giant Comp.	.95				
First Eigvalue	10.1				
Homophily	.94				
Avg Path Length	3.5				

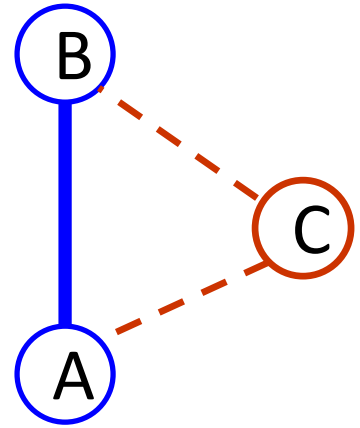
Recreate Networks (Favor, allow covariates)

	Data	Block Model	ERGM	Latent Space	SUGM
Avg. Degree	7.1	7.8	16.6	13.1	7.2
Clustering	.29	.05	.15	.07	.19
Frac. Giant Comp.	.95	.995	.91	.87	.96
First Eigvalue	10.1	9.5	21.6	16.0	9.8
Homophily	.94	0.73	.79	.91	.87
Avg Path Length	3.5	2.9	2.8	3.8	3.2

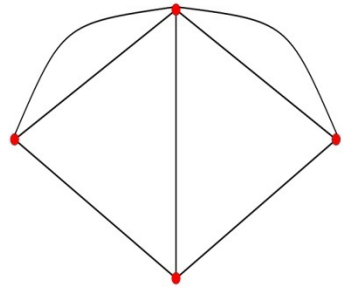
Example: Social Pressure



- Cross caste relationships
  - Are they more likely to occur ``in private'' with no friends in common
  - Or occur with same frequency in embedded relationships?

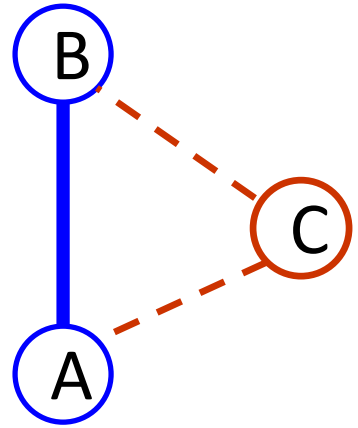


Example: Social Pressure

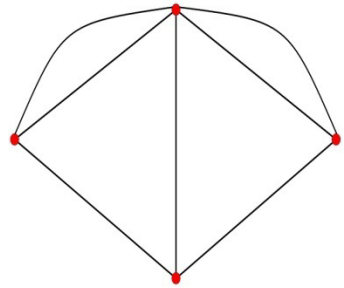


- Examine relative probabilities of triangles diff/same compared to link probabilities diff/same

- Adjust by exponent 3/2…

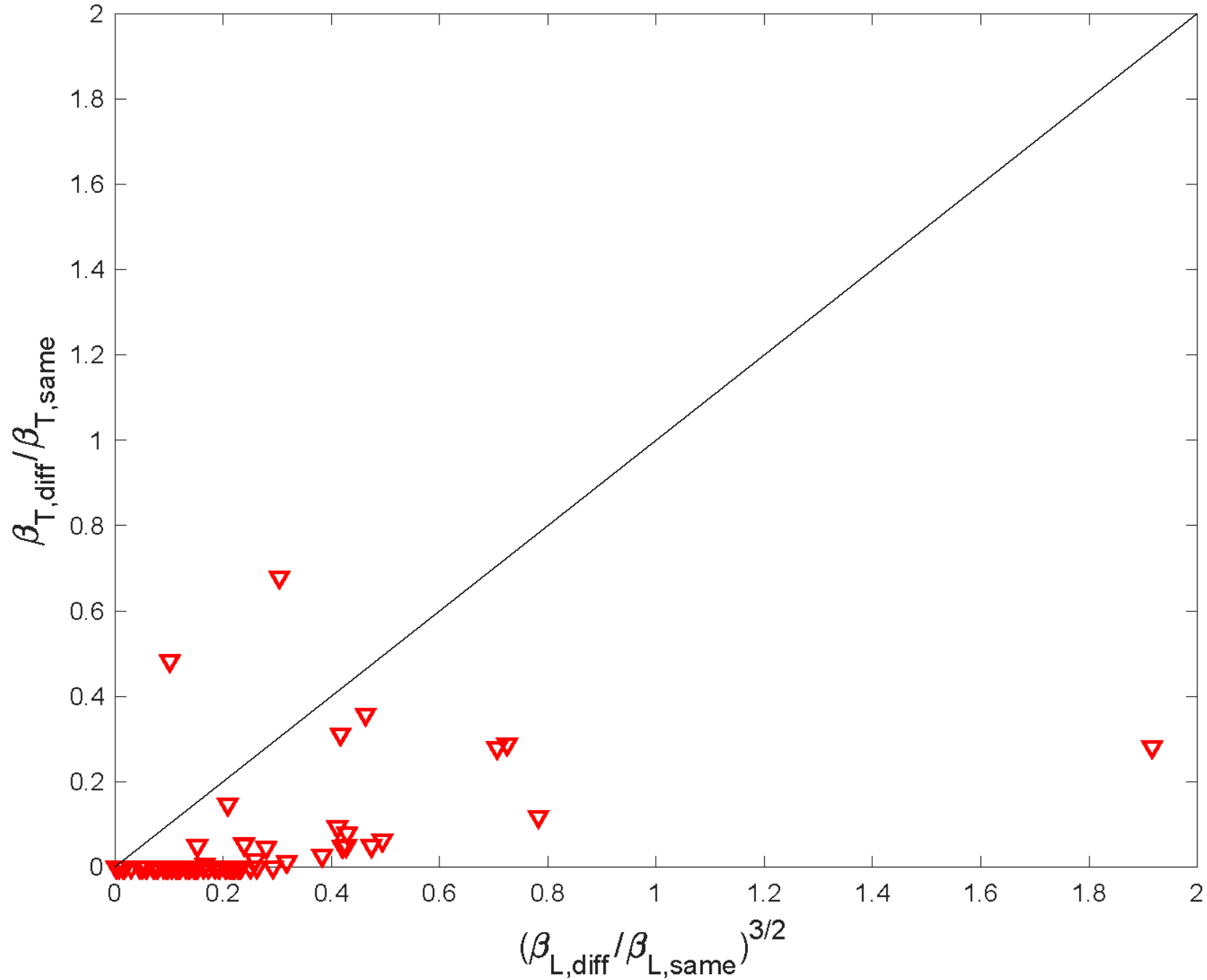


Application: Social Pressure

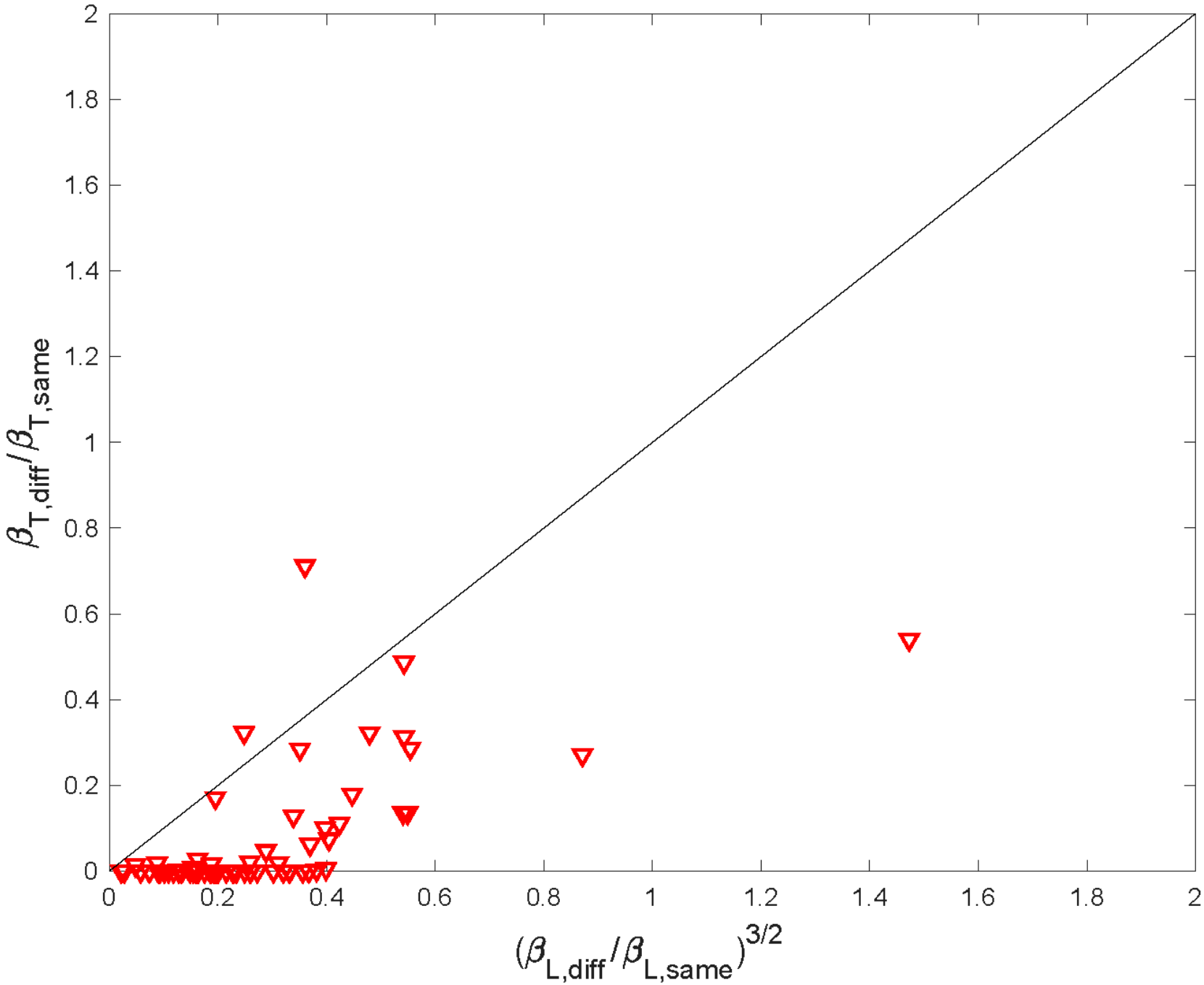


$$\frac{\text{Freq across caste triad}}{\text{Freq within caste triad}} = \frac{F(U(\text{cross triad}))^3}{F(U(\text{within triad}))^3}$$

$$\frac{\text{Freq across caste link}}{\text{Freq within caste link}} = \frac{F(U(\text{cross link}))^2}{F(U(\text{within link}))^2}$$

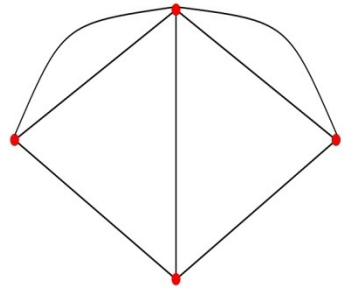


Favor Links
75 villages



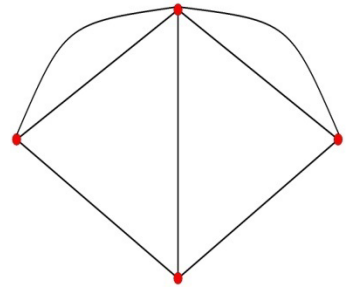
Info Links
75 villages

Summary



- Network formation important for welfare, policy EC
- Estimation problems ERGMs, spatial and block models not up to task when triangles matter.
- Subgraphs directly: identification, estimation
- Match observables, test hypotheses…
- CLT for general correlation structure

Thank You!



- Questions?