A Structural Model of Homophily and Clustering in Social Networks

Angelo Mele

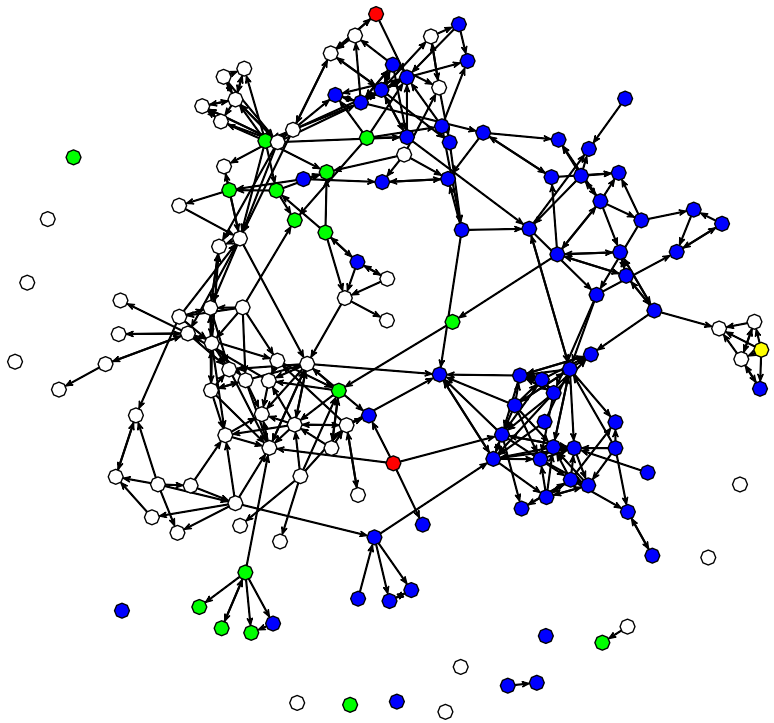
Johns Hopkins University - Carey Business School

07/26/2018
NBER Summer Institute - Labor Studies

web: **http://www.meleangelo.com**
email: **angelo.mele@jhu.edu**

A School friendship network

- School network extracted from Add Health
- School has 150 students, 58.7% females, All grades 7-12.
- Control vbls: race, gender, grade, income of parents



White = Caucasians 42%

Blue = African-Americans 45.3%

Yellow = Asians 0.7%

Green = Hispanics 10.7%

Red = Other race 1.3%

What I have in mind: Mean Girls' cafeteria

- Cliques
- Unobserved heterogeneity
- Hierarchical social system

Stylized facts about social networks

1. Social networks display **homophily**
   $\Rightarrow$ similar people more likely to link
   1. Preferences
   2. Opportunities to meet
   3. Unobserved factors
2. Social networks are usually **sparse**
   $\Rightarrow$ # links $\propto$ # people
3. Social networks display **clustering**
   $\Rightarrow$ people with common friend(s), link with high probability

Stylized facts about social networks

1. Social networks display **homophily**
   $\Rightarrow$ similar people more likely to link
   1. Preferences
   2. Opportunities to meet
   3. Unobserved factors
2. Social networks are usually **sparse**
   $\Rightarrow$ # links $\propto$ # people
3. Social networks display **clustering**
   $\Rightarrow$ people with common friend(s), link with high probability

Need model that matches these facts

Exponential Random Graphs (ERGM)

$$P(g) = \frac{\exp\left[\theta_1 t_1(g,x) + \theta_2 t_2(g,x) + ... + \theta_P t_P(g,x)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[\theta_1 t_1(\omega,x) + \theta_2 t_2(\omega,x) + ... + \theta_P t_P(\omega,x)\right]}$$

- g: network
- x: observable characteristics of people
- θ_p: parameters
- $t_p(g,x)$: sufficient statistics of the network
  1. $t_1(g,x) = \sum_{ij} g_{ij}$ = # links
  2. $t_2(g,x) = \sum_{ijk} g_{ij} g_{jk} g_{ki}$ = # triangles
  3. $t_3(g,x) = \sum_{ij} g_{ij} \mathbf{1}_{\{x_i = x_j = white\}}$ = # links among same race
  4. $t_4(g,x) = \sum_{ij} g_{ij} |x_i - x_j|$ = links weighted by difference in incomes

This paper

1. Model generates sparse networks
2. Equilibrium networks exhibit homophily and clustering
3. Unobserved heterogeneity: latent community structure
4. Estimation: Bayesian approach using exchange algorithm
5. Application to school networks
6. Model replicates properties of the observed network

Related literature

- Model → Weak dependence ERGM (Schweinberger- Handcock 2015)
- Unobserved heterogeneity without clustering (Graham 2017, Dzemski 2017); with clustering (Boucher-Mourifie 2017, Leung 2015)
- Latent community structure without microfoundations (Breza et al 2017, Airoldi et al 2008, Schweinberger-Handcock 2015)
- Exploit subnetworks for computation (Sheng 2016, DePaula et al 2017, Chandrasekhar-Jackson 2016)
- Sparsity and good statistical properties (DePaula et al 2017, Chandrasekhar-Jackson 2016, Menzel 2016)
- Homophily bias in preferences and/or meetings (Currarini et al 2009, Boucher 2015, Mayer-Puller 2008, DePaula 2017, Menzel 2016, Sheng 2016, Ridder-Sheng 2015)
- Peer effects and Lucas critique: (Carrell-Sacerdote-West 2013, Badev 2013, Goldsmith-Pinkham-Imbens 2013, Hsieh-Lee 2015)

Setup and notation

- n players
- K communities

Player i:

- $g_i = \{g_{i1}, ..., g_{in}\}$: **links**
  $g_{ij} = 1$ if i and j are friends
  $g_{ij} = 0$ otherwise
- $x_i = \{x_{i1}, ..., x_{iM}\}$: **observable attributes** (e,g, race, gender)
- $z_i = \{z_{i1}, ..., z_{iK}\}$: **unobservable communities**

Aggregate:

$g = \{g_1, .., g_n\}$: **network** (adjacency matrix)
$x = \{x_1, ..., x_n\}$: **observables**
$z = \{z_1, .., z_n\}$: **communities**

No self loops: $g_{ii} = 0$ for all i

Undirected network: $g_{ij} = g_{ji}$

Directed network can be modeled too (Mele 2017)

Communities and sequential network formation

- Time is discrete: $t = 0, 1, 2, 3, ...$
- At $t = 0$ *Nature* assigns communities

$$Z_i \overset{iid}{\sim} Multinomial(1; \eta_1, ..., \eta_K) \tag{1}$$

Remark: a community contains max $B < n$ people
Remark: each person belongs to one community only

(extensions to multiple communities possible as in Airoldi et al 2008)

- Conditional on $Z = z$, **network g is formed sequentially.** here
- In each period t
  1. Two players i and j meet
  2. Players receive random matching shock ε_{ij}
  3. Players decide whether to form/cut/keep link g_{ij}
     $\rightarrow$ **maximize surplus generated by g_{ij}**

How people meet

Players meet people of same community more often

ASSUMPTION 1 Meeting process is i.i.d. over time.

$$\text{Prob. } i \text{ and } j \text{ meet} = \begin{cases} \rho_w(g_{-ij}, x_i, x_j, n) & \text{if } z_i = z_j, \\ \rho_b(g_{-ij}, x_i, x_j) n^{-\delta} & \text{otherwise} \end{cases} \tag{2}$$

- $0 < \rho_b(g_{-ij}, x_i, x_j) \leq \rho_w(g_{-ij}, x_i, x_j, n) \leq 1$ for any n and (i, j)
- $\delta > 0$
- The sum of ρ over all pairs (i, j) is 1.

Preferences

Players care about direct and common friends (locally)

ASSUMPTION 2. Payoff of player i

$$U_i(g,x,z;\theta) = \underbrace{\sum_{j=1}^{n} g_{ij}\, u(x_i,x_j,z_i,z_j;\alpha,\beta)}_{\text{payoff direct friends}} + \underbrace{\sum_{j=1}^{n} g_{ij} \sum_{r\neq i,j}^{n} g_{jr}g_{ri}\, v(z_i,z_j,z_r;\gamma)}_{\text{payoff common friends}}$$

1. **Symmetry**: $u(x_i,x_j,z_i,z_j;\alpha,\beta) = u(x_j,x_i,z_j,z_i;\alpha,\beta)$
2. **Local transitivity**: $v(z_i,z_j,z_r;\gamma) = \begin{cases} \gamma_k & \text{if } i,j,r \text{ belong to } k \\ 0 & \text{otherwise} \end{cases}$

Matching shocks

Shocks shift preferences and give a logistic model

ASSUMPTION 3

Players receive a matching shock $(\varepsilon_{ij,0}, \varepsilon_{ij,1})$ before updating their links, i.i.d. over time and across pairs.

$$\varepsilon_{ij,1} \sim Gumbel(a, b) \qquad \varepsilon_{ij,0} \sim Gumbel(a, b) \tag{3}$$

Equilibrium: Stationary distribution

PROPOSITION

Under Assumptions 1-3 and conditional on z, the sequence of networks generated by the model is a Markov chain with unique stationary distribution $\pi(g, x, z; \theta)$:

$$\pi(g, x, z; \theta) = \frac{\exp\left[Q(g, x, z; \theta)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[Q(\omega, x, z; \theta)\right]} = \frac{\exp\left[Q(g, x, z; \theta)\right]}{c(\theta, x, z)} \tag{4}$$

where

$$Q(g, x, z; \theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} u(x_i, x_j, z_i, z_j; \alpha, \beta) + \frac{1}{6} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r \neq i,j}^{n} g_{ij} g_{jr} g_{ri} v(z_i, z_j, z_r; \gamma)$$

Equilibrium: Stationary distribution

PROPOSITION

Under Assumptions 1-3 and conditional on z, the sequence of networks generated by the model is a Markov chain with unique stationary distribution $\pi(g, x, z; \theta)$:

$$\pi(g, x, z; \theta) = \frac{\exp\left[Q(g, x, z; \theta)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[Q(\omega, x, z; \theta)\right]} = \frac{\exp\left[Q(g, x, z; \theta)\right]}{c(\theta, x, z)} \tag{4}$$

where

$$Q(g, x, z; \theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} u(x_i, x_j, z_i, z_j; \alpha, \beta) + \frac{1}{6} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r \neq i,j}^{n} g_{ij} g_{jr} g_{ri} v(z_i, z_j, z_r; \gamma)$$

Computational issue: (Mele 2017)
The set $\mathcal{G}$ contains $2^{n(n-1)/2}$ networks
If $n = 20$, then $2^{90} \approx 10^{27}$ terms

Equilibrium: Potential function

Potential function Q summarizes incentives of players
(net of the matching shock)

$$\begin{aligned} Q(g,x,z;\theta) - Q(g',x,z;\theta) &= U_i(g,x,z;\theta) + U_j(g,x,z;\theta) \\ &- \left[U_i(g',x,z;\theta) + U_j(g',x,z;\theta)\right] \end{aligned}$$

- g is network where $g_{ij} = 1$;
- g' is network where $g'_{ij} = 0$ and $g'_{-ij} = g_{-ij}$.

Maxima of Q are pairwise stable (with transfers) networks

Equilibrium properties

PROPOSITION. Equilibrium networks are sparse.

PROPOSITION. Likelihood factorizes into

$$\begin{aligned}\pi(g,x,z;\theta) &= \prod_{k=1}^{K} \frac{\exp\left[Q_{k,k}(g_{k,k},x^{(k)},z;\theta)\right]}{c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)} \\ &\times \left[\prod_{l>k}^{K}\prod_{i\in\mathcal{C}_k}\prod_{j\in\mathcal{C}_l} \frac{\exp\left[2g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta)\right]}{1+\exp\left[2u(x_i,x_j,z_i,z_j;\alpha,\beta)\right]}\right]\end{aligned}$$

REMARK. Model's equilibrium $\Rightarrow$ HERGM
(ERGM with weak dependence, Schweinberger-Handcock 2015)

Model specification

$$
\begin{aligned}
U_i(g, x, z; \theta) = \sum_{j=1}^{n} g_{ij} \Big[& \alpha_{z_i z_j} \\
+ \; & \beta_{white,white} \mathbf{1}_{\{race_i = race_j = white\}} + \beta_{black,black} \mathbf{1}_{\{race_i = race_j = black\}} \\
+ \; & \beta_{hisp,hisp} \mathbf{1}_{\{race_i = race_j = hispanic\}} + \beta_{grade7,grade7} \mathbf{1}_{\{grade_i = grade_j = 7\}} \\
+ \; & \beta_{grade8,grade8} \mathbf{1}_{\{grade_i = grade_j = 8\}} + \beta_{grade9,grade9} \mathbf{1}_{\{grade_i = grade_j = 9\}} \\
+ \; & \beta_{grade10,grade10} \mathbf{1}_{\{grade_i = grade_j = 10\}} + \beta_{grade11,grade11} \mathbf{1}_{\{grade_i = grade_j = 11\}} \\
+ \; & \beta_{grade12,grade12} \mathbf{1}_{\{grade_i = grade_j = 12\}} + \beta_{male,male} \mathbf{1}_{\{gender_i = gender_j = male\}} \\
+ \; & \beta_{female,female} \mathbf{1}_{\{gender_i = gender_j = female\}} \\
+ \; & \beta_{|income_i - income_j|} |income_i - income_j| \\
+ \; & \sum_r g_{jr} g_{rj} \gamma(z_i, z_j, z_r) \Big]
\end{aligned}
\tag{5}
$$

Model specification: Parsimony

If the model has K communities:

- α has $K(K-1)/2$ parameters
- γ has K parameters
- β has P parameters

$\Rightarrow$ **At least** $K(K+1)/2+P$ **parameters to estimate**

Potentially $K = n \Rightarrow n(n+1)/2+P$ parameters.

$$\alpha_{z_i z_j} = \begin{cases} \alpha_k & \text{if } z_i = z_j \text{ and } z_{ik} = 1, \text{ for } k = 1, 2, ..., K \\ \alpha_b & \text{otherwise} \end{cases} \tag{6}$$

Statistical Properties

PROPOSITION. If meeting parameter $\delta > 3$, asy. normal sufficient stats

$$\frac{S_p(g,x,z)}{\sqrt{V\left[S_p(g,x,z)\right]}} \xrightarrow{d} N(0,1) \text{ as } K \to \infty$$

where $V\left[S_p(g,x,z)\right]$ = variance of sufficient stats $S_p(g,x,z)$

Sufficient stats concentrate around their mean
This is good if you want to do maximum likelihood estimation

$$t_p(g,x) = \mathbb{E}_{\theta_0}\left[t_p(\omega,x)\right] \tag{7}$$

Panel data

- K fixed; n fixed.
- We observe the network over time: $g^{(1)}, g^{(2)}, ..., g^{(T)}$
- Conditioning on z, log-likelihood is

$$\log \ell(\alpha, \gamma, g) = \sum_{t=1}^{T} \sum_{k=1}^{K} (\alpha_k e_{kt} + \gamma_k t_{kt}) + \alpha_b e_{bt} - T \log(c(\alpha, \gamma)$$

  - e_{kt} = number of links of community k at time t;
  - e_{bt} = number of links across communities at time t;
  - t_{kt} = number of triangles of community k at time t.
- MLE is consistent and asymptotically normal under standard regularity condition as $T \to \infty$

One network observation

- Suppose you only observe one network
- Asymptotics is more complicated
- Let $(\alpha_k, \gamma_k) = (\alpha, \gamma)$ for all $k = 1, ...K$
- MLE is asymptotically normal under usual regularity conditions and $K \to \infty$.
- Some quibbles: communities should not too different in size

Estimation: Complete likelihood

The likelihood of the model can be written as

$$L\left(g, Z; \theta, \eta, x\right) = \sum_{z \in \mathcal{Z}} P_\theta\left(G = g | X = x, Z = z\right) P_\eta\left(Z = z\right)$$

$$= \sum_{z \in \mathcal{Z}} \prod_{k=1}^{K} \frac{\exp\left[Q_{k,k}(g_{k,k}, x^{(k)}, z)\right]}{c_{k,k}(\mathcal{G}_{k,k}, x^{(k)}; \theta)} \left[\prod_{l>k}^{K} \frac{\exp\left[Q_{k,l}(g_{k,l}, x^{(k)}, x^{(l)}, z)\right]}{c_{k,l}(\mathcal{G}_{k,l}, x^{(k)}, x^{(l)}; \theta)}\right] P_\eta\left(Z = z\right)$$

Estimation: Communities

Probability $P_\eta(Z = z)$ is i.i.d. multinomial for $i = 1, ..., n$

$$Z_i|\eta_1, ..., \eta_K \overset{iid}{\sim} Multinomial(1; \eta_1, ...\eta_K)$$

Priors for η_k

$$\begin{aligned}
\eta_1 &= V_1 \\
\eta_k &= V_k \prod_{j=1}^{k-1}(1 - V_j) \qquad k = 2, 3, 4, ... \\
V_k|\phi &\overset{iid}{\sim} Beta(1, \phi) \qquad k = 1, 2, 3, ... \\
\phi > 0 \quad &\text{and} \quad \sum_{k=1}^{\infty} \eta_k = 1 \ w.p.1
\end{aligned}$$

See Ishwaran and James (2001) and Schweinberger and Handcock (2015)

Estimation: Prior truncation

Mumber of communities is at most K_{max}:

$$\eta_1 = V_1 \tag{8}$$

$$\eta_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \qquad k = 2, 3, 4, ..., K_{max} \tag{9}$$

$$V_k|\phi \overset{iid}{\sim} Beta(1, \phi) \qquad k = 1, 2, 3, ..., K_{max} - 1 \tag{10}$$

$$V_{K_{max}} = 1 \tag{11}$$

$$\phi > 0 \quad \text{and} \quad \sum_{k=1}^{K_{max}} \eta_k = 1 \ w.p.1 \tag{12}$$

Estimation: Priors for payoffs

Priors for payoffs are multivariate normals

$$\begin{aligned} \alpha_b|\mu_b,\Sigma_b &\sim MVN\left(\mu_b,\Sigma_b\right) \\ (\alpha_k,\gamma_k)|\mu_w,\Sigma_w &\sim MVN\left(\mu_w,\Sigma_w\right) \text{ for } k=1,..,K_{max} \\ \beta|\mu_\beta,\Sigma_\beta &\sim MVN\left(\mu_\beta,\Sigma_\beta\right) \end{aligned}$$

Estimation: Posterior

The posterior distribution can be written as follows

$$p(\phi, \mu_w, \Sigma_w, \mu_b, \Sigma_b, \mu_\beta, \Sigma_\beta, \eta, \alpha, \beta, \gamma, z | g, x) \propto$$

$$\propto p(\phi, \mu_w, \Sigma_w, \mu_b, \Sigma_b, \mu_\beta, \Sigma_\beta, \eta, \alpha, \beta, \gamma) \cdot P_\eta (Z = z) \cdot P_\theta (G = g | X = x, Z = z)$$

(prior) **(communities)** **(network likelihood)**

Posterior Sampling algorithm

Exchange algorithm: samples from posterior by sequentially

1. sampling communities
2. sampling parameters
3. sampling networks

In this scheme, z is like a parameter to estimate

Identification and Label Switching Problem

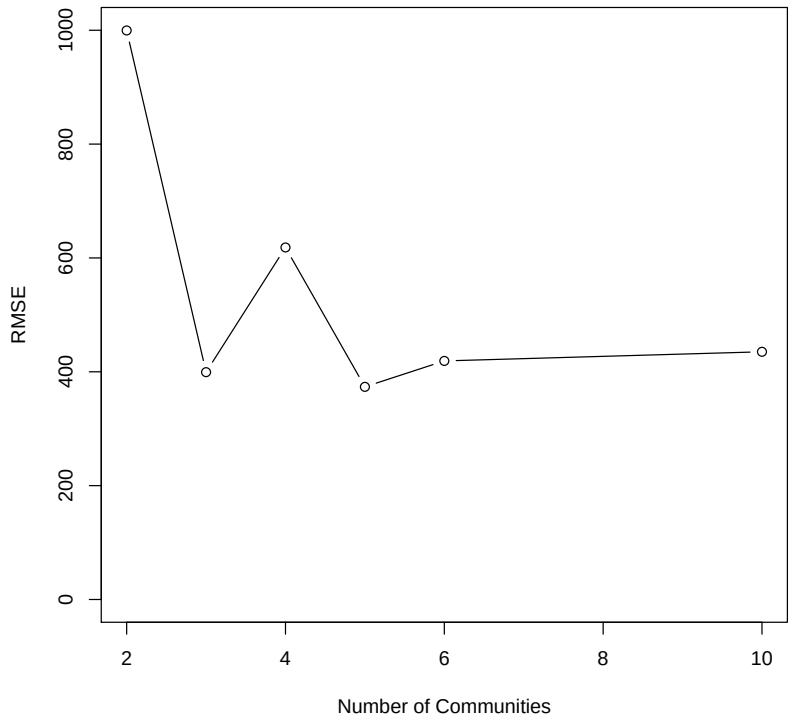
Likelihood invariant to permutations of community labels

1. Use **nonparametric priors**
   $\Rightarrow$ posterior not invariant to permutations of labels
2. Use **relabeling algorithm** of Stephens 2000 and Schweinberger-Handcock 2015 to relabel posterior simulation output algorithm
   - Alternative: ad hoc restrictions that are equivalent to prior restrictions

Choosing number of communities

- Try different $K_{max} = \{2, 3, 4, 5, ...\}$
- Check whether model replicate the number of links and triangles

RMSE Posterior predictions for number of triangles



Structural estimates

Parameter	Post. mean	Post. s.d.	Posterior quantiles 2.5%	50%	97.5%
A. Cost of link					
α_1	-4.070	0.464	-4.888	-4.086	-3.091
α_2	-3.854	0.587	-4.883	-3.895	-2.589
α_3	-2.527	1.049	-4.385	-2.609	-0.316
α_b	-5.754	0.455	-6.636	-5.763	-4.837

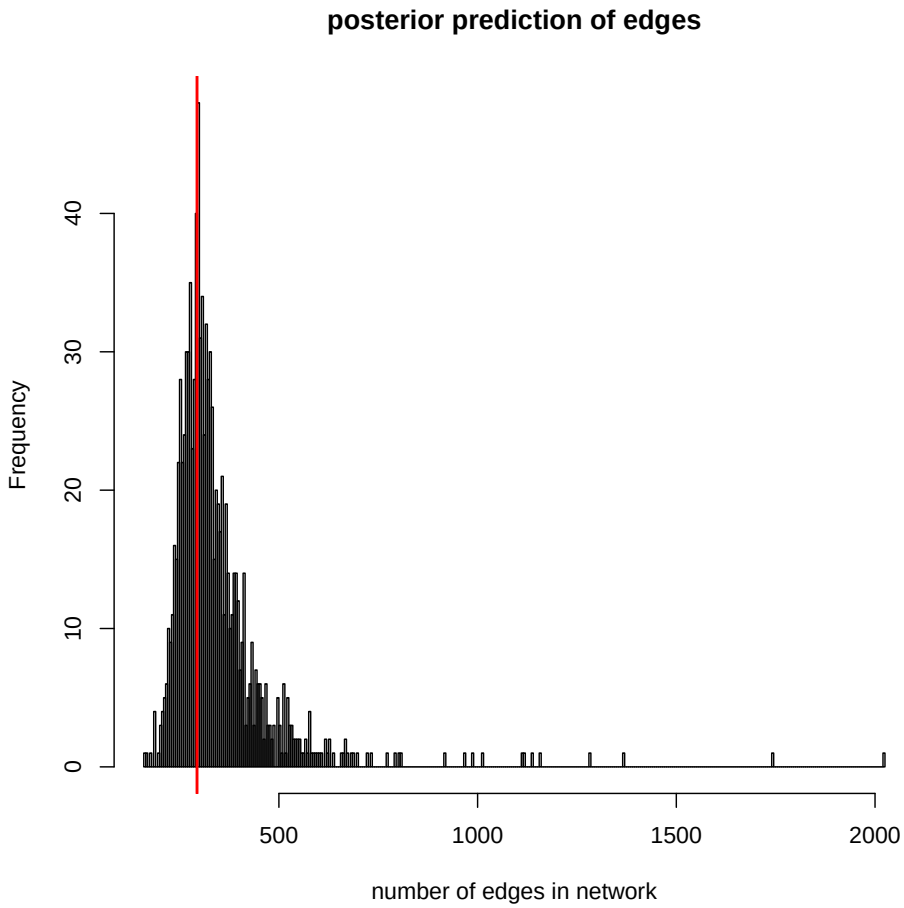
Structural Estimates

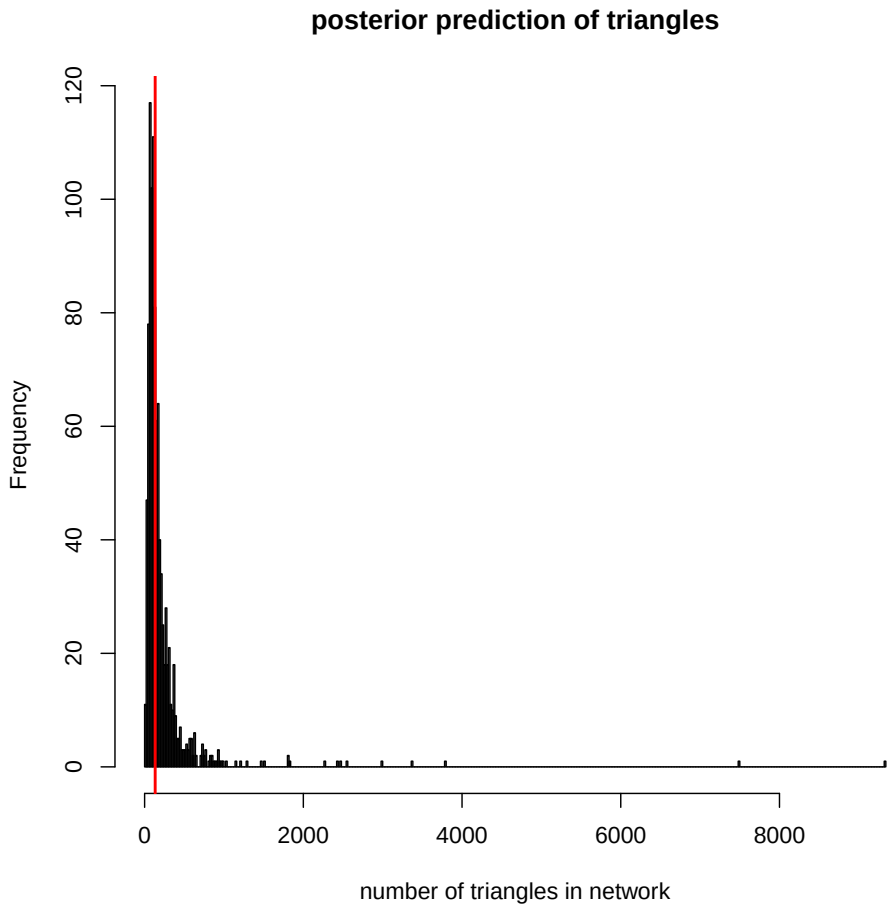
Parameter	Post. mean	Post. s.d.	Posterior quantiles 2.5%	50%	97.5%
B. Payoff from covariates					
$\beta_{white,white}$	1.002	0.246	0.500	1.017	1.420
$\beta_{black,black}$	0.923	0.252	0.424	0.938	1.364
$\beta_{hisp,hisp}$	1.965	0.628	0.789	1.920	3.128
$\beta_{grade7,grade7}$	1.371	0.290	0.685	1.409	1.831
$\beta_{grade8,grade8}$	1.321	0.311	0.627	1.327	1.892
$\beta_{grade9,grade9}$	1.203	0.332	0.568	1.172	1.883
$\beta_{grade10,grade10}$	1.140	0.446	0.207	1.127	1.929
$\beta_{grade11,grade11}$	1.241	0.433	0.249	1.291	1.973
$\beta_{grade12,grade12}$	1.029	0.281	0.435	1.033	1.562
$\beta_{male,male}$	-0.061	0.297	-0.689	-0.029	0.450
$\beta_{female,female}$	-0.170	0.254	-0.725	-0.135	0.294
$\beta_{\lvert income_i - income_j \rvert}$	-0.588	0.278	-1.208	-0.568	-0.136

Structural Estimates

Parameter	Post. mean	Post. s.d.	Posterior quantiles 2.5%	50%	97.5%
C. Payoff from common friends					
γ_1	0.969	0.149	0.644	0.977	1.244
γ_2	1.573	0.562	0.508	1.561	2.738
γ_3	0.995	0.948	-0.889	0.969	2.920

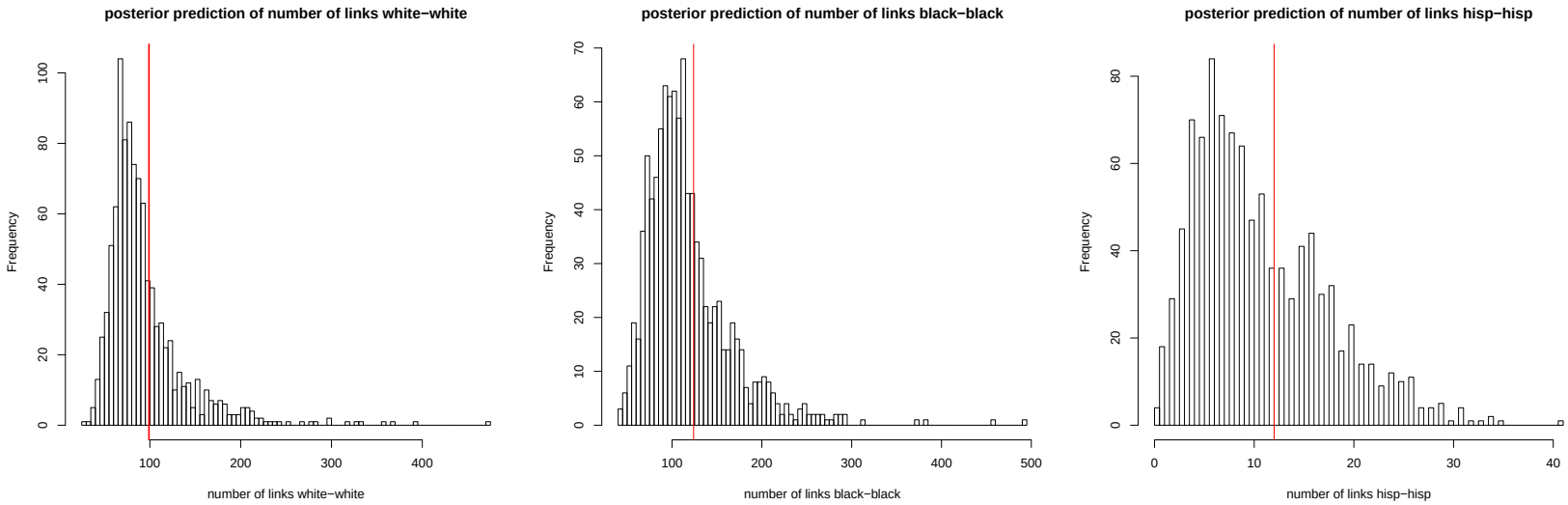
Model fit: links and triangles





Model fit: Racial homophily

Figure: Posterior predictions for racial homophily



Summary

Summary

- Equilibrium model with community structure
- Equilibrium networks are sparse
- Equilibrium networks display homophily and clustering
- Model can replicate features of real-world networks

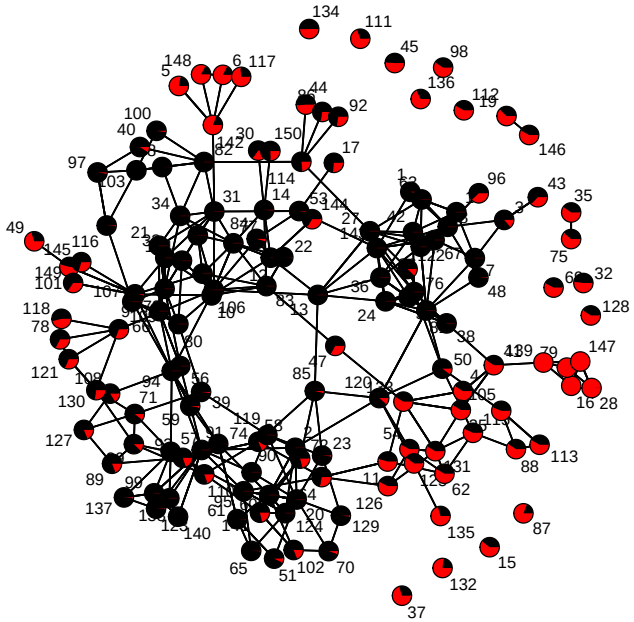
In progress

- Bayesian estimation not practical for large networks
  $\rightarrow$ Approximate Maximum Likelihood methods
- Variational approx (Mele-Zhu 2017) + simulations (Mele 2017)
- Applications: patent collaborations, Venture Capital syndicates

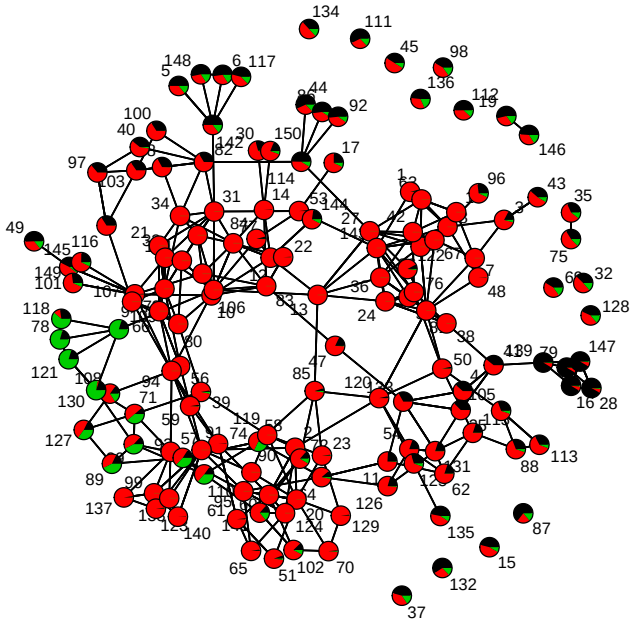
THANK YOU!

More of this at:
web: **http://www.meleangelo.com**
email: **angelo.mele@jhu.edu**

Bonus

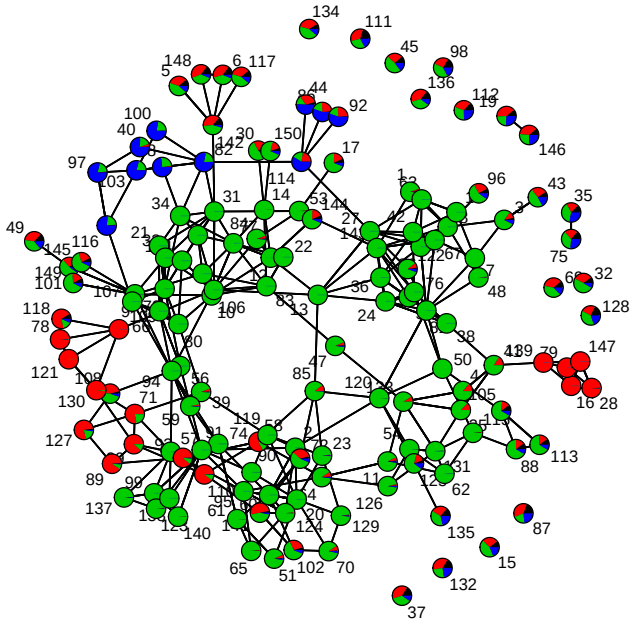


$K_{max} = 2$

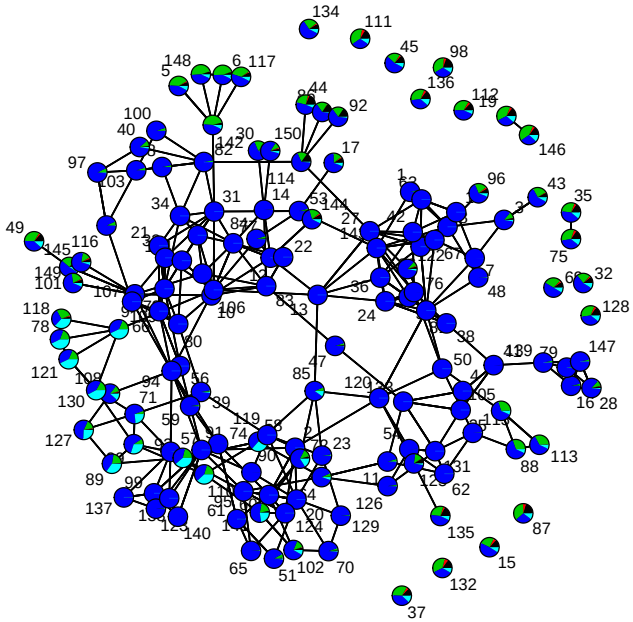


$K_{max} = 3$

Bonus

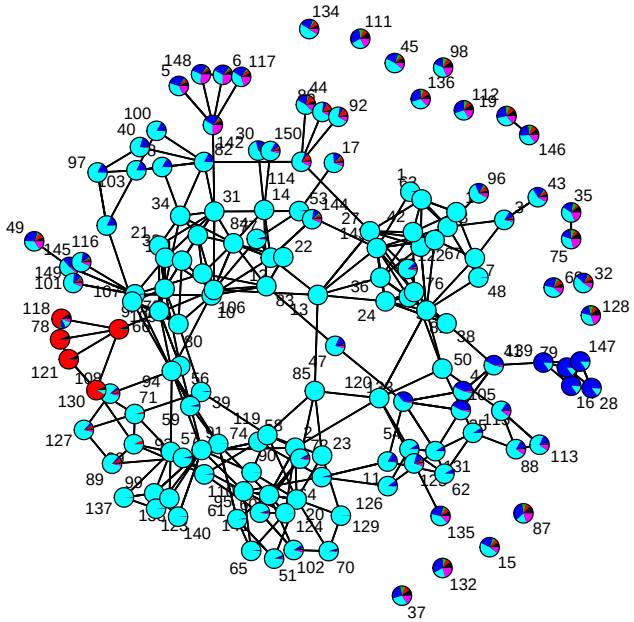


$K_{max} = 4$

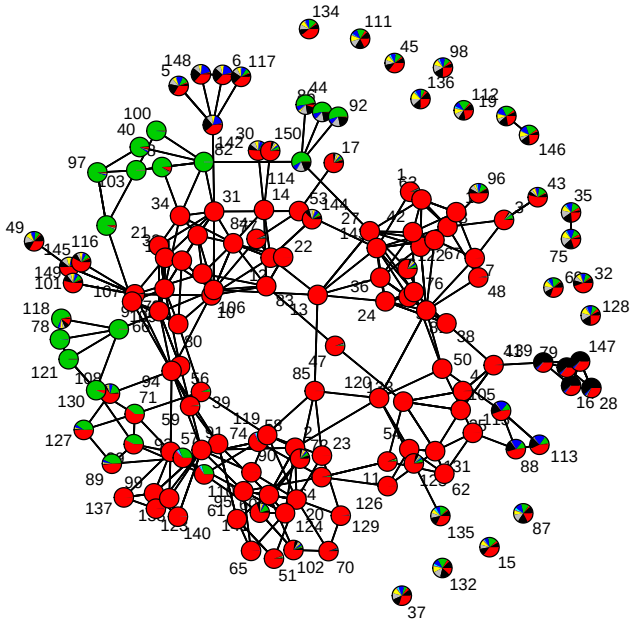


$K_{max} = 5$

Bonus



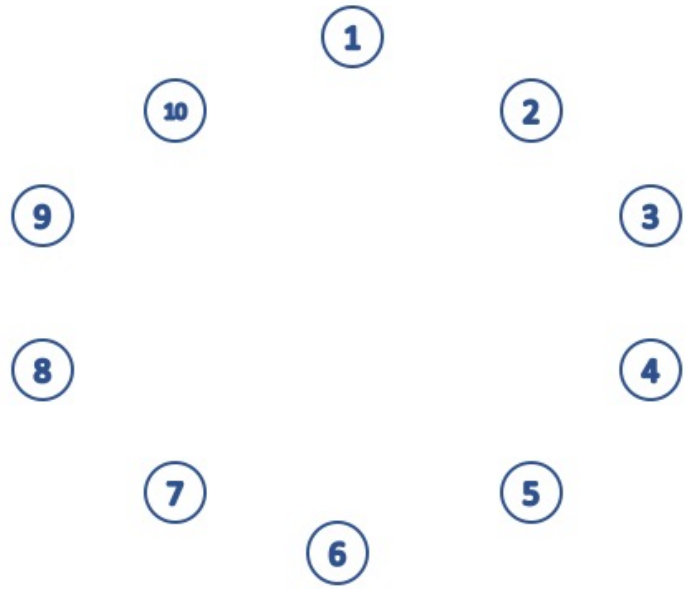
$K_{max} = 6$



$K_{max} = 10$

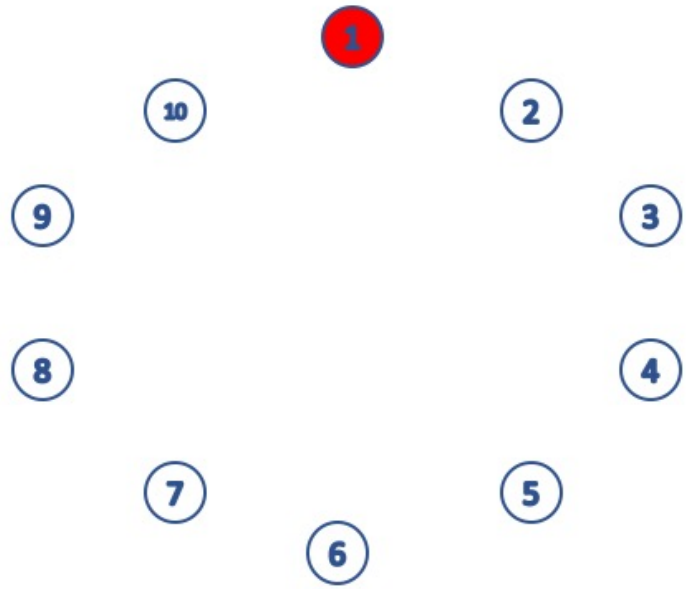
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



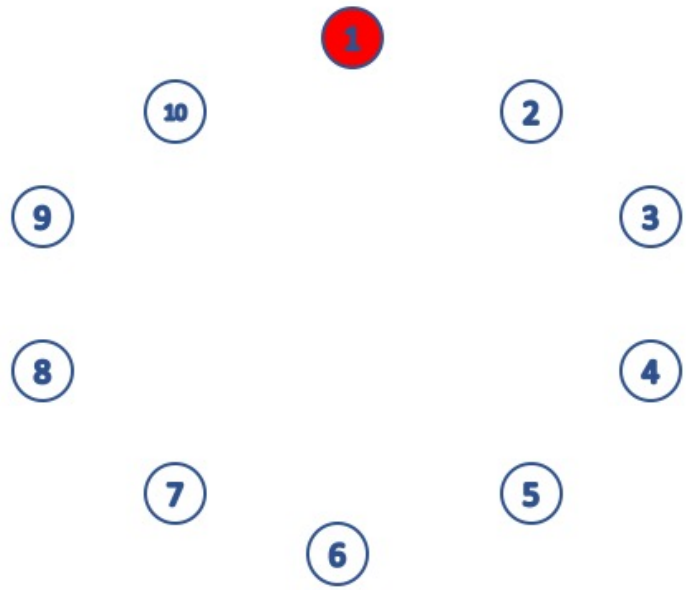
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



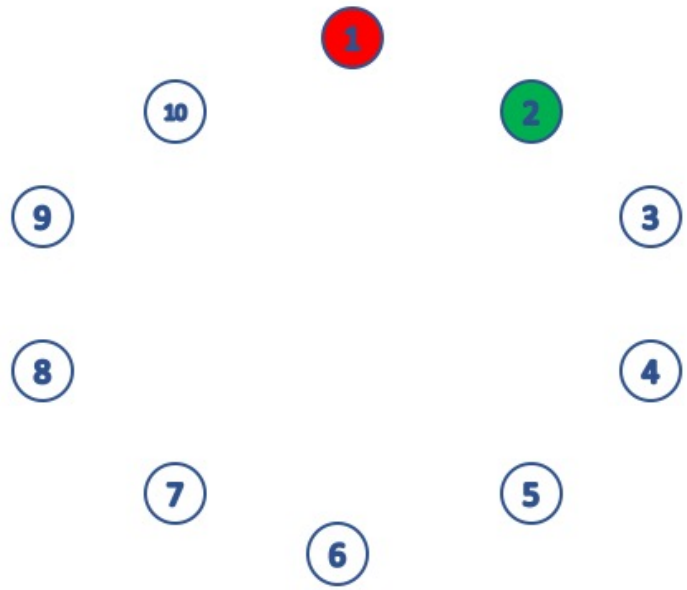
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_2 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



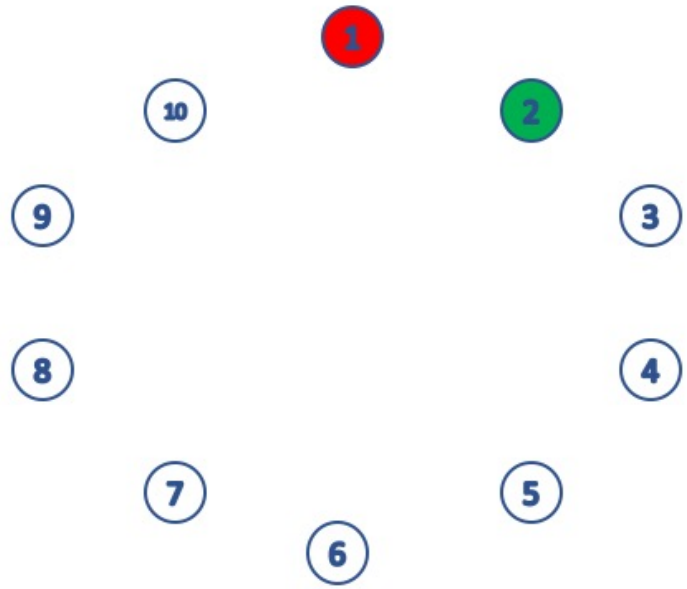
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_2 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



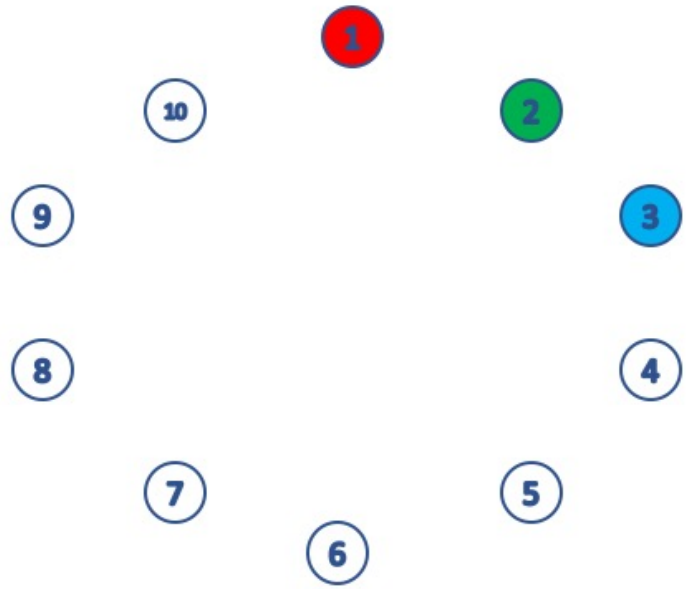
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_2 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_3 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



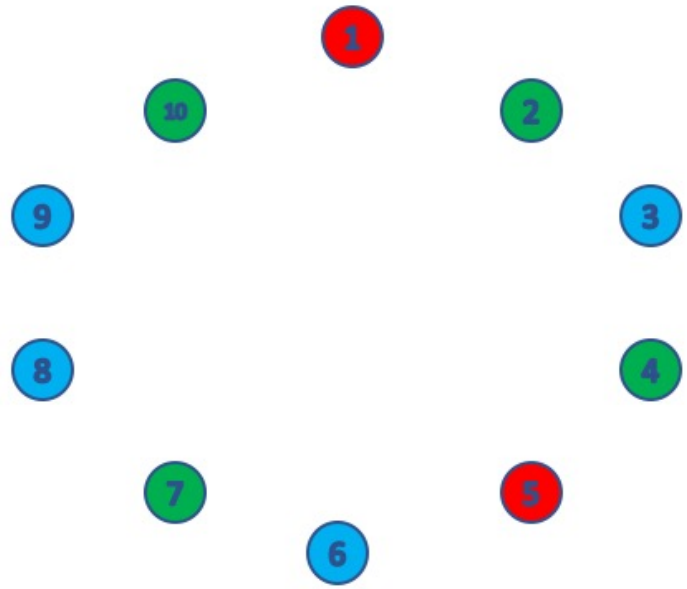
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_2 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_3 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



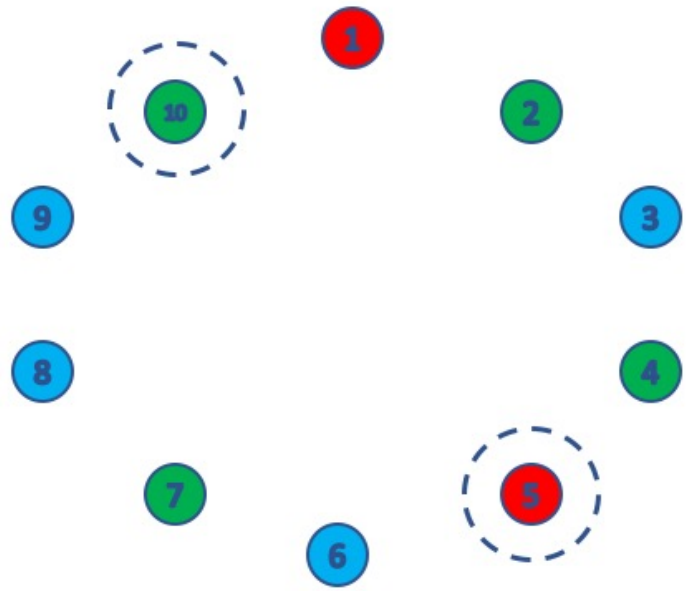
back

$$Z_1 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_2 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_3 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_4 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_5 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_6 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_7 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_8 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_9 \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$
$$Z_{10} \sim Multinomial(1; \eta_1, \eta_2, \eta_3)$$



back

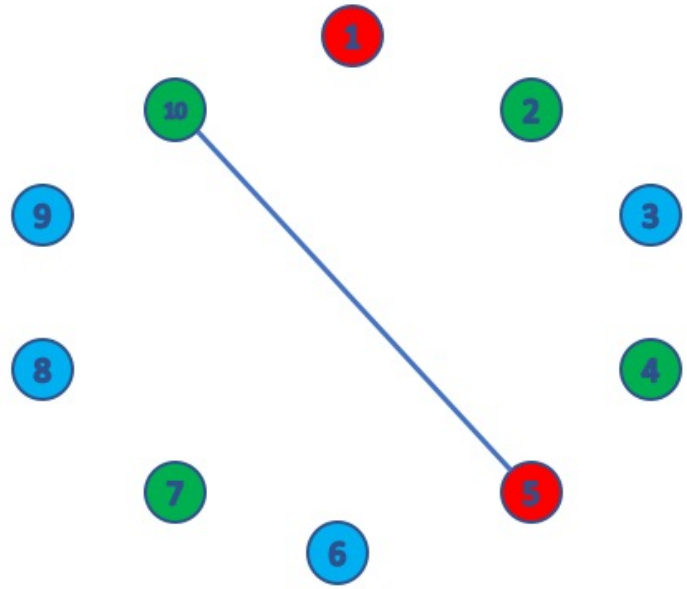
5 and 10 meet



back

5 and 10 meet

5 and 10 form link

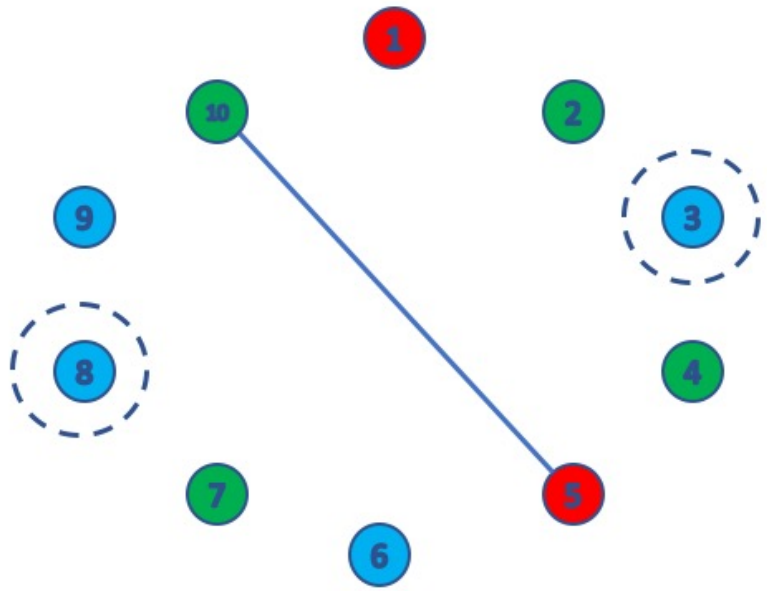


back

5 and 10 meet

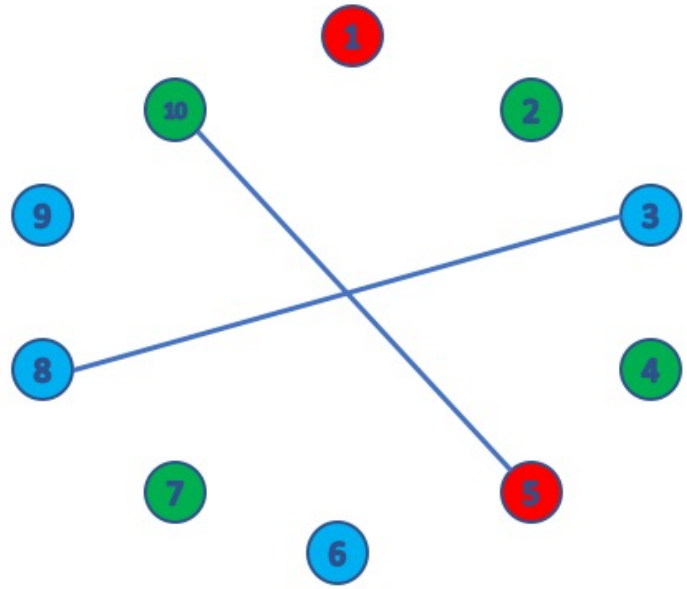
5 and 10 form link

3 and 8 meet



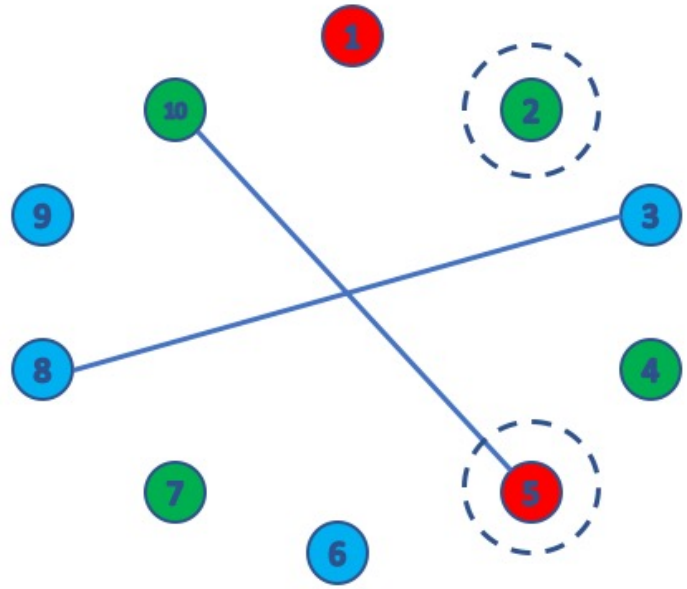
back

5 and 10 meet
5 and 10 form link
3 and 8 meet
3 and 8 form link



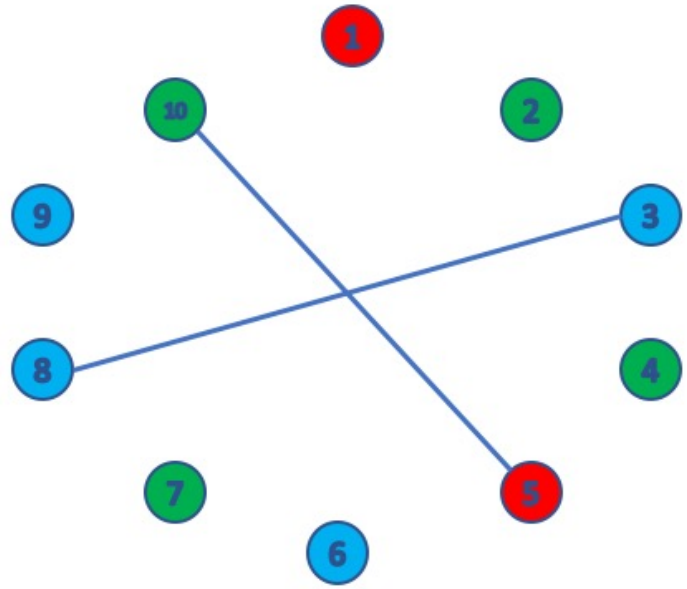
back

5 and 10 meet
5 and 10 form link
3 and 8 meet
3 and 8 form link
2 and 5 meet



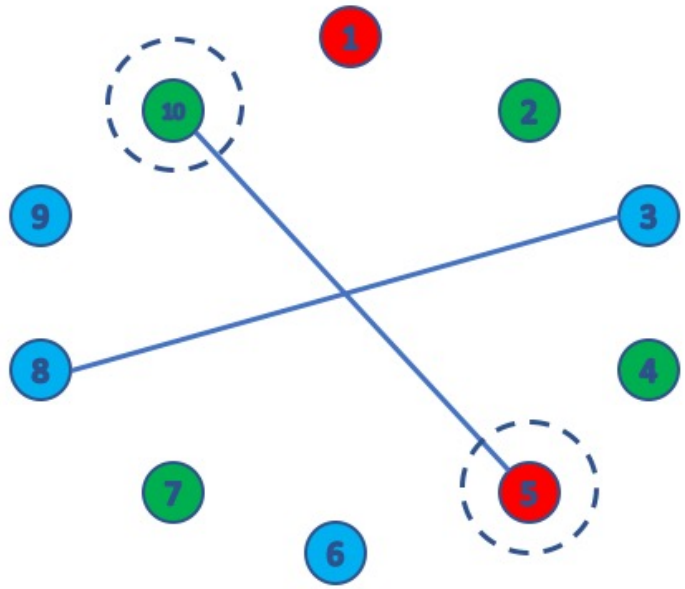
back

5 and 10 meet
5 and 10 form link
3 and 8 meet
3 and 8 form link
2 and 5 meet
2 and 5 do not form link



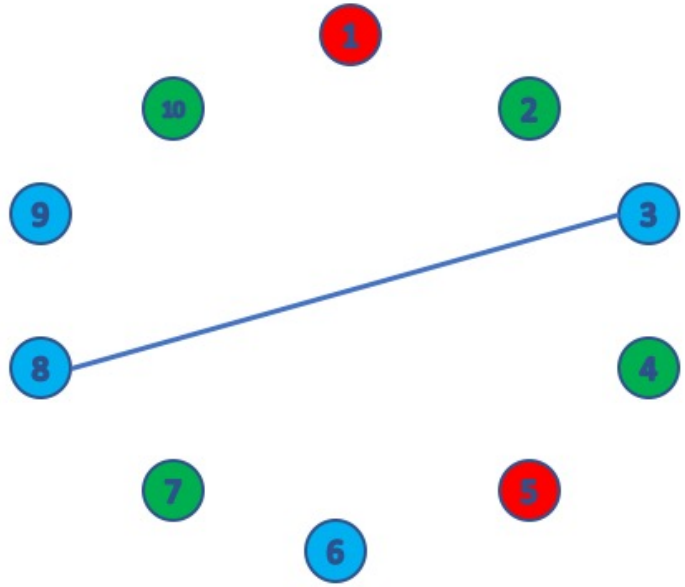
back

5 and 10 meet
5 and 10 form link
3 and 8 meet
3 and 8 form link
2 and 5 meet
2 and 5 do not form link
5 and 10 meet



back

5 and 10 meet
5 and 10 form link
3 and 8 meet
3 and 8 form link
2 and 5 meet
2 and 5 do not form link
5 and 10 meet
5 and 10 cut their link



Relabeling Algorithm

back

- Get MCMC posterior simulation $\{\theta^s, z^s\}_{s=1}^S$
- Algorithm minimizes the loss function

$$L(\xi, \nu(Z)) = \min_{\nu} L_0 \left[\xi, \nu(Z)\right] \tag{13}$$

where

$$L_0 \left[\xi, \nu(Z)\right] = -\log \prod_{i=1}^{n} \xi_{i,c_i} \tag{14}$$

- ξ is $n \times K$ matrix
- $\xi_{i,k}$ = prob that i reported to be in k;
- $\nu(Z)$ = permutation of the community structure Z.

Relabeling Algorithm

back

- Goal of relabeling: ξ that minimizes the posterior expectation of $L\left[\xi, \nu(Z)\right]$.
- In practice posterior expectation approximated by the MC

$$\frac{1}{S}\sum_{s=1}^{S}\min_{\nu_s}\left[L_0\left[\xi, \nu_s(z^s)\right]\right] = \min_{\nu_1,\ldots,\nu_S}\left[\frac{1}{S}\sum_{s=1}^{S}L_0\left[\xi, \nu_s(z^s)\right]\right] \tag{15}$$

- Iterations are:
  1. choose $\hat{\xi}$ to minimize $\sum_{s=1}^{S}\left[L_0\left[\xi, \nu_s(z^s)\right]\right]$ subject to the constraint $\sum_{k=1}^{K_{max}}\xi_{i,k} = 1$ for $i = 1, ..., n$;
  2. for $s = 1, ..., S$ choose ν_s to minimize $L_0\left[\xi, \nu_s(z^s)\right]$
- Step 2 infeasible unless K_{max} very small.
- Use Simulated Annealing to perform the S minimizations in parallel (Stephens 2000, Schweinberger-Handcock 2015)

Posterior sampling

At each iteration: back

1. Sample (θ^*, z^*) from auxiliary distribution $q(\theta^*, z^*|\eta, \theta, z, g)$
2. Sample g^* from $\pi(\omega, x, z^*; \theta^*)$ using MH sampler (Mele 2017)
3. Accept swap (θ, z) to (θ^*, z^*) with prob $\min\{1, exch\}$

$$exch = \frac{P_\eta(Z = z^*)}{P_\eta(Z = z)} \frac{q(\theta, z|\eta, \theta^*, z^*, g)}{q(\theta^*, z^*|\eta, \theta, z, g)}$$
$$\times \frac{\pi(g, x, z^*; \theta^*)}{\pi(g, x, z; \theta)} \frac{\pi(g^*, x, z; \theta)}{\pi(g^*, x, z^*; \theta^*)} \frac{\prod_{k=1}^{K_{max}} p(\alpha_k^*, \gamma_k^*|\mu_w, \Sigma_w)}{\prod_{k=1}^{K_{max}} p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)}$$

- $P_\eta(Z = z)$: community structure,
- $\pi(g, x, z; \theta)$: network likelihood
- $p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)$: priors.

Posterior sampling

At each iteration: back

1. Sample (θ^*, z^*) from auxiliary distribution $q(\theta^*, z^*|\eta, \theta, z, g)$
2. Sample g^* from $\pi(\omega, x, z^*; \theta^*)$ using MH sampler (Mele 2017)
3. Accept swap (θ, z) to (θ^*, z^*) with prob $\min\{1, exch\}$

$$exch = \frac{P_\eta(Z = z^*)}{P_\eta(Z = z)} \frac{q(\theta, z|\eta, \theta^*, z^*, g)}{q(\theta^*, z^*|\eta, \theta, z, g)}$$
$$\times \frac{e^{Q(g,x,z^*,\theta^*)}}{e^{Q(g,x,z;\theta)}} \frac{c(\theta, z)}{c(\theta^*, z^*)} \frac{e^{Q(g^*,x,z;\theta)}}{e^{Q(g^*,x,z^*;\theta^*)}} \frac{c(\theta^*, z^*)}{c(\theta, z)} \frac{\prod_{k=1}^{K_{max}} p(\alpha_k^*, \gamma_k^*|\mu_w, \Sigma_w)}{\prod_{k=1}^{K_{max}} p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)}$$

- $P_\eta(Z = z)$: community structure,
- $\pi(g, x, z; \theta)$: network likelihood
- $p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)$: priors.

Posterior sampling

At each iteration: back

1. Sample (θ^*, z^*) from auxiliary distribution $q(\theta^*, z^*|\eta, \theta, z, g)$
2. Sample g^* from $\pi(\omega, x, z^*; \theta^*)$ using MH sampler (Mele 2017)
3. Accept swap (θ, z) to (θ^*, z^*) with prob $\min\{1, exch\}$

$$exch = \frac{P_\eta(Z = z^*)}{P_\eta(Z = z)} \frac{q(\theta, z|\eta, \theta^*, z^*, g)}{q(\theta^*, z^*|\eta, \theta, z, g)} \times \frac{e^{Q(g,x,z^*,\theta^*)}}{e^{Q(g,x,z;\theta)}} \frac{e^{Q(g^*,x,z;\theta)}}{e^{Q(g^*,x,z^*;\theta^*)}} \frac{\prod_{k=1}^{K_{max}} p(\alpha_k^*, \gamma_k^*|\mu_w, \Sigma_w)}{\prod_{k=1}^{K_{max}} p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)}$$

- $P_\eta(Z = z)$: community structure,
- $\pi(g, x, z; \theta)$: network likelihood
- $p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)$: priors.