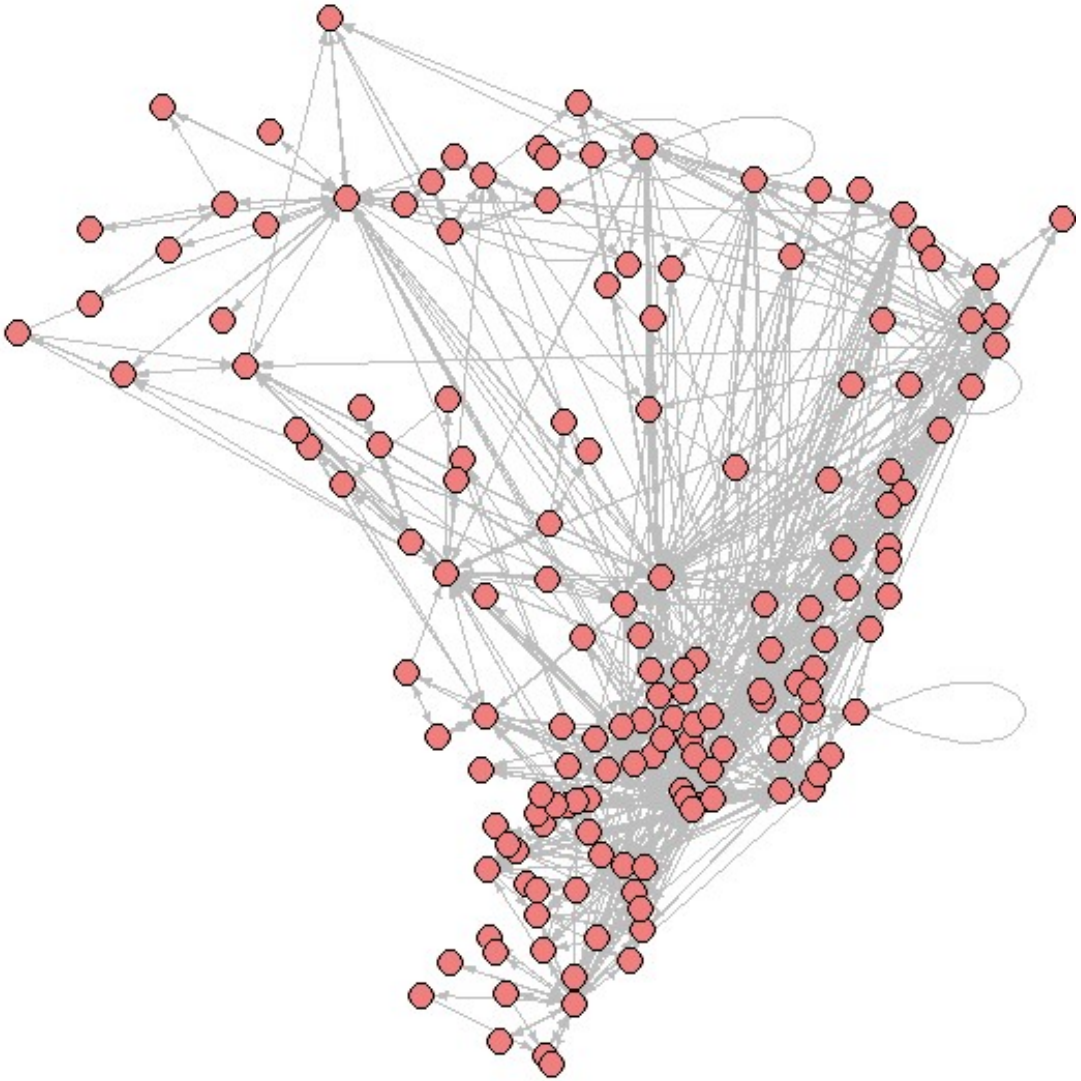
Statistical Inference for network models

Andressa Cerqueira

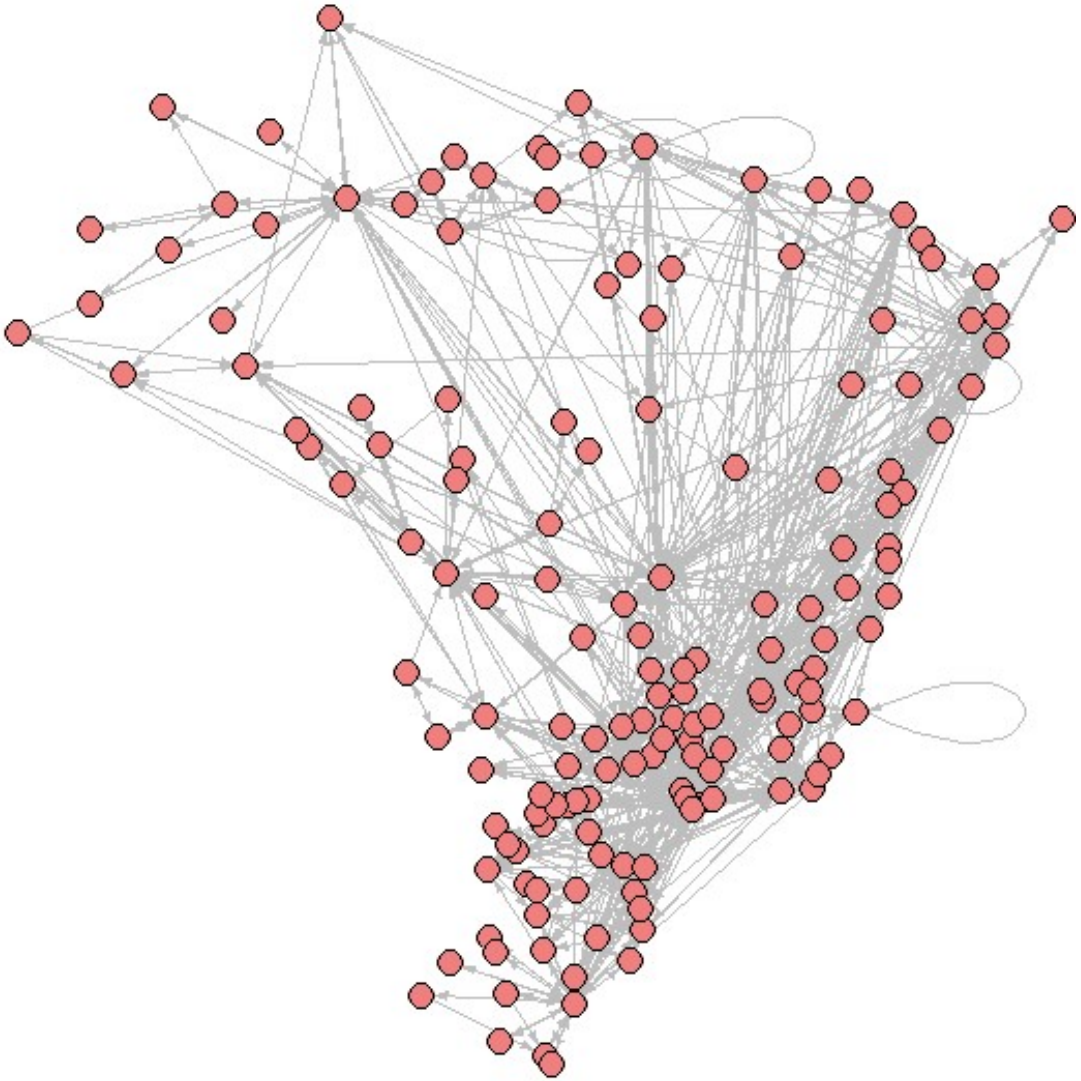
Universidade Federal de São Carlos

September 18, 2020

Motivation: Graphs/Networks



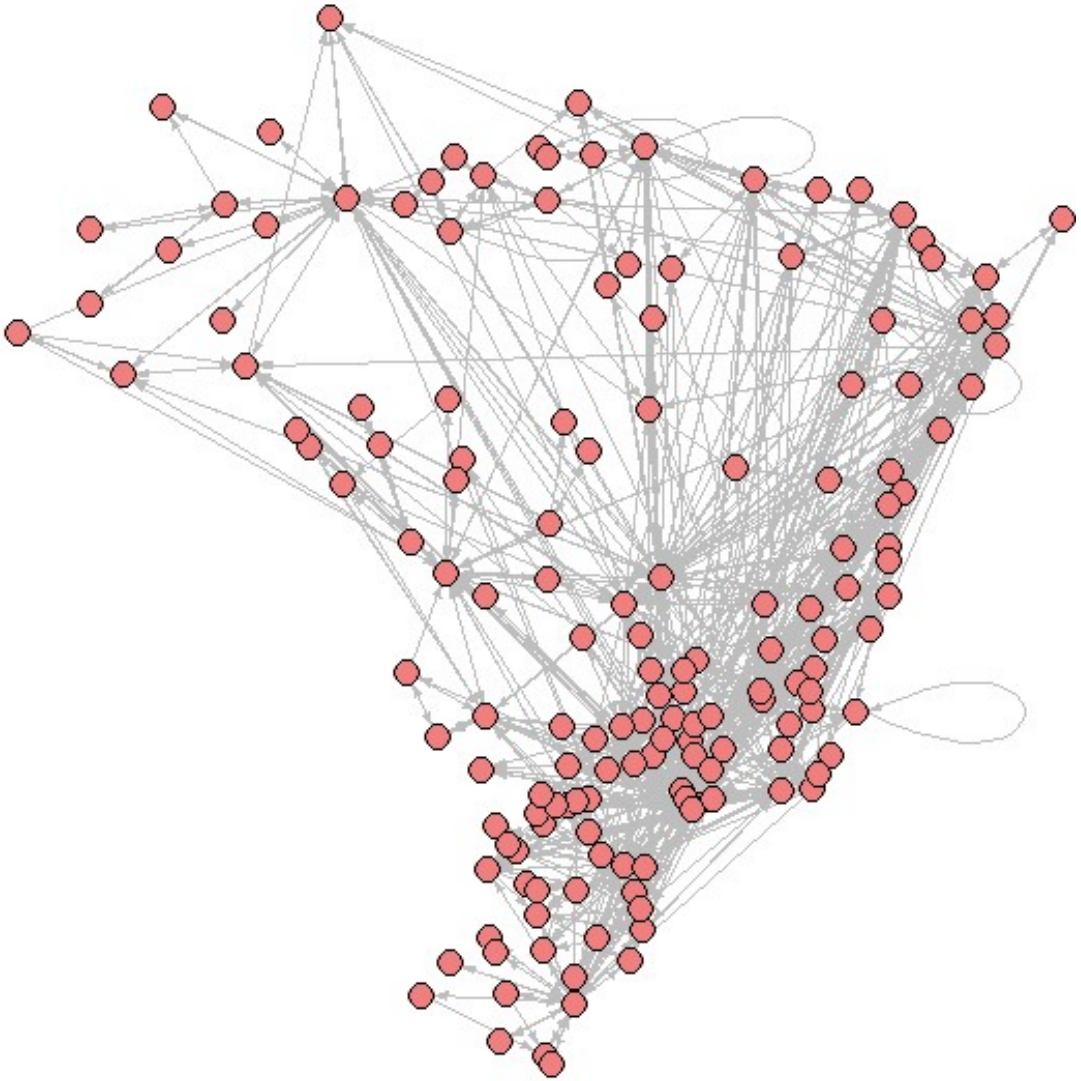
Motivation: Graphs/Networks



Statistics

- statistical model
- inferring the parameters
- hypothesis testing
- clustering nodes

Motivation: Graphs/Networks



Statistics

- statistical model
- inferring the parameters
- hypothesis testing
- clustering nodes

Probability

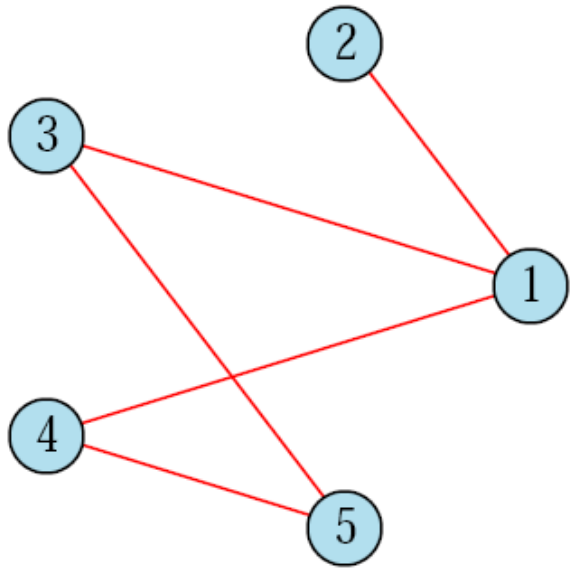
- probabilistic model
- study asymptotic properties of the model

Graph

- A simple graph is a pair (V, E), where V is a finite set of vertices and $E \subseteq V \times V$ is a set of edges;

- The graph can be represented by its adjacency matrix, where

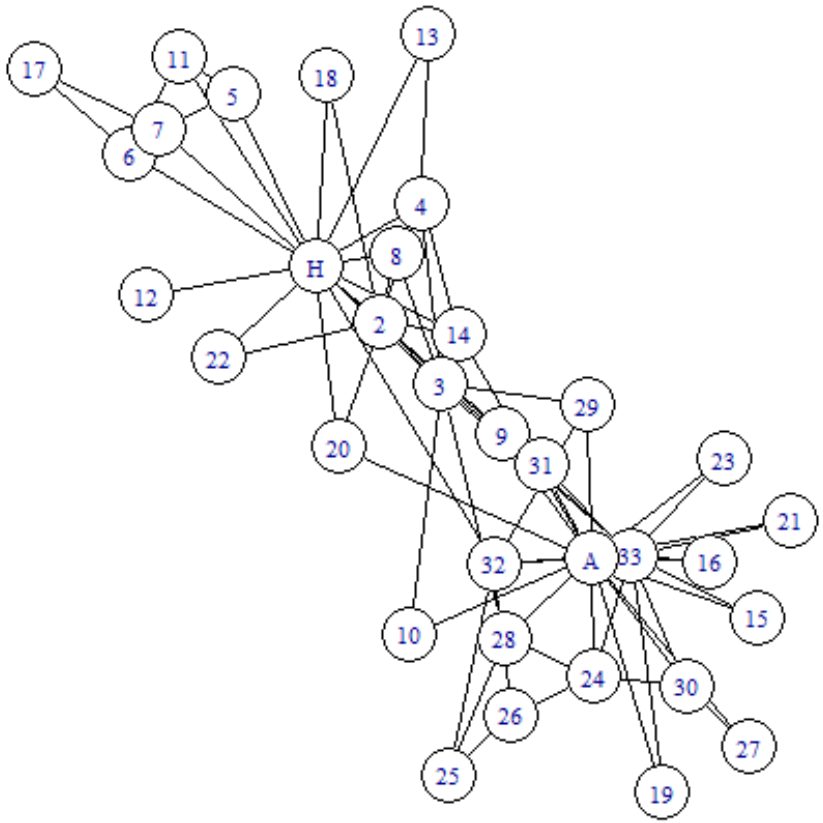
$$x_{ij} = \begin{cases} 1, & \text{if there is an edge between } i \text{ and } j \\ 0, & \text{otherwise} \end{cases}$$



$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

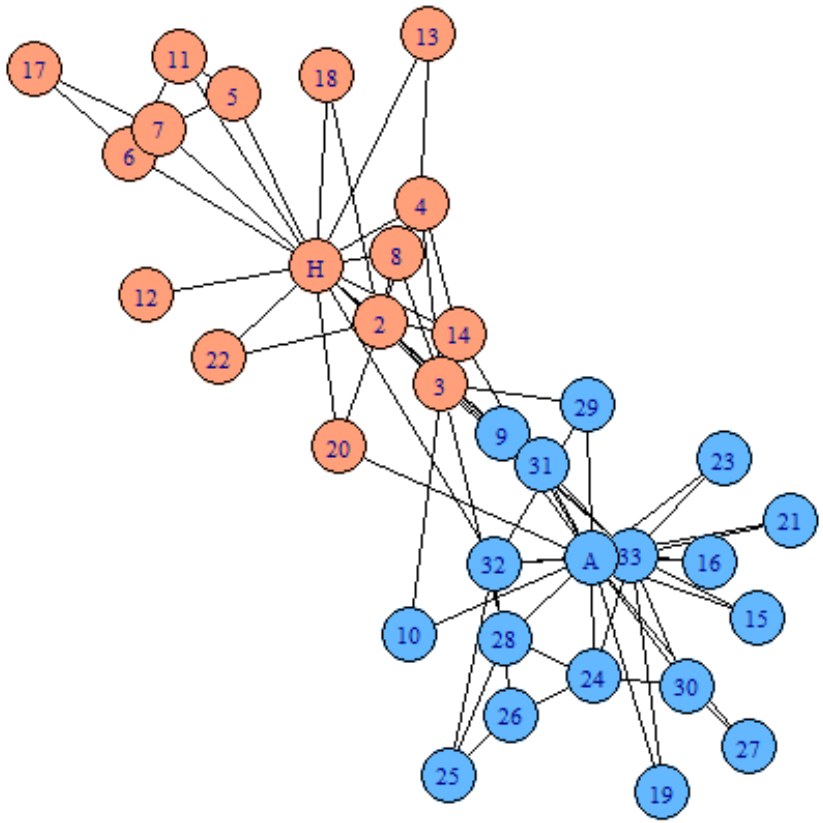
Motivation: Community detection

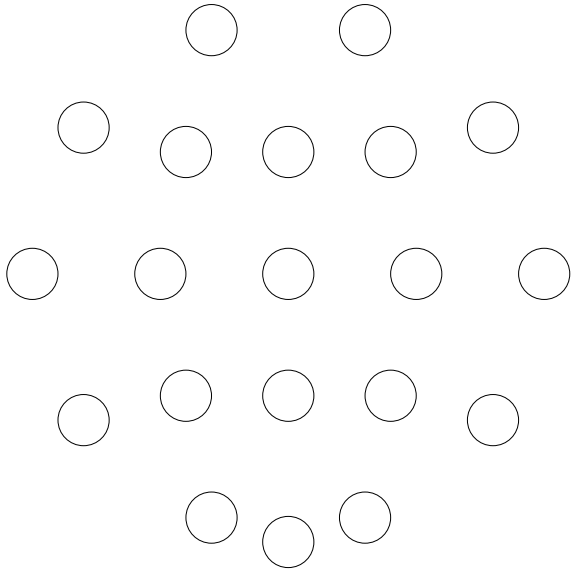
- Zachary's karate club: social relationship between 34 members of a karate club

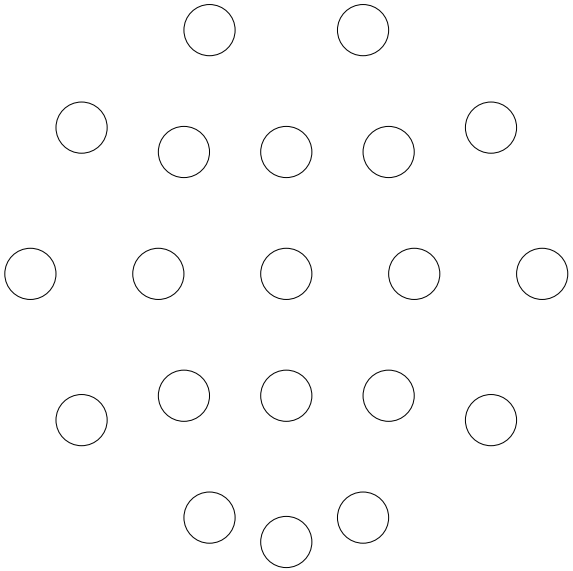


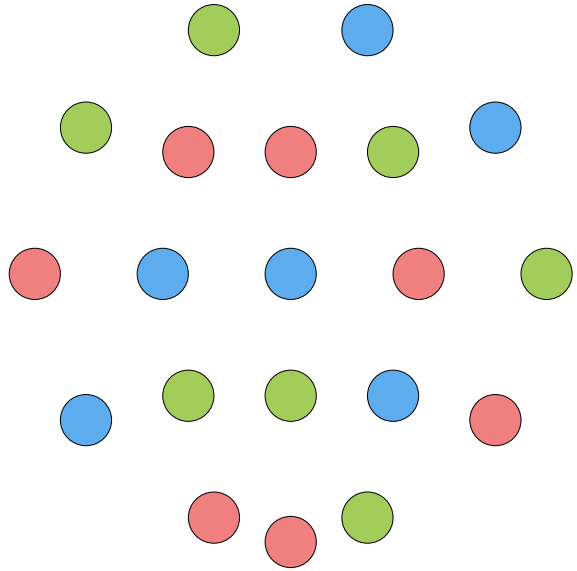
Motivation: Community detection

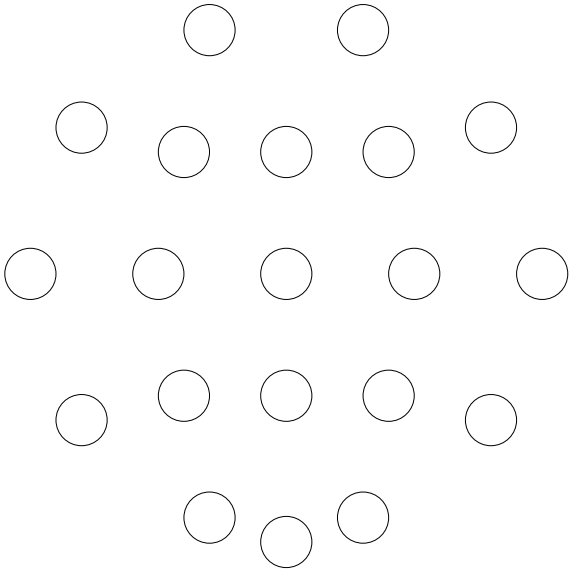
- Zachary's karate club: social relationship between 34 members of a karate club

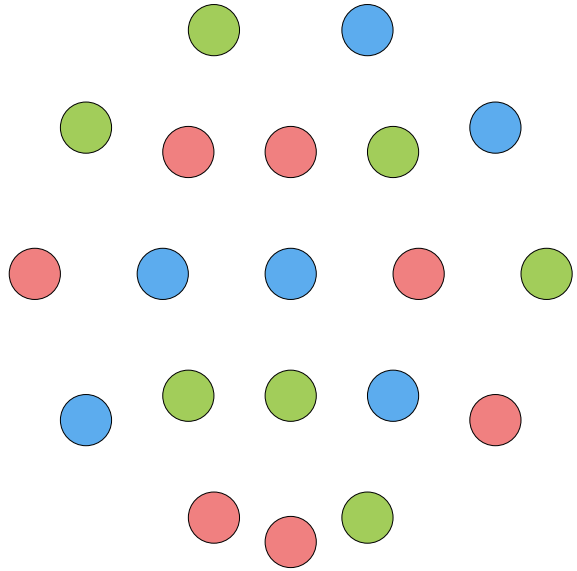


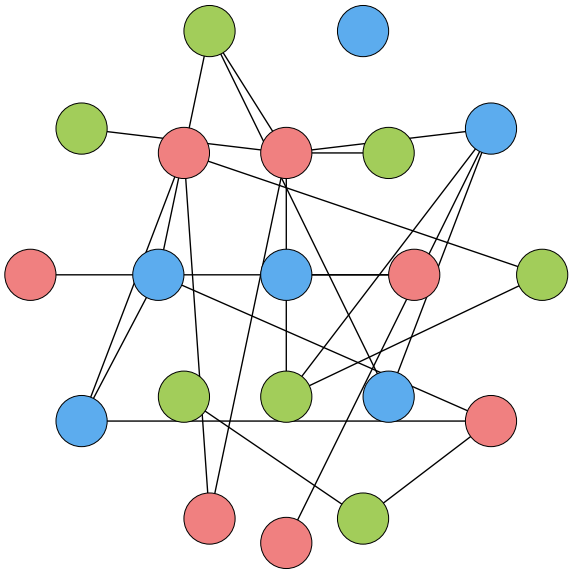


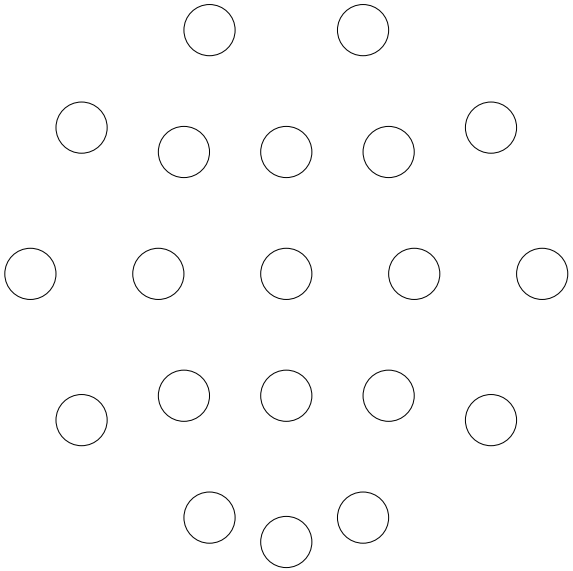


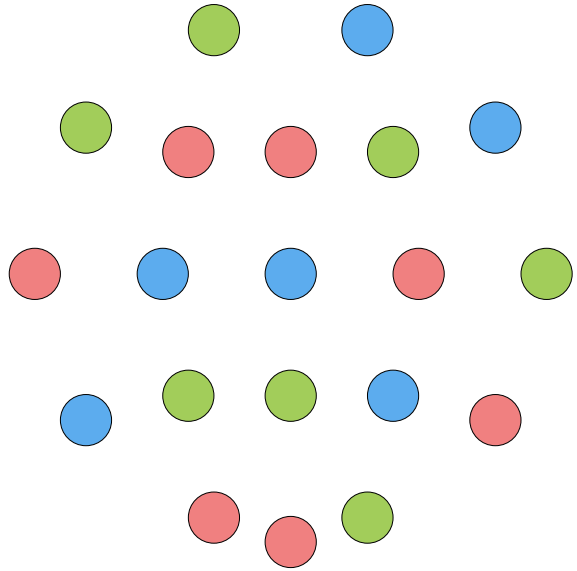


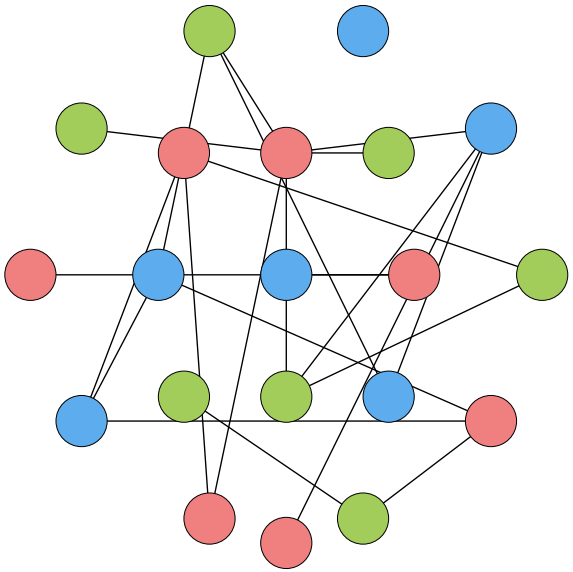


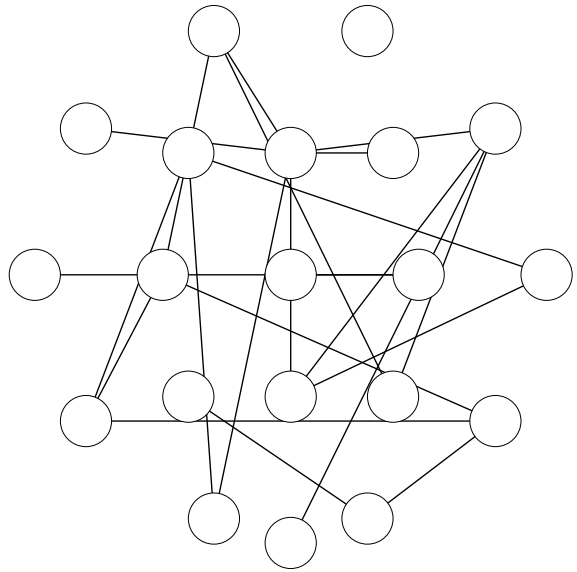












Stochastic Block Models

Latent variables: $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)$, $Z_i \in [k_0] = \{1, \cdots, k_0\}$. k_0 is unknown.
vector of probabilities: $\pi^0 = (\pi^0_1, \cdots, \pi^0_{k_0})$, $\mathbb{P}(Z_i = a) = \pi^0_a$.

Stochastic Block Models

Latent variables: $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)$, $Z_i \in [k_0] = \{1, \cdots, k_0\}$. k_0 is unknown.
vector of probabilities: $\pi^0 = (\pi^0_1, \cdots, \pi^0_{k_0})$, $\mathbb{P}(Z_i = a) = \pi^0_a$.

Observed variables: $\mathbf{X}_{n\times n}$ adjacent matrix of the graph with n vertices. $X_{ij} \in \{0, 1\}$ and

$$X_{ij}|(Z_i = a, Z_j = b) \sim Bernoulli(P^0_{a,b})$$

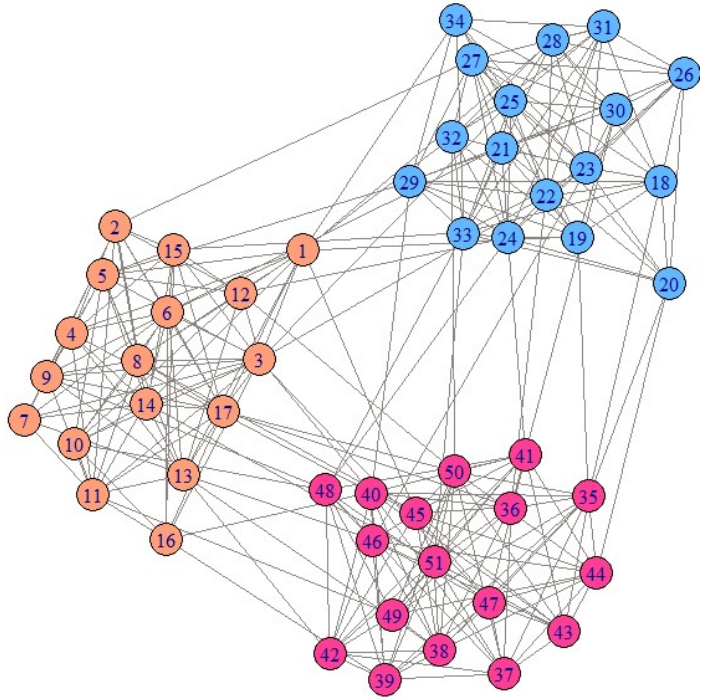
P^0 is a symmetric matrix $k_0 \times k_0$.

Blockmodel: Theory

Example of a network with 50 nodes generated from the SBM

Blockmodel: Theory

Example of a network with 50 nodes generated from the SBM

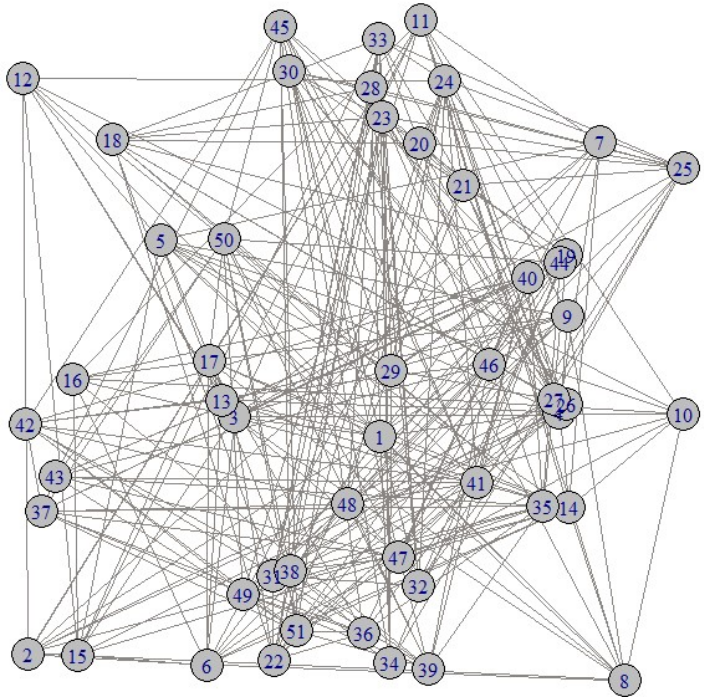


Blockmodel: Data

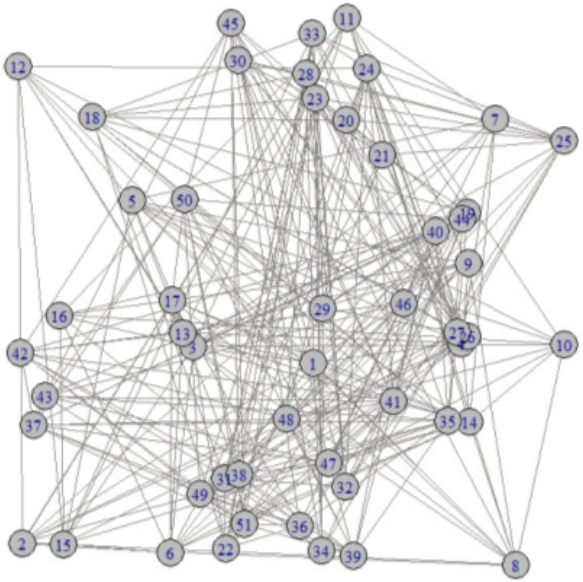
Example of a network with 60 nodes generated from the SBM

Blockmodel: Data

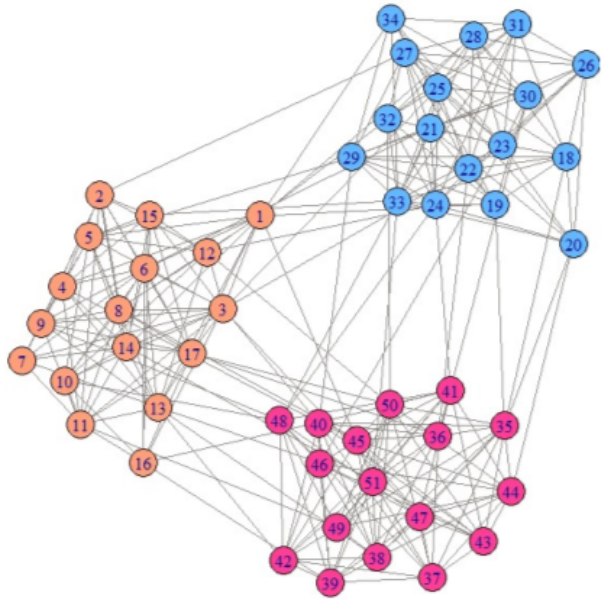
Example of a network with 60 nodes generated from the SBM



Blockmodel: Community detection



Community detection

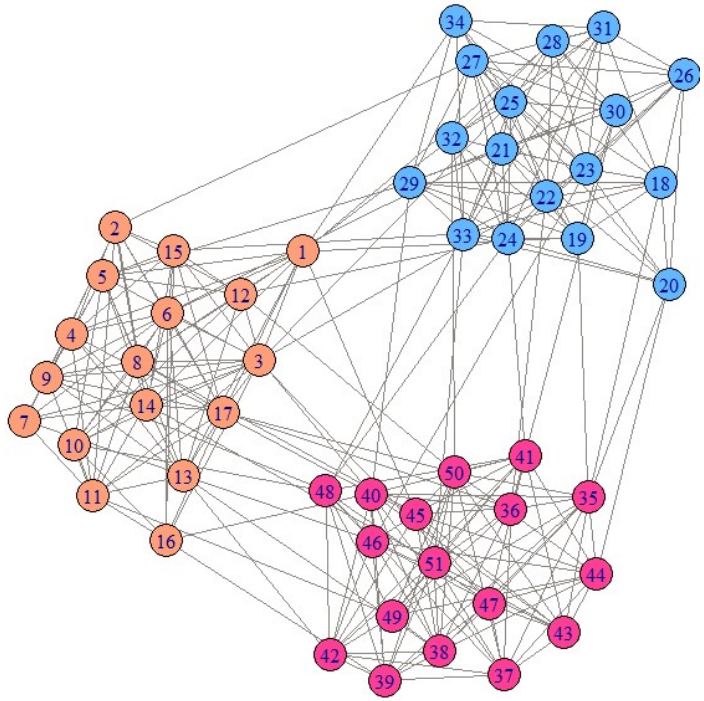


Blockmodel: Balanced Networks

$$n = 50$$

$$\pi = (1/3, 1/3, 1/3)$$

$$P^0 = \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$

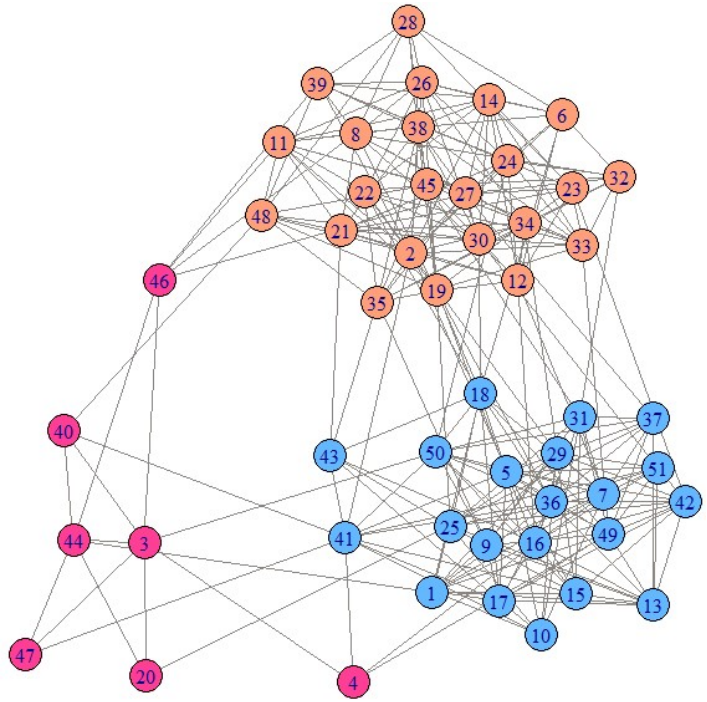


Blockmodel: Unbalanced Networks

$$n = 50$$

$$\pi = (0.4, 0.4, 0.2)$$

$$P^0 = \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$



Blockmodel: Sparsity

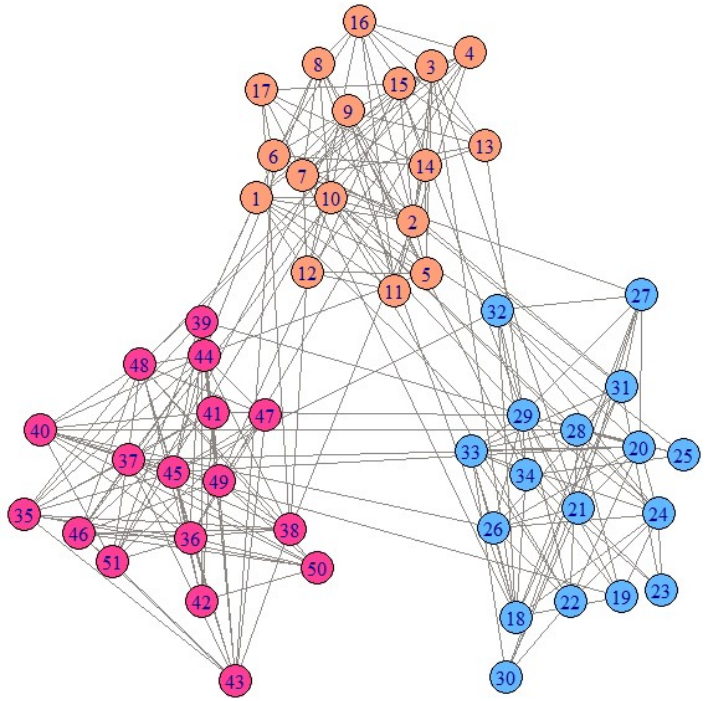
$$n = 50$$

$$\pi = (1/3, 1/3, 1/3)$$

$$P^0 = \rho_n \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$

$$\rho_n = 1$$

$$\rho_n \to 0?$$



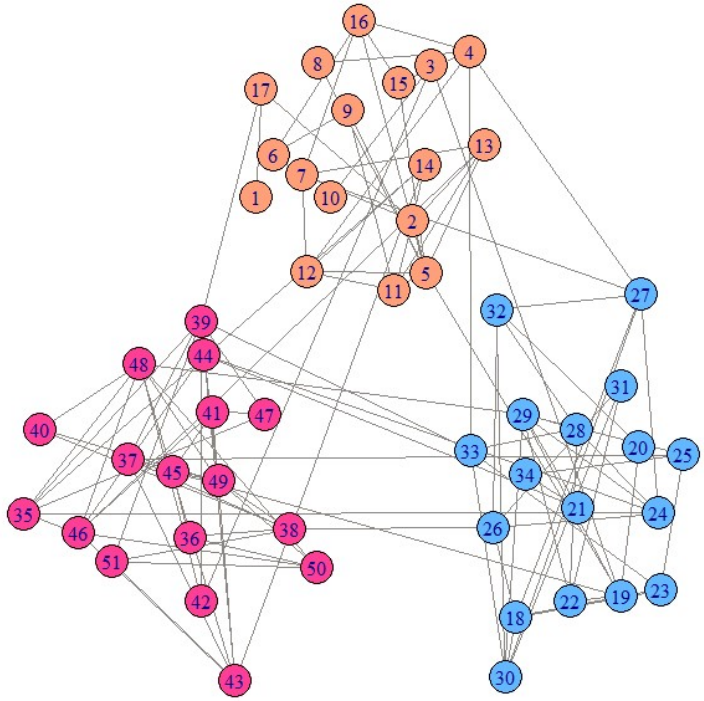
Blockmodel: Sparsity

$$n = 50$$

$$\pi = (1/3, 1/3, 1/3)$$

$$P^0 = \rho_n \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$

$$\rho_n = 0.5$$



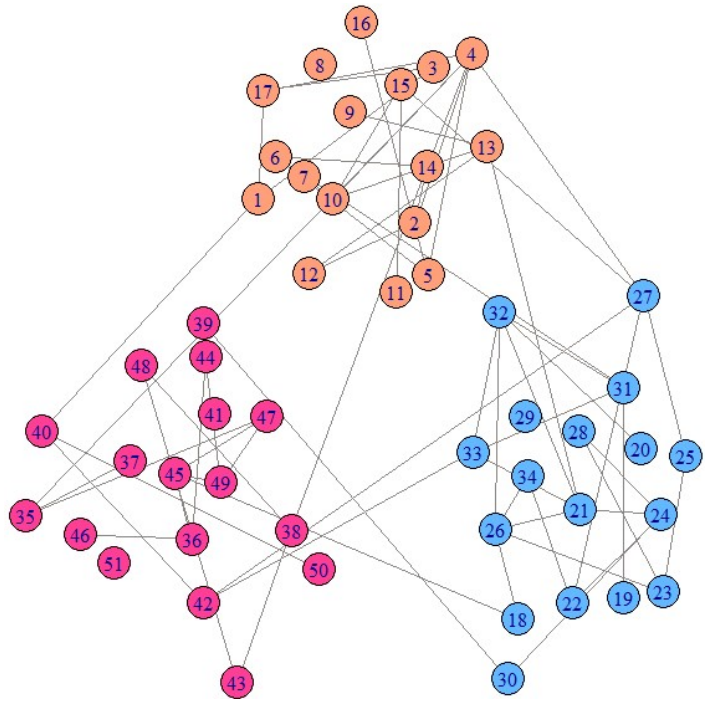
Blockmodel: Sparsity

$$n = 50$$

$$\pi = (1/3, 1/3, 1/3)$$

$$P^0 = \rho_n \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$

$$\rho_n = 0.3$$



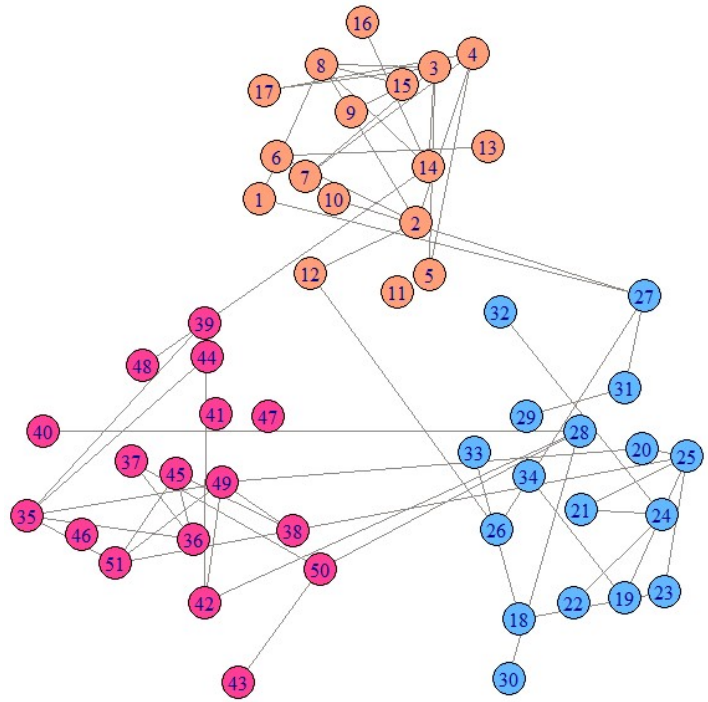
Blockmodel: Sparsity

$$n = 50$$

$$\pi = (1/3, 1/3, 1/3)$$

$$P^0 = \rho_n \begin{bmatrix} 0.5 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.5 \end{bmatrix}$$

$$\rho_n = 0.2$$



Let $(\mathbf{x}_{n\times n}, z_n^0)$ be a sample of the SBM with k_0 blocks and parameter $\theta_0 = (\pi_0, P_0)$.

Let $(\mathbf{x}_{n\times n}, z_n^0)$ be a sample of the SBM with k_0 blocks and parameter $\theta_0 = (\pi_0, P_0)$.

Problems:

1. Estimation of nodes' labels $z_n^0 = (z_1^0, \cdots, z_n^0)$ (Community detection);

Let $(\mathbf{x}_{n\times n}, z_n^0)$ be a sample of the SBM with k_0 blocks and parameter $\theta_0 = (\pi_0, P_0)$.

Problems:

1. Estimation of nodes' labels $z_n^0 = (z_1^0, \cdots, z_n^0)$ (Community detection);

2. Estimation of the parameter $\theta_0 = (\pi_0, P_0)$;

Let $(\mathbf{x}_{n\times n}, z_n^0)$ be a sample of the SBM with k_0 blocks and parameter $\theta_0 = (\pi_0, P_0)$.

Problems:

1. Estimation of nodes' labels $z_n^0 = (z_1^0, \cdots, z_n^0)$ (Community detection);
2. Estimation of the parameter $\theta_0 = (\pi_0, P_0)$;
3. Estimation of the number of blocks k_0 (Model selection problem).

Let $(\mathbf{x}_{n\times n}, z_n^0)$ be a sample of the SBM with k_0 blocks and parameter $\theta_0 = (\pi_0, P_0)$.

Problems:

1. Estimation of nodes' labels $z_n^0 = (z_1^0, \cdots, z_n^0)$ (Community detection);
2. Estimation of the parameter $\theta_0 = (\pi_0, P_0)$;
3. Estimation of the number of blocks k_0 (Model selection problem).

For the labels $\mathbf{z}_n \in [k]^n$ and for $\theta \in \Theta^k$ we write the joint distribution as

$$\mathbb{P}_\theta(\mathbf{X}_{n\times n} = \mathbf{x}_{n\times n}, \mathbf{Z}_n = \mathbf{z}_n)$$

For the labels $\mathbf{z}_n \in [k]^n$ and for $\theta \in \Theta^k$ we write the joint distribution as

$$\mathbb{P}_\theta(\mathbf{X}_{n\times n} = \mathbf{x}_{n\times n}, \mathbf{Z}_n = \mathbf{z}_n) = \prod_{a=1}^{k} \pi_a^{n_a} \prod_{a,b=1}^{k} P_{a,b}^{O_{a,b}/2} (1 - P_{a,b})^{(n_{a,b} - O_{a,b})/2},$$

For the labels $\mathbf{z}_n \in [k]^n$ and for $\theta \in \Theta^k$ we write the joint distribution as

$$\mathbb{P}_\theta(\mathbf{X}_{n\times n} = \mathbf{x}_{n\times n}, \mathbf{Z}_n = \mathbf{z}_n) = \prod_{a=1}^{k} \pi_a^{n_a} \prod_{a,b=1}^{k} P_{a,b}^{O_{a,b}/2} (1 - P_{a,b})^{(n_{a,b} - O_{a,b})/2} ,$$

where the counters $n_a = n_a(\mathbf{z}_n)$, $n_{a,b} = n_{a,b}(\mathbf{z}_n)$ and $O_{a,b} = O_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n})$ are given by

$$n_a(\mathbf{z}_n) = \sum_{i=1}^{n} \mathbb{1}\{z_i = a\} , \qquad 1 \le a \le k$$

$$n_{a,b}(\mathbf{z}_n) = \begin{cases} n_a(\mathbf{z}_n) n_b(\mathbf{z}_n) , & 1 \le a, b \le k \,;\, a \ne b \\ n_a(\mathbf{z}_n)(n_a(\mathbf{z}_n) - 1) & 1 \le a, b \le k \,;\, a = b \end{cases}$$

$$O_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n}) = \sum_{i,j=1}^{n} \mathbb{1}\{z_i = a, z_j = b\} x_{ij} , \quad 1 \le a, b \le k .$$

For $\theta \in \Theta^k$ the marginal distribution of $\mathbf{X}_{n\times n}$ is given by

$$\mathbb{P}_\theta(\mathbf{X}_{n\times n} = \mathbf{x}_{n\times n}) = \sum_{\mathbf{z}_n \in [k]^n} \mathbb{P}_\theta(\mathbf{X}_{n\times n} = \mathbf{x}_{n\times n}, \mathbf{Z}_n = \mathbf{z}_n)$$

Estimation of the number of communities

- Likelihood-based Model Selection Criterion

Wang, Y. R., & Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2), 500-528.

Define the estimator

$$\widehat{k}_0^{\text{PML}} = \arg\max_k \left\{ \sup_{\theta \in \Theta^k} \log \mathbb{P}_\theta(\mathbf{x}_{n\times n}) - \lambda \frac{k(k+1)}{2} n \log n \right\}$$

where λ is a tuning parameter.

$\widehat{k}_0^{\text{PML}}$ is a **weakly** consistent estimator of k_0.

Estimation of the number of communities

- Likelihood-based Model Selection Criterion

Wang, Y. R., & Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2), 500-528.

Define the estimator

$$\widehat{k}_0^{\text{PML}} = \arg\max_k \left\{ \sup_{\theta \in \Theta^k} \log \mathbb{P}_\theta(\mathbf{x}_{n\times n}) - \lambda \frac{k(k+1)}{2} n \log n \right\}$$

where λ is a tuning parameter.

$$\mathbb{P}_\theta(\mathbf{x}_{n\times n}) = \sum_{\mathbf{z}_n \in [k]^n} \mathbb{P}_\theta(\mathbf{x}_{n\times n}, \mathbf{z}_n)$$

Estimation of the number of communities

- Likelihood-based Model Selection Criterion

Wang, Y. R., & Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2), 500-528.

Define the estimator

$$\widehat{k}_0^{\text{PML}} = \arg\max_k \left\{ \sup_{\theta \in \Theta^k} \log \mathbb{P}_\theta(\mathbf{x}_{n\times n}) - \lambda \frac{k(k+1)}{2} n \log n \right\}$$

where λ is a tuning parameter.

$\widehat{k}_0^{\text{PML}}$ is a **weakly** consistent estimator of k_0.

Estimation of the number of communities

- Likelihood-based Model Selection Criterion

Wang, Y. R., & Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2), 500-528.

Define the estimator

$$\widehat{k}_0^{\text{PML}} = \arg\max_k \left\{ \max_{\mathbf{z}_n \in [k]^n} \sup_{\theta \in \Theta^k} \log \mathbb{P}_\theta(\mathbf{x}_{n\times n}, \mathbf{z}_n) - \lambda \frac{k(k+1)}{2} n \log n \right\}$$

where λ is a tuning parameter.

$\widehat{k}_0^{\text{PML}}$ is a **weakly** consistent estimator of k_0.

- Corrected Bayesian Information Criterion

Hu, J., Qin, H., Yan, T., & Zhao, Y. (2019). Corrected Bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, 1-13.

Define the estimator

$$\hat{k}_0^{\text{CBIC}} = \arg\max_k \left\{ \max_{\mathbf{z}_n \in [k]^n} \sup_{\theta \in \Theta^k} \log \mathbb{P}_\theta(\mathbf{x}_{n\times n}, \mathbf{z}_n) - \left[\lambda n \log k + \frac{k(k+1)}{2} \log n \right] \right\}$$

where λ is a tuning parameter.

$\hat{k}_0^{\text{CBIC}}$ is a **weakly** consistent estimator of k_0.

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\mathbb{Q}_k(\mathbf{x}_{n\times n}) = \int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta$$

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\mathbb{Q}_k(\mathbf{x}_{n\times n}) = \underbrace{\int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta}_{\mathbb{E}_{\nu_k}[\,\mathbb{P}_\theta(\mathbf{x}_{n\times n})\,]}$$

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\underbrace{\mathbb{Q}_k(\mathbf{x}_{n\times n})}_{\mathrm{KT}_k(\mathbf{x}_{n\times n})} = \underbrace{\int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta}_{\mathbb{E}_{\nu_k}[\,\mathbb{P}_\theta(\mathbf{x}_{n\times n})\,]}$$

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\underbrace{\mathbb{Q}_k(\mathbf{x}_{n\times n})}_{\mathrm{KT}_k(\mathbf{x}_{n\times n})} = \underbrace{\int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta}_{\mathbb{E}_{\nu_k}[\,\mathbb{P}_\theta(\mathbf{x}_{n\times n})\,]}$$

$\mathrm{KT}_k(\mathbf{x}_{n\times n})$ is called Krichevsky-Trofimov (KT) mixture distribution.

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\underbrace{\mathbb{Q}_k(\mathbf{x}_{n\times n})}_{\mathrm{KT}_k(\mathbf{x}_{n\times n})} = \underbrace{\int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta}_{\mathbb{E}_{\nu_k}[\,\mathbb{P}_\theta(\mathbf{x}_{n\times n})\,]}$$

$\mathrm{KT}_k(\mathbf{x}_{n\times n})$ is called Krichevsky-Trofimov (KT) mixture distribution.

$$\nu_k(\theta) = \left[\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k}\prod_{a=1}^{k}\pi_a^{-1/2}\right]\left[\prod_{a=1}^{k}\prod_{a\leq b\leq k}\frac{1}{\Gamma(\frac{1}{2})^2}P_{a,b}^{-1/2}(1-P_{a,b})^{-1/2}\right]$$

Cerqueira, A., & Leonardi, F. (2020). Estimation of the Number of Communities in the Stochastic Block Model. *IEEE Transactions on Information Theory*.

$$\underbrace{\mathbb{Q}_k(\mathbf{x}_{n\times n})}_{\mathrm{KT}_k(\mathbf{x}_{n\times n})} = \underbrace{\int_{\Theta} \mathbb{P}_\theta(\mathbf{x}_{n\times n})\nu_k(\theta)d\theta}_{\mathbb{E}_{\nu_k}[\,\mathbb{P}_\theta(\mathbf{x}_{n\times n})\,]}$$

$\mathrm{KT}_k(\mathbf{x}_{n\times n})$ is called Krichevsky-Trofimov (KT) mixture distribution.

$$\nu_k(\theta) = \left[\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k}\prod_{a=1}^{k}\pi_a^{-1/2}\right]\left[\prod_{a=1}^{k}\prod_{a\leq b\leq k}\frac{1}{\Gamma(\frac{1}{2})^2}P_{a,b}^{-1/2}(1-P_{a,b})^{-1/2}\right]$$

The penalized estimator is given by

$$\hat{k}_0^{\mathrm{KT}} = \arg\max_k \{\, \log \mathrm{KT}_k(\mathbf{x}_{n\times n}) - \mathrm{pen}(k,n)\,\}\ .$$

How to compute $\mathbf{KT}_k(\mathbf{x}_{n\times n})$?

How to compute $\mathbf{KT}_k(\mathbf{x}_{n\times n})$?

Latouche, P., Birmele, E., & Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12(1), 93-115.

This algorithm is implemented in the R package *mixer*.

How to compute $\mathbf{KT}_k(\mathbf{x}_{n\times n})$?

Latouche, P., Birmele, E., & Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12(1), 93-115.

This algorithm is implemented in the R package *mixer*.

Integrated Likelihood Variational Bayes estimator:

$$\hat{k}_0^{\text{ILvb}} = \arg\max_k \left\{ \log \mathrm{KT}_k(\mathbf{x}_{n\times n}) \right\} .$$

Theorem

Suppose the SBM has k_0 communities and parameters π_0 and P_0. Then, for a penalty function of the form

$$pen(k,n) = \left[\frac{k(k-1)(2k-1)}{12} + \frac{k(k-1)}{2} + \frac{(3+\epsilon)(k-1)}{2}\right]\log n$$

for some $\epsilon > 0$, we have that

$$\hat{k}_0^{KT}(\mathbf{x}_{n\times n}) = k_0$$

eventually almost surely as $n \to \infty$.

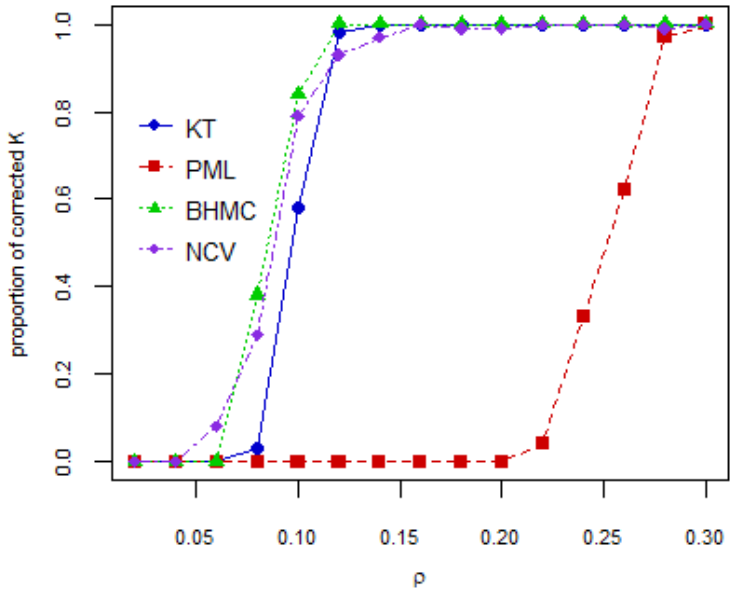
Sparsity

Estimator	$P^0 = \rho_n S^0$	penalty
$\widehat{k}_0^{\text{PML}}$	$\rho_n \to 0$ at rate $\frac{n\rho_n}{\log n} \to \infty$	$\frac{k(k+1)}{2} n \log n$
$\hat{k}_0^{\text{CBIC}}$	$\rho_n \to 0$ at rate $\frac{n\rho_n}{\log n} \to \infty$	$n \log k + \frac{k(k+1)}{2} \log n$
$\hat{k}_0^{\text{KT}}$		

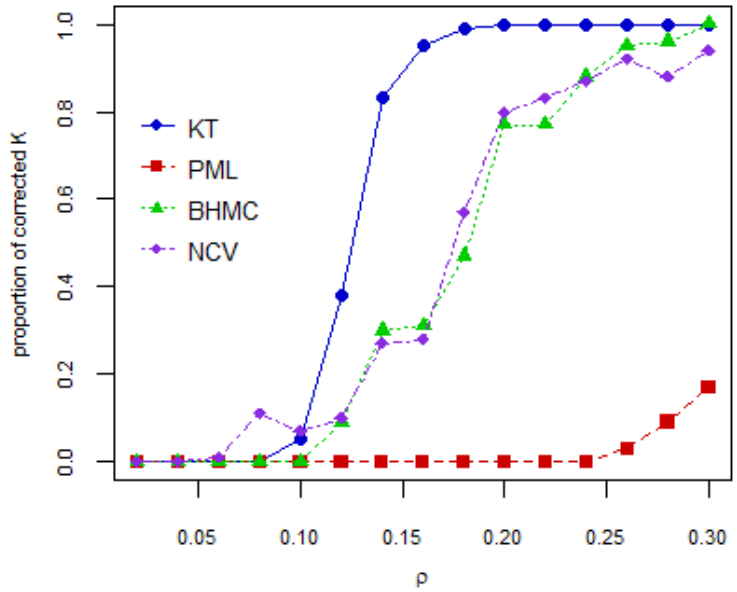
Sparsity

Estimator	$P^0 = \rho_n S^0$	penalty
$\widehat{k}_0^{\text{PML}}$	$\rho_n \to 0$ at rate $\frac{n\rho_n}{\log n} \to \infty$	$\frac{k(k+1)}{2} n \log n$
$\hat{k}_0^{\text{CBIC}}$	$\rho_n \to 0$ at rate $\frac{n\rho_n}{\log n} \to \infty$	$n \log k + \frac{k(k+1)}{2} \log n$
$\hat{k}_0^{\text{KT}}$	$\rho_n \to 0$ at rate $n\rho_n \to \infty$	$\approx k^3 \log n$

Simulations

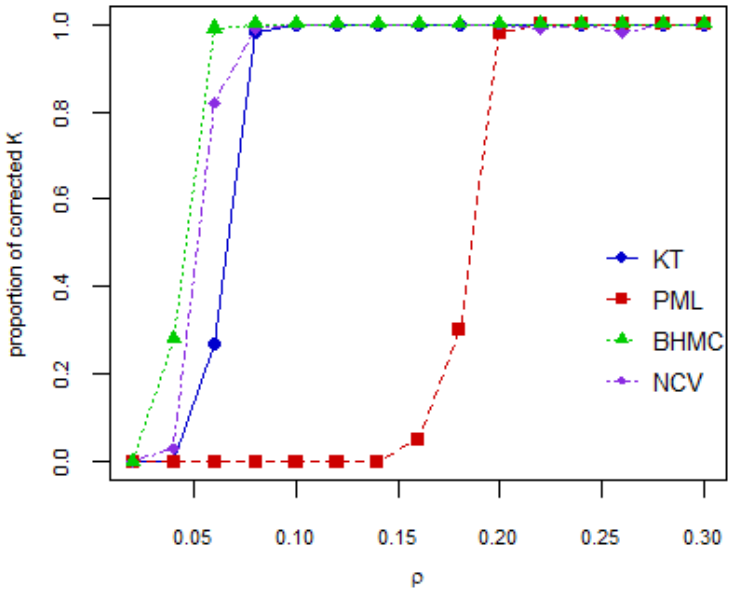


(a) $n = 300$ and $\pi = (1/3, 1/3, 1/3)$

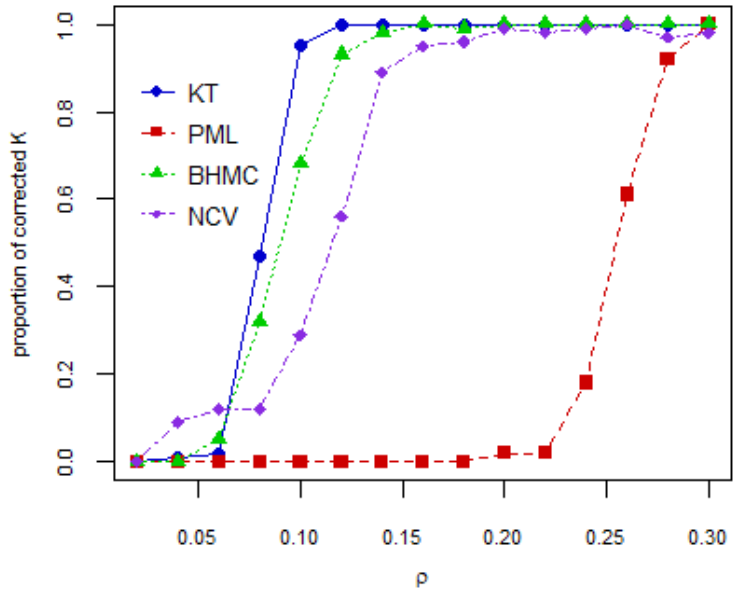


(b) $n = 300$ and $\pi = (0.2, 0.5, 0.3)$

Simulations



(a) $n = 500$ and $\pi = (1/3, 1/3, 1/3)$



(b) $n = 500$ and $\pi = (0.2, 0.5, 0.3)$

$\pi = (1/3, 1/3, 1/3),\ n = 300$							
ρ	0.06	0.08	0.10	0.12	0.14	0.16	0.18
KT	0.00	0.00	0.41	0.98	1.00	1.00	1.00
ILvb	0.00	0.15	0.54	0.99	1.00	1.00	1.00

$\pi = (0.5, 0.2, 0.3),\ n = 300$							
ρ	0.06	0.08	0.10	0.12	0.14	0.16	0.18
KT	0.00	0.00	0.04	0.38	0.80	0.98	1.00
ILvb	0.05	0.09	0.26	0.60	0.86	0.98	1.00

$$\hat{k}_0^{\text{ILvb}} = \arg\max_k \{ \log \text{KT}_k(\mathbf{x}_{n\times n}) \} \ .$$

$$\hat{k}_0^{\text{KT}} = \arg\max_k \{ \log \text{KT}_k(\mathbf{x}_{n\times n}) - \text{pen}(k, n) \} \ .$$

Final Remarks

- weighted networks
- mixed-membership blockmodels
- time-evolving networks
- networks with node covariates