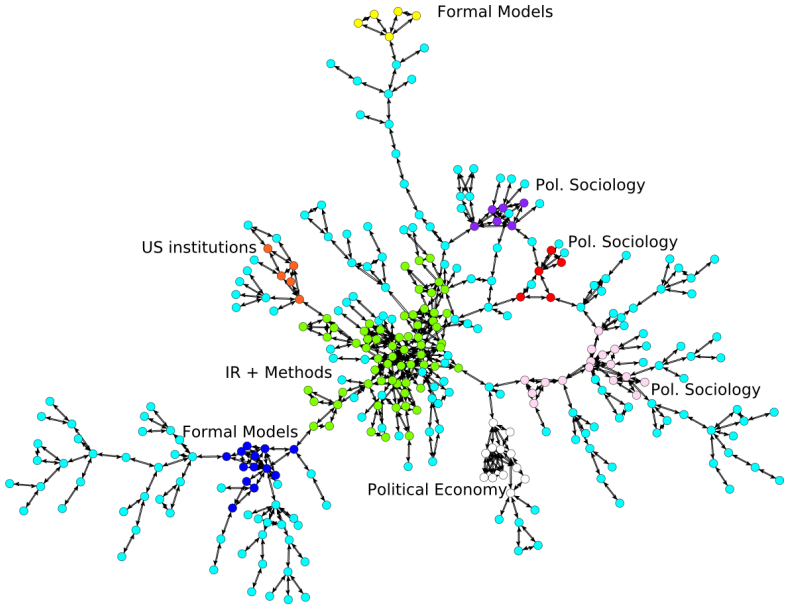
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Pairwise covariates-adjusted block model for community detection

Yang Feng
Department of Statistics, Columbia University

Network Data

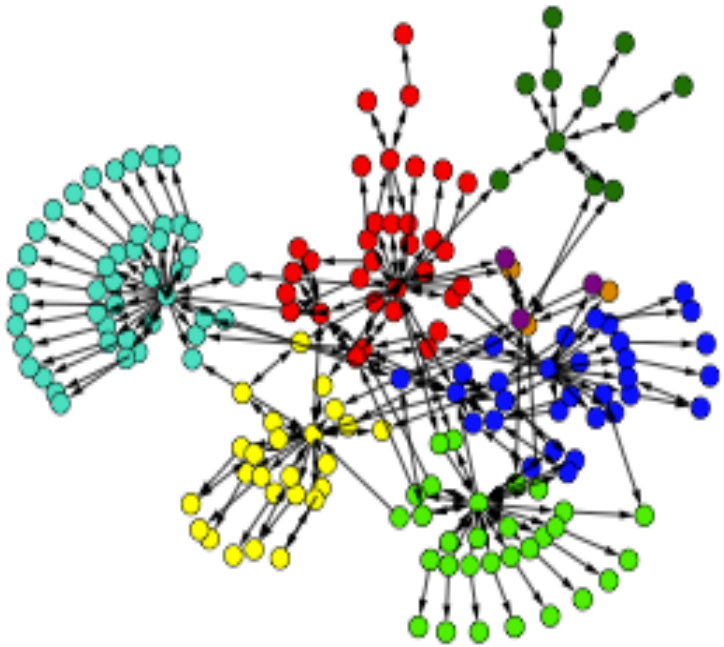
Many systems can be represented as networks.



- online social network, friendship between users
- citation networks
- Hyperlinks between web pages
- email networks

Note: images credit to `crmswitch.com` and `www.kai-arzheimer.com/networkpics/mutualworld.png`.

Community detection (Fortunato, 2010; Malliaros et al, 2013): Identifying organization of vertices in clusters, based on the network topology.



community structures detected (Girvan & Newman, 2003) in a web network

Community Detection in Networks

Communities are functional or structural units in a networked system

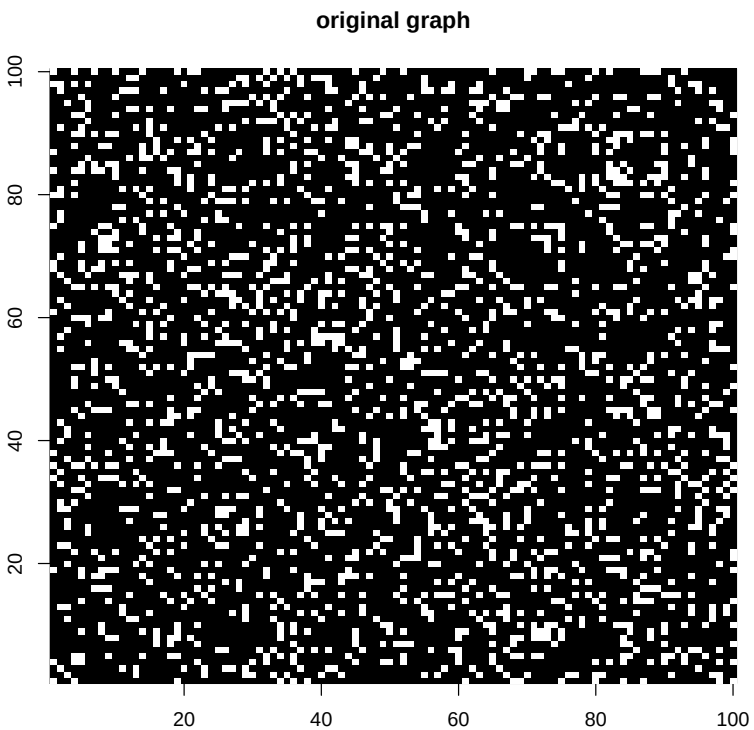
- social groupings in a social network
- related papers on a single topic in a citation network
- customers of similar interest in a network of purchase relationships
- ...

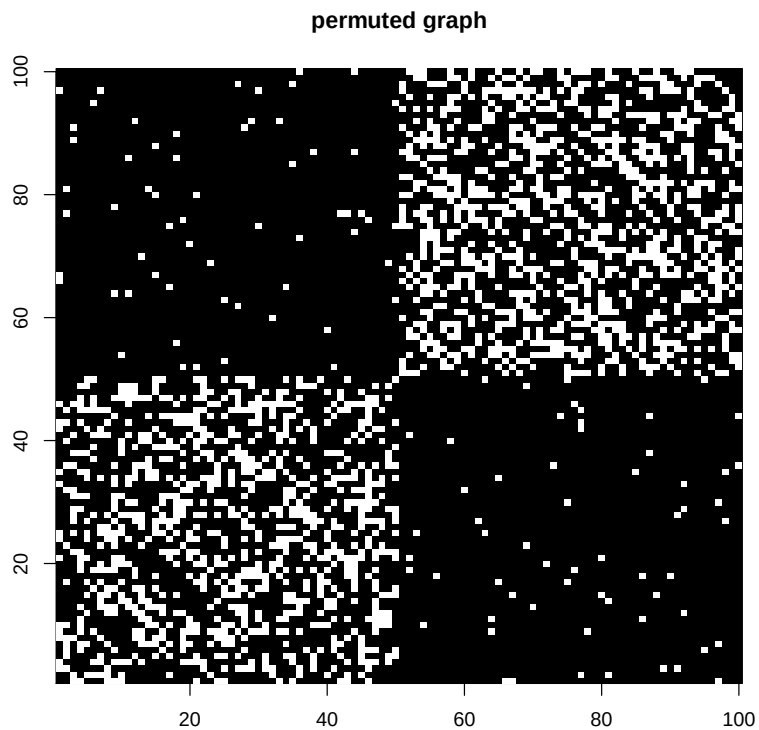
Mathematical Formulation

- A graph $G(V, E)$ with vertex set V and edge set E
- The adjacency matrix $A \in \{0, 1\}^{n \times n}$, where $A_{ij} = 1$ if and only if $(i, j) \in E$.

Stochastic Block Model (Holland, 1983)

- n nodes, K non-overlapping communities, membership $\mathbf{c} = \{c_i\}_{i=1}^n$, connectivity matrix $B \in [0,1]^{K \times K}$
- $P(A_{ij} = 1) = B_{c_i c_j} \quad (1 \leq i < j \leq n)$
- Extensions: Degree-Corrected Stochastic Block Model (Karrer & Newman, 2010), Mixed Membership Stochastic Block Model (Airoldi et al., 2009).





Community Detection with Covariates

Networks often appear with additional covariate information

- online social networks with users' personal profile information
- citation networks with papers' authors, keywords and abstracts
- ...

Motivations

1. more accurate community detection
2. identify relevant covariates
3. interplay between node and edge information

Some existing methods: Nallapati and Cohen, 2008; Chang and Blei, 2010; Akoglu et al., 2012; Ruan et al., 2013; Yang et al., 2013; Zhang et al, 2013; Binkiewicz et al, 2015; etc. Mostly heuristic or based on algorithms, not much model based work.

Community Detection with Covariates

Networks often appear with additional covariate information

- online social networks with users' personal profile information
- citation networks with papers' authors, keywords and abstracts
- ...

Motivations

1. more accurate community detection
2. identify relevant covariates
3. interplay between node and edge information

Some existing methods: Nallapati and Cohen, 2008; Chang and Blei, 2010; Akoglu et al., 2012; Ruan et al., 2013; Yang et al., 2013; Zhang et al, 2013; Binkiewicz et al, 2015; etc. Mostly heuristic or based on algorithms, not much model based work.

Community Detection with Covariates

Networks often appear with additional covariate information

- online social networks with users' personal profile information
- citation networks with papers' authors, keywords and abstracts
- ...

Motivations

1. more accurate community detection
2. identify relevant covariates
3. interplay between node and edge information

Some existing methods: Nallapati and Cohen, 2008; Chang and Blei, 2010; Akoglu et al., 2012; Ruan et al., 2013; Yang et al., 2013; Zhang et al, 2013; Binkiewicz et al, 2015; etc. Mostly heuristic or based on algorithms, not much model based work.

Network with covariates

Suppose, in addition to the adjacency matrix A, the available covariates information is formulated in a covariate matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times d}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th node's covariate vector.

Goal: estimate $\mathbf{c}$ from the observations A and X.

Question: How to integrate the nodal information with the adjacency matrix?

Network with covariates

Suppose, in addition to the adjacency matrix A, the available covariates information is formulated in a covariate matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times d}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th node's covariate vector.

Goal: estimate $\mathbf{c}$ from the observations A and X.

Question: How to integrate the nodal information with the adjacency matrix?

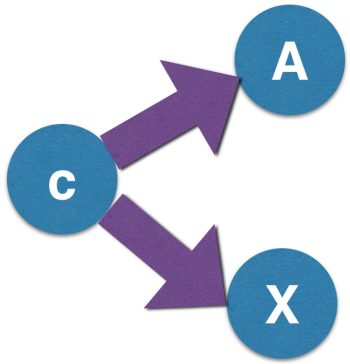
Suppose, in addition to the adjacency matrix A, the available covariates information is formulated in a covariate matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times d}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th node's covariate vector.

Goal: estimate $\mathbf{c}$ from the observations A and X.

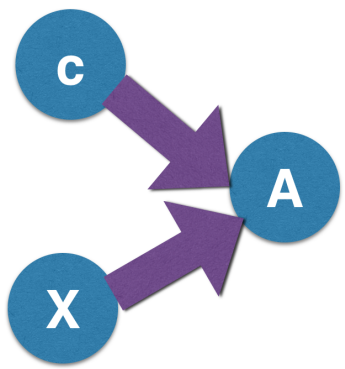
Question: How to integrate the nodal information with the adjacency matrix?

Two possible model assumptions

- Assuming both A and X depend on $\mathbf{c}$.
- Example: citation network; $\mathbf{c}$: research topic, X: keywords.

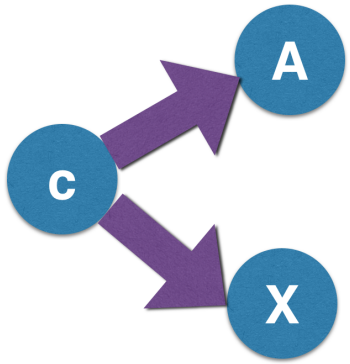


- Assuming A depends on both $\mathbf{c}$ and X.
- Example: high-school friendship network; $\mathbf{c}$: grade, X: ethnicity.

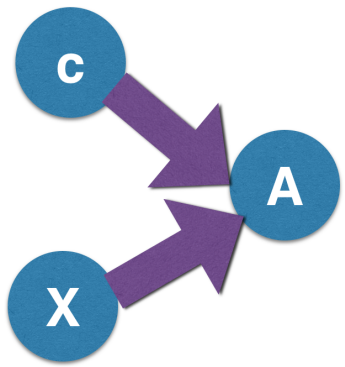


Two possible model assumptions

- Assuming both A and X depend on $\mathbf{c}$.
- Example: citation network; $\mathbf{c}$: research topic, X: keywords.



- Assuming A depends on both $\mathbf{c}$ and X.
- Example: high-school friendship network; $\mathbf{c}$: grade, X: ethnicity.



First model: a generic two-step inference procedure (Weng and F., 2017)

Assuming $A \perp X \mid \mathbf{c}$,

(1) Parameter estimation based on $P(A, X; \boldsymbol{\theta})$:

$$P(A, X; \boldsymbol{\theta}) = \sum_{\mathbf{c}} P(A \mid \mathbf{c}) P(X \mid \mathbf{c}) P(\mathbf{c}) \tag{1}$$

$$= P(X; \boldsymbol{\theta}_1) \sum_{\mathbf{c}} P(A \mid \mathbf{c}; \boldsymbol{\theta}_2) P(\mathbf{c} \mid X; \boldsymbol{\theta}_3), \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ indexes a family of generative models (here, we are not assuming specific parametric forms)

(2) Posterior inference according to $P(\mathbf{c} \mid A, X)$.

$$P(\mathbf{c} \mid A, X) = \frac{P(A \mid \mathbf{c}) P(X \mid \mathbf{c}) P(\mathbf{c})}{\sum_{\mathbf{c}} P(A \mid \mathbf{c}) P(X \mid \mathbf{c}) P(\mathbf{c})} = \frac{P(A \mid \mathbf{c}) P(\mathbf{c} \mid X) P(X)}{\sum_{\mathbf{c}} P(A \mid \mathbf{c}) P(\mathbf{c} \mid X) P(X)}$$

$$= \frac{P(A \mid \mathbf{c}) P(\mathbf{c} \mid X)}{\sum_{\mathbf{c}} P(A \mid \mathbf{c}) P(\mathbf{c} \mid X)}. \tag{3}$$

Node-coupled Stochastic Block Model (NSBM)

(a) SBM:

$$P(A \mid \mathbf{c}) = \prod_{i<j} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1 - A_{ij}}$$

(b) multi-logistic regression:

$$P(\mathbf{c} \mid X) = \prod_i \frac{\exp(\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i)}{\sum_{k=1}^K \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)},$$

where $B = (B_{ab}) \in [0,1]^{K \times K}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T \in \mathbb{R}^{Kp}$. Here, $\boldsymbol{\beta}_K = 0$ for identifiability.

Community detection consistency

$$\text{(\textit{strong consistency})} \qquad P(\hat{\mathbf{c}} = \mathbf{c}) \to 1, \text{ as } n \to \infty,$$

$$\text{(\textit{weak consistency})} \qquad P\Big(\frac{1}{n}\sum_{i=1}^{n} 1(\hat{c}_i \neq c_i) < \epsilon\Big) \to 1, \text{ as } n \to \infty, \forall \epsilon > 0.$$

Here, $B = \rho_n \bar{B}$ with $\bar{B}$ fixed and sparsity level $\rho_n = P(A_{ij} = 1) \to 0$ as $n \to \infty$ (Bickel and Chen (2009); Zhao et al. (2012); Bickel et al. (2013))

Note: both versions of consistency are interpreted upon a permutation of the community labels.

Marginalization over $\mathbf{c}$.

$$(\hat{\boldsymbol{\beta}}, \hat{B}) = \underset{\substack{\boldsymbol{\beta}_K = \mathbf{0},\ \boldsymbol{\beta} \in \mathbb{R}^{Kp} \\ B \in [0,1]^{K \times K}, B^T = B}}{\arg\max} \sum_{\mathbf{c}} \prod_{i<j} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1 - A_{ij}} \cdot \prod_i \frac{e^{\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i}}{\sum_{k=1}^K e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i}}, \quad (4)$$

$$\hat{\mathbf{c}} = \underset{\mathbf{c} \in \{1,\ldots,K\}^n}{\arg\max} \prod_{i<j} \hat{B}_{c_i c_j}^{A_{ij}} (1 - \hat{B}_{c_i c_j})^{1 - A_{ij}} \cdot \prod_i \frac{e^{\hat{\boldsymbol{\beta}}_{c_i}^T \boldsymbol{x}_i}}{\sum_{k=1}^K e^{\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i}}. \quad (5)$$

Technical conditions

Condition 1

$\bar{B}$ has no two identical columns.

Condition 2

$(c_1, \boldsymbol{x}_1), \ldots, (c_n, \boldsymbol{x}_n) \overset{iid}{\sim} (c, \boldsymbol{x})$ with $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) \succ 0$, where $\succ 0$ represents the matrix being positive definite.

Condition 3

There exist constants κ_1 and κ_2 such that for sufficiently large t, we have

$$P(\|\boldsymbol{x}\|_2 > t) \leq \kappa_1 e^{-\kappa_2 t}.$$

Theorem 1

Assume the data (A, X) follows NSBM and Conditions 1-3 hold. In addition, assume $\frac{n\rho_n}{\log n} \to \infty$ as $n \to \infty$. Then, we have as $n \to \infty$

$$P(\hat{\boldsymbol{c}} = \boldsymbol{c}) \to 1, \quad \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta})),$$

where $I(\boldsymbol{\beta})$ is the Fisher information for the multi-logistic regression problem of regressing $\boldsymbol{c}$ on X.

Remark: condition on expected degree the same as Bickel and Chen (2009) and Zhao et al. (2012).

Remark: Computational infeasible even to evaluate the objective function at any non-degenerate point.

Theorem 1

Assume the data (A, X) follows NSBM and Conditions 1-3 hold. In addition, assume $\frac{n\rho_n}{\log n} \to \infty$ as $n \to \infty$. Then, we have as $n \to \infty$

$$P(\hat{\boldsymbol{c}} = \boldsymbol{c}) \to 1, \quad \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta})),$$

where $I(\boldsymbol{\beta})$ is the Fisher information for the multi-logistic regression problem of regressing $\boldsymbol{c}$ on X.

Remark: condition on expected degree the same as Bickel and Chen (2009) and Zhao et al. (2012).

Remark: Computational infeasible even to evaluate the objective function at any non-degenerate point.

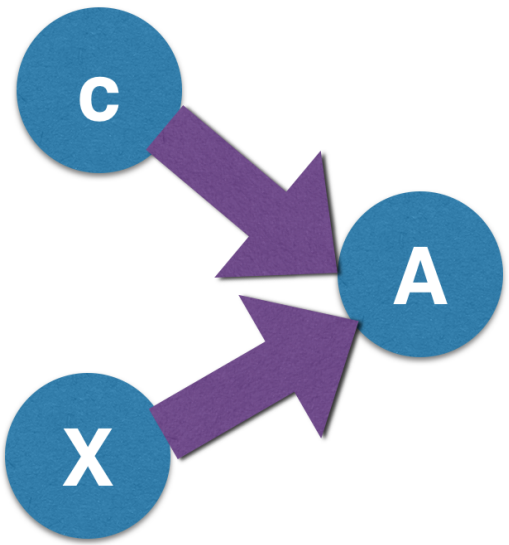
Alternative estimators

More in: Weng and **F.** (2016), Community detection with nodal information, `https://arxiv.org/abs/1610.09735`.

- Variational EM
- Profile likelihood
- Semi-definite Programming

Second assumption

Assuming A depends on both $\mathbf{c}$ and X.



Pairwise Covariate-Adjusted Block Model (PCABM)

- Given $X_{n\times p}$, define a p-dimensional covariate $Z_{ij} = g(\mathbf{X}_i, \mathbf{X}_j)$. It is a measure of similarity of nodes i and j in different aspects.
- For example, $|X_{i1} - X_{j1}|$ measures the distance of the covariates for two nodes.
- Given community assignment c and pairwise covariates Z, assume A_{ij} are independent Poisson distributed with rate

$$\lambda_{ij} := P(A_{ij} = 1) = B_{c_i c_j} e^{Z_{ij}^T \gamma^0}$$

- Adjustment factor: $\exp\{Z_{ij}^T \gamma^0\}$.
- Reduces to the standard SBM when $\gamma^0 = 0$.
- Let $B = \rho_n \bar{B}$ with $\bar{B}$ fixed and $\rho_n \to 0$.

- The likelihood function of the complete graph model (CGM) is defined as

$$f(A, e|Z, \gamma, B, \pi) = \prod_{i=1}^{n} \pi_{e_i} \prod_{i<j} B_{e_i e_j}^{A_{ij}} e^{A_{ij} Z_{ij}^T \gamma} e^{-B_{e_i e_j} e^{Z_{ij}^T \gamma}},$$

  where $\pi_a = P(e = a)$.
- We then define the following maximum likelihood estimators for graph model (GM):

$$\hat{\gamma} = \arg\max_{\gamma \in \Gamma} \sum_{e} f(A, e|Z, \gamma), \ \ \hat{c} = \arg\max_{e} f(A, e|Z, \hat{\gamma}),$$

  where Γ is a compact set in $\mathbb{R}^p$. Basically, GM is the community marginalization of CGM.

Technical Conditions

$n_k(\mathbf{e}) = \sum_{i=1}^{n} \mathbf{1}_{e_i=k}$, and $s_{\mathbf{e}}(k,l) = \{(i,j)|e_i = k, e_j = l, i \neq j\}$.

Condition 1 (non-degeneracy)

$n_k(\mathbf{e}) \to \infty$ for any k.

Condition 2 (non-collinearity)

There exists a community pair (k,l) s.t. $|s_{\mathbf{e}}(k,l)| \geq p+1$ and the matrix $[(\mathbf{Z}_{ij}^T \ \mathbf{1})]_{(i,j)\in s_{\mathbf{e}}(k,l)} \in \mathbb{R}^{|s_{\mathbf{e}}(k,l)|\times(p+1)}$ is full column rank.

For any given e, let

$$\tilde{\gamma}(e) = \arg\max_{\gamma \in \Gamma} f(A, e|Z, \gamma).$$

Theorem 3

Under Conditions 1 and 2, as $n \to \infty$, $\tilde{\gamma}(e) \xrightarrow{p} \gamma^0$.

Now, let

$$\hat{c} = \arg\max_{e'} f(A, e'|Z, \tilde{\gamma}(e)).$$

Theorem 4

Under Conditions 1 and 2, as $n \to \infty$, if $N\rho_n \to \infty$, the asymptotical distribution of $\sqrt{N\rho_n}(\tilde{\gamma}(\mathbf{e}) - \gamma^0)$ is multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $I^{-1}(\gamma^0)$, where $I(\gamma^0) = \sum_{ab} \bar{B}_{ab}\pi_a\pi_b(\nu_2 - \nu_0^{-1}\nu_1\nu_1^T)$.

For any given e, let

$$\tilde{\gamma}(e) = \arg\max_{\gamma \in \Gamma} f(A, e|Z, \gamma).$$

Theorem 3

Under Conditions 1 and 2, as $n \to \infty$, $\tilde{\gamma}(e) \xrightarrow{p} \gamma^0$.

Now, let

$$\hat{c} = \arg\max_{e'} f(A, e'|Z, \tilde{\gamma}(e)).$$

Theorem 4

Under Conditions 1 and 2, as $n \to \infty$, if $N\rho_n \to \infty$, the asymptotical distribution of $\sqrt{N\rho_n}(\tilde{\gamma}(\mathbf{e}) - \gamma^0)$ is multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $I^{-1}(\gamma^0)$, where $I(\gamma^0) = \sum_{ab} \bar{B}_{ab}\pi_a\pi_b(\nu_2 - \nu_0^{-1}\nu_1\nu_1^T)$.

Theorem 5

Under PCABM, when $\tilde{\gamma}(\mathbf{e})$ is consistent, $\tilde{\mathbf{c}}(\mathbf{e})$ is weakly consistent if $\varphi_n \to \infty$ and strongly consistent if $\varphi_n / \log n \to \infty$.

Spectral clustering with adjustment (SCWA)

Observation

$$E\left[\frac{A_{ij}}{\exp\{Z_{ij}^T\gamma^0\}}\right] = B_{c_i c_j}.$$

Intuitively, we would like to perform spectral clustering on $A' = \{A'_{ij}\}$, where

$$A'_{ij} = \frac{A_{ij}}{\exp\{Z_{ij}^T\tilde{\gamma}\}}.$$

Remark

We need to choose a lower bound $v(n)$ for the adjusted factor, to control large variances of the adjusted variable. When $e^{\mathbf{Z}_{ij}^T\tilde{\gamma}} < v(n)$, we use $v(n)$ as the adjustment factor.

Theorem 6 (Spectral bound of poisson random matrices).

Let A be the adjacency matrix generated by PCABM (M, B, Z, γ^0), and the adjusted adjacency matrix A' is derived with some proper threshold value $v(n) > 0$. Assume that $nP_{\max} \leq d$ and $d \geq C_1 \log n$ for $C_1 > 0$. Then, for any $r > 0$, there exists a constant C such that

$$\|A' - P\| \leq C\sqrt{d}/v(n)$$

with probability at least $1 - n^{-r \times v(n)}$.

We define two measure of estimation error, an overall relative error and a worst case relative error:

$$L_1(\hat{M}, M) = n^{-1} \min_{Q \in \mathcal{S}_K} \|\hat{M}Q - M\|_0,$$

$$L_2(\hat{M}, M) = \min_{Q \in \mathcal{S}_K} \max_{1 \leq k \leq K} n_k^{-1} \|(\hat{M}Q)_{G_k \cdot} - M_{G_k \cdot}\|_0,$$

Theorem 7 (Error bound)

Let A be an adjacency matrix generated from a PCABM (M, B, Z, γ^0), where $B = \rho_n \bar{B}$ for some $\rho_n \geq C_1 \log n/n$ and with $\bar{B}'s$ minimum absolute eigenvalue bounded below by $\tau > 0$ and $\max_{kl} \bar{B}(k, l) = 1$. The adjusted adjacency matrix A' is derived with some threshold value $v(n) > 0$. Let $\hat{M}$ be the output of spectral clustering using $(1 + \epsilon)$ approximate k-means. For any $r > 0$, there exists an absolute constant C such that if

$$(2 + \epsilon)\frac{Kn}{n_{\min}^2 \tau^2 \rho_n v^2(n)} < C,$$

then with probability at least $1 - n^{-r \times v(n)}$,

$$L_2(\hat{M}, M) \leq C^{-1}(2 + \epsilon)\frac{Kn}{n_{\min}^2 \tau^2 \rho_n v^2(n)},$$

$$L_1(\hat{M}, M) \leq C^{-1}(2 + \epsilon)\frac{Kn'_{\max}}{n_{\min}^2 \tau^2 \rho_n v^2(n)}.$$

Simulation

- $B_0 = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.
- $n = 100, 200, 300, 400, 500$.
- $\rho_n = 2(\log n)^{1.5}/n$.
- $\gamma^0 = (0.4, 0.8, 1.2, 1.6, 2)$.
- The components of $\mathbf{X}_i$ are i.i.d. and follow the distributions: $B(1, 0.1)$, $Pois(0.1)$, $U(0, 1)$, $Exp(0.3)$, $N(0, 0.3)$.
- The pairwise covariates $\mathbf{Z}_{ij} = |\mathbf{X}_i - \mathbf{X}_j|$

We use two quantitative measures for evaluating the community detection performance.

Normalized Mutual Information (Ana and Jain, 2003):

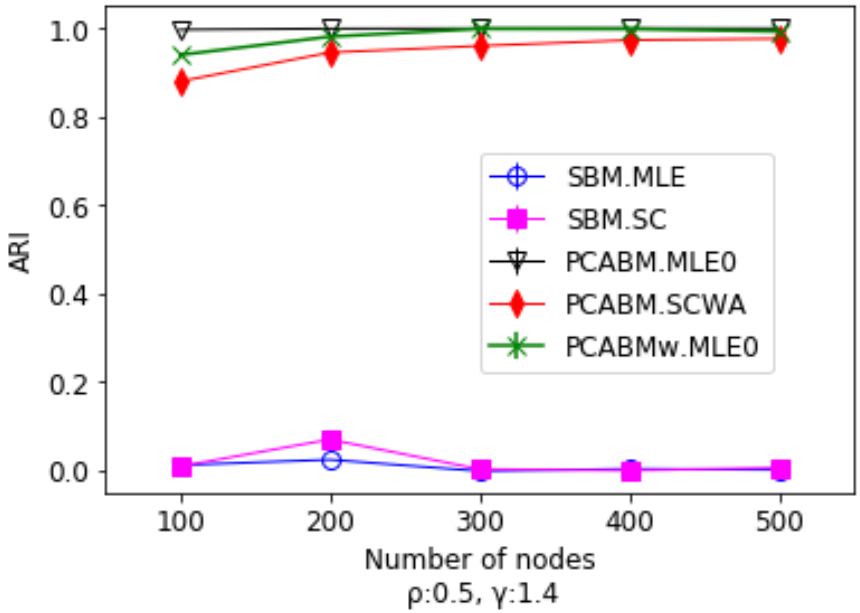
$$\mathrm{NMI} = \frac{-2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij\cdot} n}{n_{i\cdot} n_{\cdot j}}\right)}{\sum_i n_{i\cdot} \log\left(\frac{n_{i\cdot}}{n}\right) + \sum_j n_{\cdot j} \log\left(\frac{n_{\cdot j}}{n}\right)}.$$

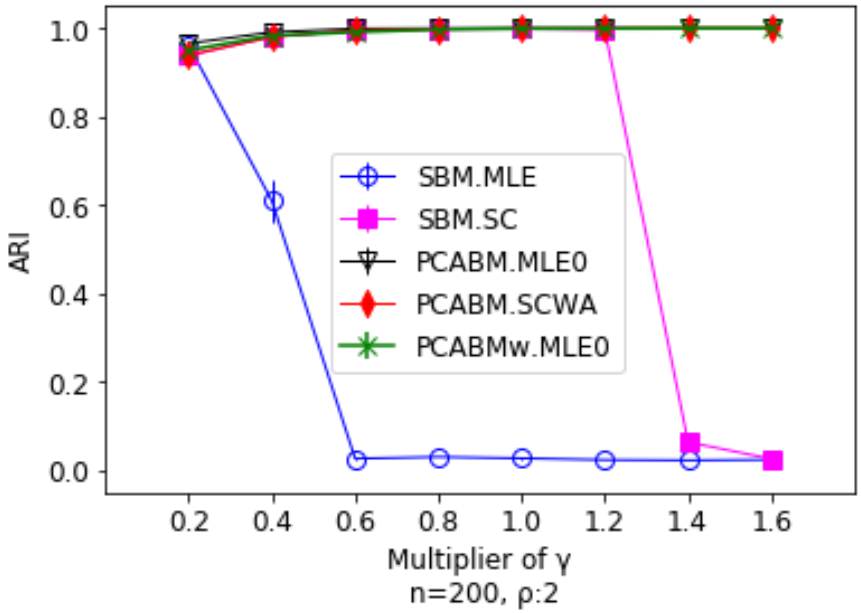
Adjusted Rand Index (Hubert and Arabie, 1985):

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\sum_i \binom{n_{i\cdot}}{2} + \frac{1}{2}\sum_j \binom{n_{\cdot j}}{2} - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}},$$

where $n_{i\cdot}$ denotes the true number of nodes in community i, $n_{\cdot j}$ represents the number of nodes in the estimated community j and n_{ij} is the number of nodes belonging to community i but estimated to be in community j. In [0,1].

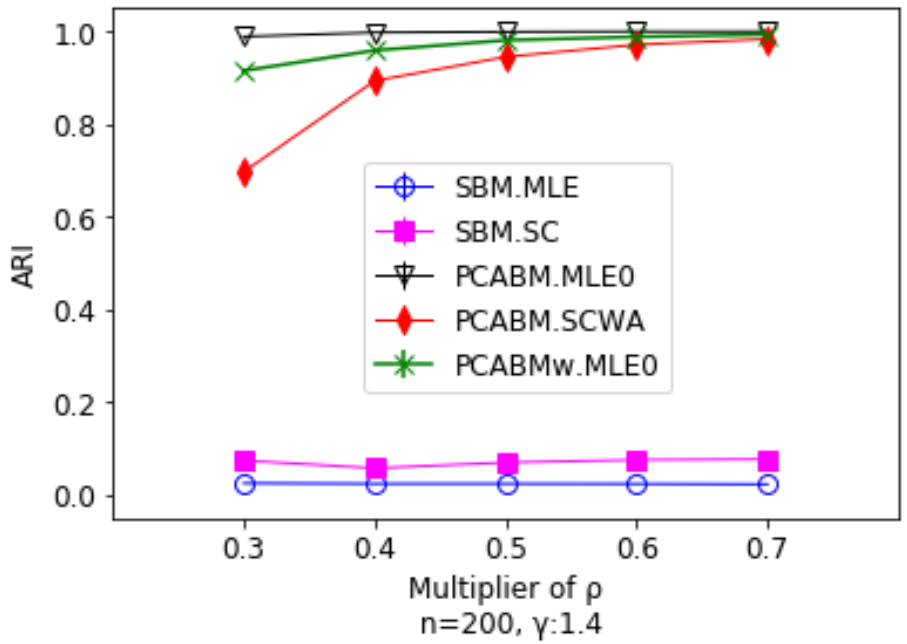
Results



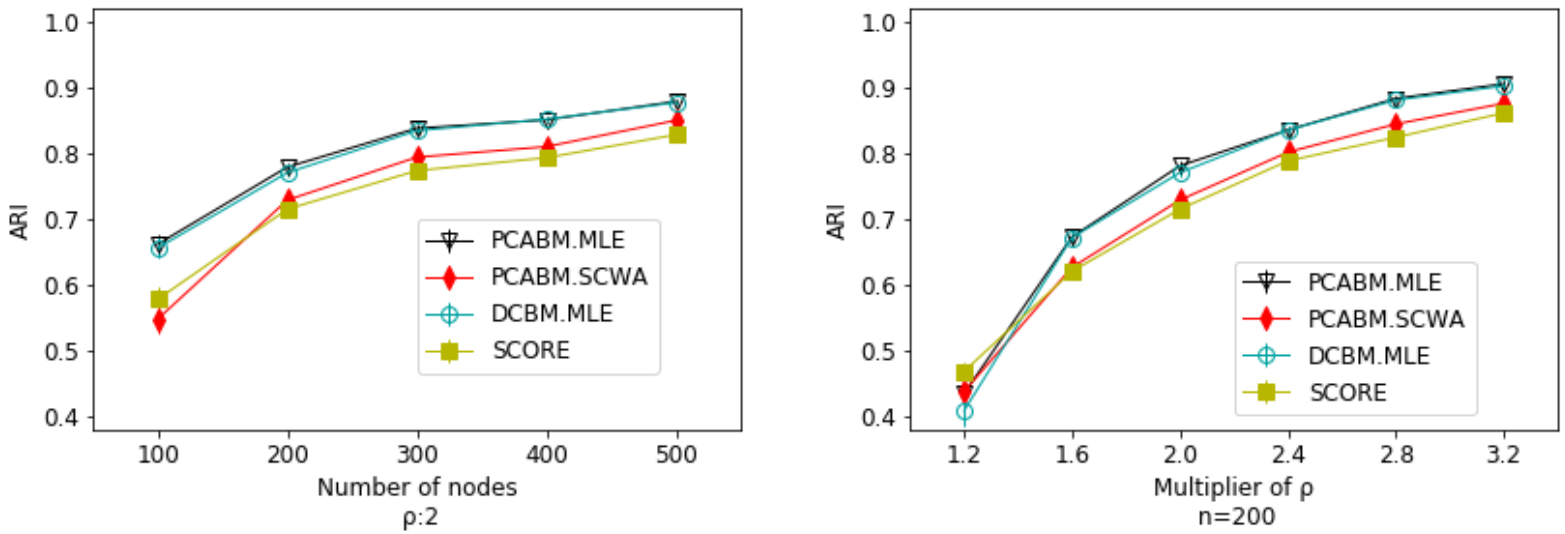


- SBM.MLE: MLE for SBM
- SBM.SC: Spectral clustering for SBM
- PCABM.MLE0: MLE for PCABM with random initialization
- PCABM.SCWA: Spectral clustering with adjustment for PCABM
- PCABMw.MLE0: MLE for PCABM with random initialization and misspecified pairwise function

Results


1.0
0.8
0.6
0.4
0.2
0.0
ARI
SBM.MLE
SBM.SC
PCABM.MLE0
PCABM.SCWA
PCABMw.MLE0
0.3
0.4
0.5
0.6
0.7
Multiplier of ρ
n=200, γ:1.4

DCBM setting



Simulation results under DCBM for different parameter settings

Political blog data

- The nodes are blogs about US politics and edges represent hyperlinks between them.
- 1,222 nodes and 16,714 edges.
- We let $Z_{ij} = \log(d_i \times d_j)$, where d_i is the degree for the i-th node.

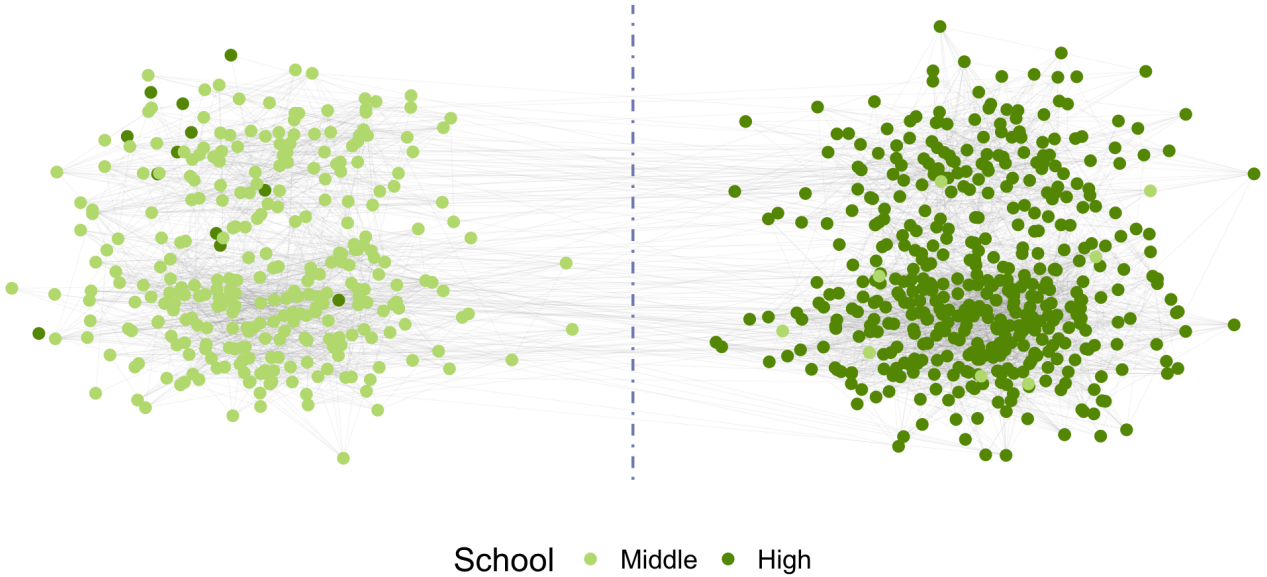
	ARI	NMI	Errors
Karrer(2011)	-	0.72	-
Zhao(2012)	0.819	-	-
SCORE	0.819	0.725	58
PCABM.MLE	0.825	0.737	56

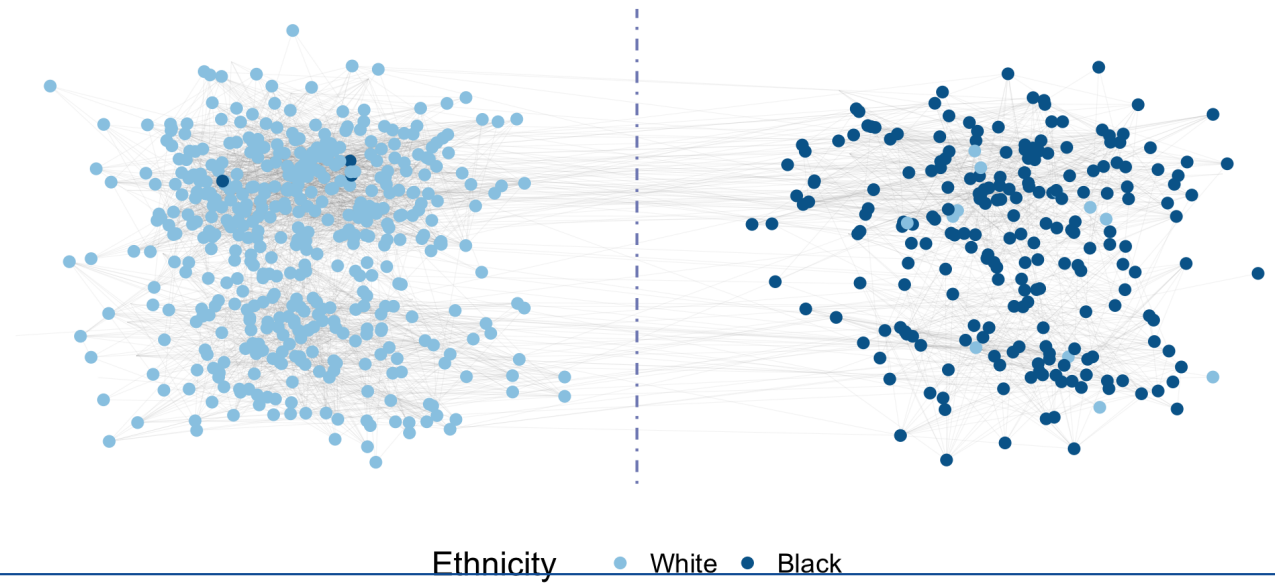
Performance comparison on political blogs data

School friendship data

- Network with 777 nodes and 4124 edges.
- Covariates: grade, gender, ethnicity and number of friends nominated.
- In each experiment, out of the three nodal covariates, school, ethnicity and gender, we viewed one covariate as the indicator for the underlying community, and performed community detection by using the pairwise covariates constructed with the other two covariates.
- For gender, school and ethnicity, we created dummy variables to represent whether the corresponding covariate values were the same for the pair of nodes. In addition, for number of nominated friends, we use $Z_{ij} = \log(n_i + 1) + \log(n_j + 1)$.

Results





NMI comparison for different methods

	School	Race	Gender
SBM.SC	0.043	-0.024	0
SBM.MLE	0.048	0.138	-0.001
PCABM.MLE	0.894	0.914	0.03

Future Directions

- Combine the two models by dividing the covariates into two categories automatically.
- In the setting with high-dimensional covariates, penalized likelihood methods are more appealing for both community detection and variable selection.
- For very sparse networks, considering $n\rho_n = O(1)$ seems to be a more realistic asymptotic framework. Under such asymptotics, community detection consistency is impossible.
- In this work, we assume the number of communities K is known. How to select K is an important problem in community detection. Some recent efforts towards this direction include Saldana, Yu and F. (2015); Le and Levina (2015); Wang and Bickel (2015); Lei (2016); Chen and Lei (2016); Li and Zhu (2017), etc.

Thank you very much!

References:

- Weng and **F.** (2016), Community detection with nodal information. `https://arxiv.org/abs/1610.09735`.
- Huang and **F.** (2018), Pairwise Covariates-adjusted Block Model for Community Detection, manuscript.

Variational Inference

- SBM: (Daudin et al., 2008; Celisse et al., 2012; Bickel et al., 2013).
-

$$\max_{\boldsymbol{\theta}} \log P(A, X; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}, Q(\cdot)} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \boldsymbol{\theta}) - \log Q(\mathbf{c})], \tag{6}$$

  where $Q(\cdot)$ denotes any joint distribution on $\mathbf{c}$.
- Here, we consider the mean-field variational approach (Jordan et al., 1999),

$$\max_{\theta, Q \in \mathcal{Q}} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \theta) - \log Q(\mathbf{c})], \tag{7}$$

  where $\mathcal{Q} = \{Q : Q(\mathbf{c}) = \prod_{i=1}^{n} q_{ic_i}, \sum_k q_{ik} = 1, 1 \le i \le n\}$. The subset $\mathcal{Q}$ contains all the distributions under which $\mathbf{c}$ is jointly independent.
- Denote the maximizer in (7) by $(\check{B}, \check{\beta})$ and $\check{\boldsymbol{c}}$.

- SBM: (Daudin et al., 2008; Celisse et al., 2012; Bickel et al., 2013).
-

$$\max_{\boldsymbol{\theta}} \log P(A, X; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}, Q(\cdot)} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \boldsymbol{\theta}) - \log Q(\mathbf{c})], \tag{6}$$

  where $Q(\cdot)$ denotes any joint distribution on $\mathbf{c}$.
- Here, we consider the mean-field variational approach (Jordan et al., 1999),

$$\max_{\boldsymbol{\theta}, \boldsymbol{Q} \in \mathcal{Q}} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \boldsymbol{\theta}) - \log Q(\mathbf{c})], \tag{7}$$

  where $\mathcal{Q} = \{Q : Q(\mathbf{c}) = \prod_{i=1}^{n} q_{ic_i}, \sum_k q_{ik} = 1, 1 \le i \le n\}$. The subset $\mathcal{Q}$ contains all the distributions under which $\mathbf{c}$ is jointly independent.
- Denote the maximizer in (7) by $(\check{B}, \check{\beta})$ and $\check{\boldsymbol{c}}$.

- SBM: (Daudin et al., 2008; Celisse et al., 2012; Bickel et al., 2013).
-

$$\max_{\boldsymbol{\theta}} \log P(A, X; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}, Q(\cdot)} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \boldsymbol{\theta}) - \log Q(\mathbf{c})], \tag{6}$$

  where $Q(\cdot)$ denotes any joint distribution on $\mathbf{c}$.
- Here, we consider the mean-field variational approach (Jordan et al., 1999),

$$\max_{\boldsymbol{\theta}, \boldsymbol{Q} \in \mathcal{Q}} \mathbb{E}_Q[\log P(A, X, \mathbf{c}; \boldsymbol{\theta}) - \log Q(\mathbf{c})], \tag{7}$$

  where $\mathcal{Q} = \{Q : Q(\mathbf{c}) = \prod_{i=1}^{n} q_{ic_i}, \sum_k q_{ik} = 1, 1 \le i \le n\}$. The subset $\mathcal{Q}$ contains all the distributions under which $\mathbf{c}$ is jointly independent.
- Denote the maximizer in (7) by $(\check{B}, \check{\boldsymbol{\beta}})$ and $\check{\boldsymbol{c}}$.

Theorem 2

Suppose the conditions of Theorem 1 hold. Then as $n \to \infty$

$$P(\check{\boldsymbol{c}} = \boldsymbol{c}) \to 1, \quad \sqrt{n}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\beta})),$$

where $I(\boldsymbol{\beta})$ is the Fisher information for the multi-logistic regression problem of regressing $\boldsymbol{c}$ on X.

Remark: the same asymptotics as the MLE when $\frac{n\rho_n}{\log n} \to \infty$ as long as the network has sufficient edge information for doing approximate inference.

Idea: treat the community assignments $\mathbf{c}$ as an unknown parameter to estimate.

$$(\tilde{\boldsymbol{\beta}}, \tilde{B}, \tilde{\mathbf{c}}) = \underset{\substack{\boldsymbol{\beta}_K = \mathbf{0},\ \boldsymbol{\beta} \in \mathbb{R}^{Kp} \\ \mathbf{c} \in \{1,\ldots,K\}^n \\ B \in [0,1]^{K \times K}, B^T = B}}{\arg\max} \prod_{i<j} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1 - A_{ij}} \cdot \prod_i \frac{e^{\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i}}{\sum_{k=1}^K e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i}}. \tag{8}$$

Reference for SBM and DCBM: Bickel and Chen (2009); Zhao et al. (2012). For theoretical convenience, we consider a slightly different formulation:

$$(\tilde{\boldsymbol{\beta}}, \tilde{B}, \tilde{\mathbf{c}}) = \underset{\substack{\boldsymbol{\beta}_K = \mathbf{0},\ \boldsymbol{\beta} \in \mathbb{R}^{Kp} \\ \mathbf{c} \in \{1,\ldots,K\}^n \\ B \in \mathbb{R}^{K \times K}, B^T = B}}{\arg\max} \prod_{i<j} e^{-B_{c_i c_j}} B_{c_i c_j}^{A_{ij}} \cdot \prod_i \frac{e^{\boldsymbol{\beta}_{c_i}^T \boldsymbol{x}_i}}{\sum_{k=1}^K e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i}}. \tag{9}$$

Computing MPLE

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{pK},\boldsymbol{\beta}_K=\mathbf{0}}{\arg\max} \sum_i \left[\boldsymbol{\beta}_{c_i^t}^T \boldsymbol{x}_i - \log\left(\sum_k e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i}\right)\right], \tag{10}$$

$$B_{ab}^{t+1} = \underset{B_{ab}}{\arg\max}\ \log B_{ab} \cdot \sum_{i<j} A_{ij} 1(c_i^t = a, c_j^t = b) - B_{ab} \cdot \sum_{i<j} 1(c_i^t = a, c_j^t = b), \tag{11}$$

$$\begin{aligned} \boldsymbol{c}^{t+1} = \underset{\boldsymbol{c}\in\{1,\ldots,K\}^n}{\arg\max} \sum_{ab} \Big[& \log B_{ab}^{t+1} \cdot \sum_{i<j} A_{ij} 1(c_i = a, c_j = b) \\ & - B_{ab}^{t+1} \cdot \sum_{i<j} 1(c_i = a, c_j = b)\Big] + \sum_i (\boldsymbol{\beta}_{c_i}^{t+1})^T \boldsymbol{x}_i. \end{aligned} \tag{12}$$

Theorem 3

Assume the data (A, X) follows NSBM and Conditions 1 and 2 hold.

(i) If $n\rho_n \to \infty$ and $\mathbb{E}\|\boldsymbol{x}\|_2^\alpha < \infty$ $(\alpha > 1)$, then there exists a constant $\gamma > 0$ such that, as $n \to \infty$

$$P\Big(\frac{1}{n}\sum_{i=1}^{n} 1(\tilde{c}_i \neq c_i) \leq \gamma(n\rho_n)^{-1/2}\Big) \to 1, \quad \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O_p((n\rho_n)^{\frac{1-\alpha}{2\alpha}}).$$

(ii) Assume Condition 3 is satisfied. If $\frac{n\rho_n}{\log n} \to \infty$, then as $n \to \infty$

$$P(\tilde{\boldsymbol{c}} = \boldsymbol{c}) \to 1, \quad \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, I^{-1}(\boldsymbol{\beta})),$$

where $I(\boldsymbol{\beta})$ is the Fisher information for the multi-logistic regression problem of regressing $\boldsymbol{c}$ on X.

Key challenges for likelihood based methods

- Non-convex. Multiple local optima may exist and the global solution is often impossible to accurately allocate.
- A good convex initialization is desirable.
- A popular class of convex initialization: semidefinite programming (SDP). Chen et al. (2012); Amini and Levina (2014); Cai and Li (2015); Montanari and Sen (2015); Guedon and Vershynin (2015).

$$\begin{aligned} \hat{Z} = \arg\max_{Z} \quad & \langle A + \gamma_n XX^T, Z \rangle \\ \text{subject to} \quad & Z \succeq 0, Z \in \mathbb{R}^{n \times n} \\ & 0 \leq Z_{ij} \leq 1, 1 \leq i, j \leq n \\ & \sum_{ij} Z_{ij} = \lambda_n, \end{aligned} \tag{13}$$

where $\gamma_n, \lambda_n > 0$ are two tuning parameters; $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices. Similar spirit as Binkiewicz et al. (2014) for spectral clustering. We then obtain the communities by running K-means on $\hat{Z}$ (treating each row of $\hat{Z}$ as a data point in $\mathbb{R}^n$).

Solving (13) via alternating direction method of multipliers (ADMM) (Boyd et al., 2011).

Theorem 4

Assume condition (b) in NSBM holds and $(c_1, \boldsymbol{x}_1), \ldots, (c_n, \boldsymbol{x}_n) \overset{iid}{\sim} (c, \boldsymbol{x})$. Let $\{\bar{c}_i\}_{i=1}^n$ be the community estimates returned by running K-means on $\hat{Z}$. If $\min_a \bar{B}_{aa} > \max_{a \neq b} \bar{B}_{ab}$, $n\rho_n \to \infty$ and $\|\boldsymbol{x}\|_2$ is sub-Gaussian, then by choosing $\gamma_n = O(n^{-1})$ and $\lambda_n = \sum_{k=1}^K (\sum_{i=1}^n 1(c_i = k))^2$, we have

$$\frac{1}{n} \sum_{i=1}^n 1(\bar{c}_i \neq c_i) \xrightarrow{P} 0.$$

Remark:

- The crucial assumption $\min_a \bar{B}_{aa} > \max_{a \neq b} \bar{B}_{ab}$ requires denser edge connections within communities than between them.
- The tuning γ_n trades off the information from two different sources : network edge and nodal covariates. The specification $\gamma_n = O(n^{-1})$ in Theorem 4 implies that a larger weight should be given to network edge information.

Solving (13) via alternating direction method of multipliers (ADMM) (Boyd et al., 2011).

- Input: initialize $Z^0 = A + \gamma_n XX^T$, $W^0 = Y^0 = U^0 = V^0 = 0$, number of iterations $\mathcal{T}$, step size ξ.
- For $t = 0, \ldots, \mathcal{T} - 1$
  - (a) $Y^{t+1} = \min\{\max\{0, \frac{1}{2}(W^t + Z^t - U^t - V^t)\}, 1\}$
  - (b) $W^{t+1} = Y^{t+1} + U^t + \frac{\lambda_n - \sum_{ij}(Y_{ij}^{t+1} + U_{ij}^t)}{n^2}\mathbf{1}\mathbf{1}^T$
  - (c) $Z^{t+1} = P\Lambda_+ P^T$, where $Y^{t+1} + V^t + \xi^{-1} \cdot (A + \gamma_n XX^T) = P\Lambda P^T$
  - (d) $U^{t+1} = U^t + Y^{t+1} - W^{t+1}$, $V^{t+1} = V^t + Y^{t+1} - Z^{t+1}$
- Output $Z^{\mathcal{T}+1}$