Collaborative Spectral Clustering in Attributed Networks

Xiaodong Jiang and Pengsheng Ji

UNIVERSITY OF GEORGIA

Introduction

We propose a novel spectral clustering algorithm for attributed networks, where each node has p-dimensional meta-covariates from various formats such as text, image, speech, etc. The connectivity matrix $W_{n\times n}$ is constructed with the adjacency matrix $A_{n\times n}$ and covariate matrix $X_{n\times p}$, and $W = (1-\alpha)A + \alpha K(X, X')$, where $\alpha \in [0, 1]$ and K is a kernel to measure the covariate similarities. We then perform the classical k-means algorithm on the element-wise ratio matrix of the first K leading eigenvector of W. Theoretical and simulation studies show the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in unbalanced networks where most community detection algorithms fail.

Traditional Community Detection

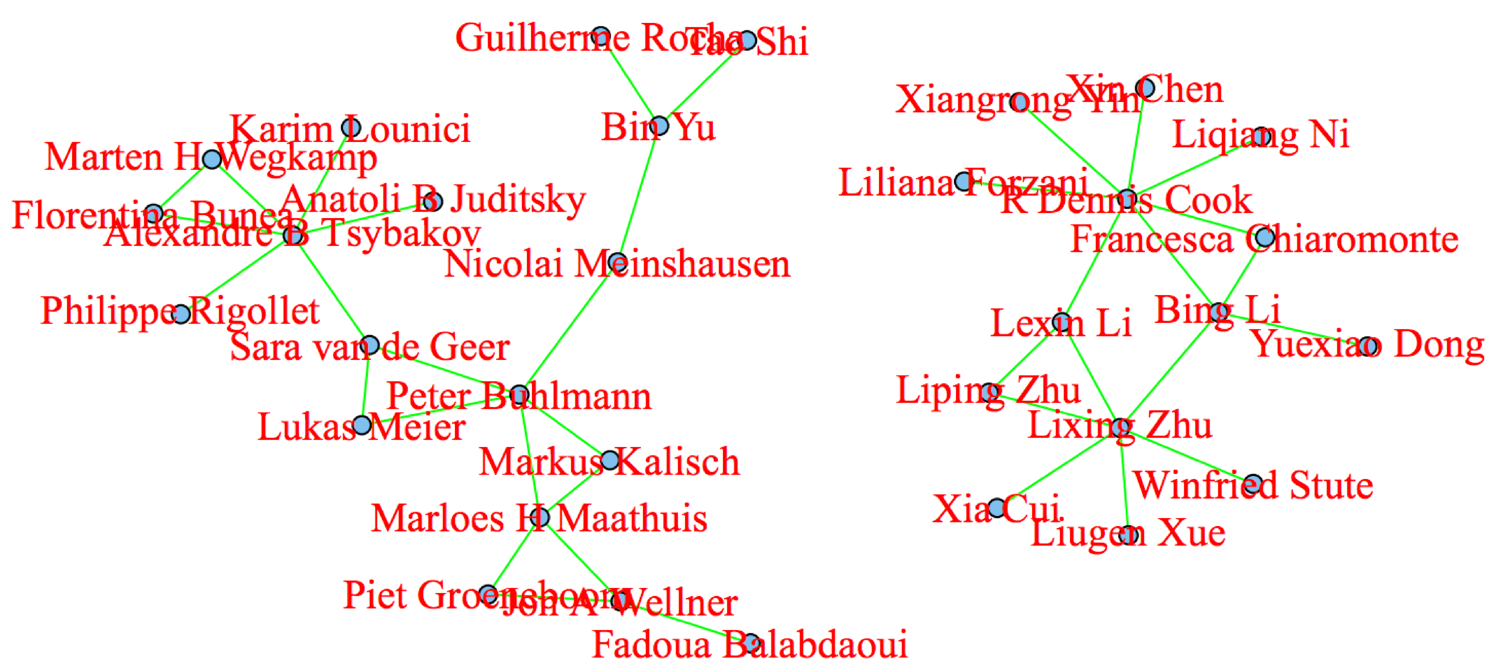

Figure 1: Theoretical ML and Dimension Reduction, Ji and Jin (2016)

Attributed Network Structure

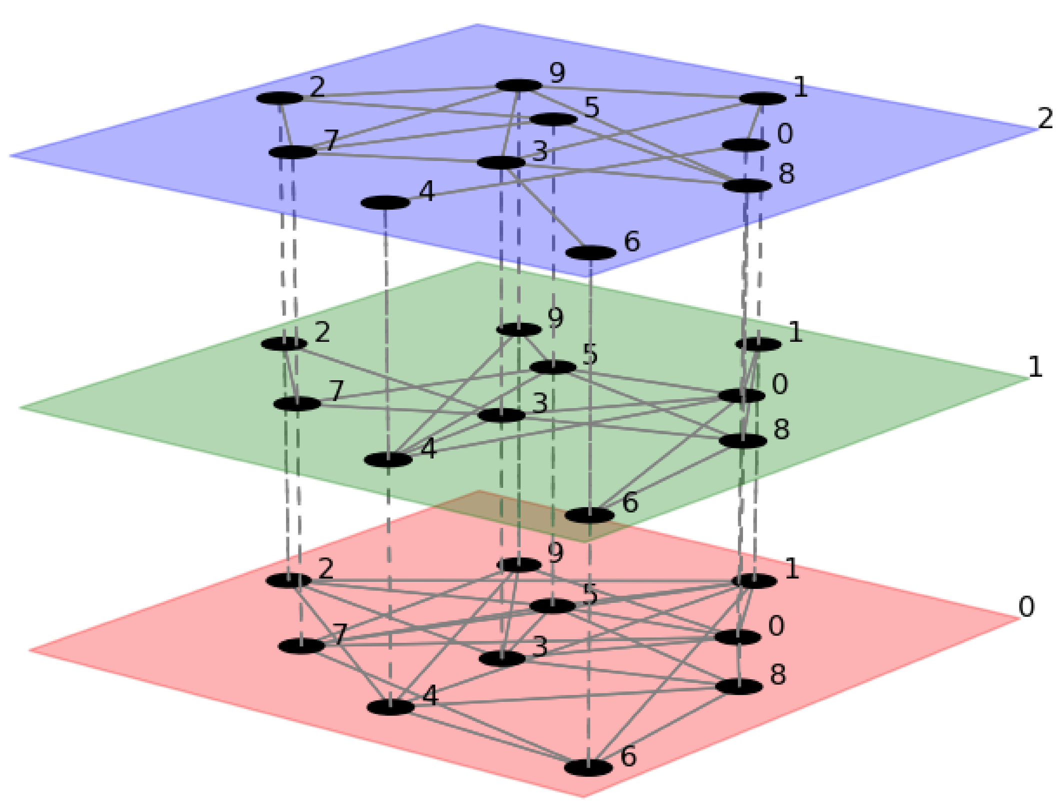

Figure 2: Network Structure with Node Attributes

Node Attributed Stochastic Block Model (NSBM)

Let f be a mixture of R p-dimensional distributions

$$f(x) = \sum_{r=1}^{R} \lambda_r f_r(x; \psi_r)$$

The adjacent and node attribute matrices of nSBM are generated as follows,

a. The adjacent matrix $A = (a_{ij})_{n\times n}$ is generated as $a_{ij} \sim Bernoulli(P_{g_i g_j})$ independently for $i \neq j$, otherwise 0.

b. The $n \times p$ node attribute matrix X is generated as $X_{i\cdot} \sim f_r$ if $g_i = r$.

Node Attributed Degree-Corrected Model (NDCBM)

The adjacency matrix and node attribute matrix are generated as follows:

a. The adjacency matrix $A = (a_{ij})_{n\times n}$ is generated as

$$a_{ij} \sim \text{Bernoulli}(\theta_i \theta_j P_{g_i g_j})$$

independently for $i \neq j$, otherwise 0.

b. The $n \times p$ node attribute matrix X is generated as $X_{i\cdot} \sim f_r$ if $g_i = r$.

The Algorithm - Collaborative Spectral Clustering

Algorithm 1 CSC with Row-Normalization

procedure CSC(A, X, α, R)

Obtain the sum of column variance

$$\hat{\sigma}^2 = \sum_{j=1}^{p} \text{Var}[x_{\cdot j}]$$

Calculate $K = (k(x_{i\cdot}, x_{j\cdot}))_{n\times n}$, where

$$k(x_{i\cdot}, x_{j\cdot}) = \exp(-\frac{||x_{i\cdot} - x_{j\cdot}||^2}{2\hat{\sigma}^2})$$

Obtain leading eigenvectors $U = \{u_1, ..., u_R\}$ of $W = (1-\alpha)A + \alpha K$

Obtain U^* s.t.

$$U^*(i, j) = \frac{U(i, j)}{\sqrt{\sum_{j=1}^{R} U^2(i, j)}}$$

Apply k-means to U^*.

end procedure

Theoretical Results

Lemma 1. *(Principal subspace perturbation bound) Let W and $E(W)$ have eigenvalues $\hat{\lambda}_1, ..., \hat{\lambda}_n$ and $\lambda_1, ..., \lambda_n$ respectively. Let the first R leading eigenvectors corresponding to the R largest leading eigenvalues be $\hat{U}$ and U for W and $E[W]$, then there exists an orthogonal matrix $\hat{O}$, such that,*

$$||\hat{U}\hat{O} - U||_F \leq \frac{2^{3/2}\sqrt{nr}\max\{C\sqrt{n}\sqrt{d}, c\sqrt{\frac{\log p}{p}}\}}{\min\{\lambda_{R-1} - \lambda_R, \lambda_R - \lambda_{R+1}\}}$$

Theorem 1. *(Concentration bound on connectivity matrix under NSBM) Let $W = (1-\alpha)A + \alpha K$, then $||W - E[W]||_\infty \leq (1-\alpha)C\sqrt{n}\sqrt{d} + \alpha c\sqrt{\frac{\log p}{p}}$ with probability at least $1 - \max\{n^{-r}, n^2 p^{-\rho c^2}\}$*

Theorem 2. *(Error bound of k-means on leading eigenvectors) Under NSBM with Gaussian distributions in $\mathcal{F}$, the error bound of k-means on the first R leading eigenvectors is*

$$\frac{||Z||}{N} \leq \frac{64mnr\max\{C\sqrt{n}\sqrt{d}, c\sqrt{\frac{\log p}{p}}\}^2}{N\min\{\lambda_{R-1} - \lambda_R, \lambda_R - \lambda_{R+1}\}^2}$$

where $m = \max(M^T M)_{ii}$.

Results with Paper Citation Network with Abstracts

Table 2: Community detection results from *CSC-SCORE* with $\alpha = 0.8$

ID	Title	Author	Year
1	Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models	Omiros Papaspiliopoulos and Roberts	2007
1	An ANOVA model for dependent random measures	Iorio et al	2004
1	Bayesian nonparametric spatial modeling with Dirichlet process mixing	Gelfand et al	2005
1	Hierarchical Dirichlet processes	Teh et al	2005
1	Bayesian density regression	Dunson and Pillai	2007
1	A method for combining inference across related nonparametric Bayesian models	Müller et al	2004
2	Empirical Bayes selection of wavelet thresholds	Johnstone	2005
2	Covariance matrix selection and estimation via penalised normal likelihood	Huang et al	2006
2	New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis	Fan and Li	2004
2	One-step sparse estimates in nonconcave penalized likelihood models	Zou and Li	2008
2	Nonconcave penalized likelihood with a diverging number of parameters	Fan and Peng	2004
3	A stochastic process approach to false discovery control	Genovese and Wasserman	2004
3	Regularization and variable selection via the elastic net	Zou and Hastie	2005
3	The Dantzig selector: statistical estimation when p is much larger than n	Candes and Tao	2005
3	High-dimensional graphs and variable selection with the lasso	Meinshausen and Bühlmann	2006
3	The adaptive lasso and its oracle properties	Zou	2006

Three communities are well built, where Community 1 is *Bayesian Statistics*, Community 2 is *Nonparametrics*, and Community 3 is lasso related group.

Figure 3: Results with Paper Citation Network with Abstracts

Conclusions

- In this work we proposed a novel and flexible model for node-attributed network data for both degree-free and degree-corrected versions.
- We proposed two types of algorithms to perform clustering on node-attributed network data.
- We also provided theoretical guarantees for the performance of our algorithm.
- Our algorithm outperform all existing works in extensive simulation studies. We also tested with real word data including paper-paper citation network with abstracts.

Contacts

- Xiaodong Jiang (xiaodong@uga.edu), Department of Statistics, University of Georgia
- Pengsheng Ji (psji@uga.edu), Department of Statistics, University of Georgia