Community Preserving Network Embedding

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, Shiqiang Yang

Mesoscopic community structure:

which is one of the most prominent feature of networks, is largely ignored

Modularized Nonnegative Matrix Factorization (M-NMF)

incorporate the community structure into network embedding

M-NMF

microscopic structure:

first- and second-order proximities of nodes

mesoscopic structure:

the communities are detected by a modularity constraint term

Application:

node clustering and classification

M-NMF Model:

undirected network $G = (V, E)$ n nodes e edges,

adjacency matrix $\mathbf{A} = [A_{ij}] \in \mathbb{R}^{n \times n}$,

representations of nodes $\mathbf{U} \in \mathbb{R}^{n \times m}$, m $(m \leq n)$

Modularity algorithms:

Newman, M. E. 2006b. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103(23):8577 - 8582

http://www.pnas.org/content/pnas/103/23/8577.full.pdf

Modeling community structure:

Two communities:

$$Q = \frac{1}{4e}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2e})h_i h_j, \qquad (1)$$

where k_i is the degree of node i and $h_i = 1$ if node i belongs to the first community, otherwise, $h_i = -1$.

the modularity matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ $B_{ij} = A_{ij} - \frac{k_i k_j}{2e}$, $Q = \frac{1}{4e}\mathbf{h}^T\mathbf{B}\mathbf{h}$,

K communities(k>2):

$$Q = \frac{1}{2m}\sum_{vw}[A_{vw} - \frac{k_v k_w}{2m}]\delta(c_v, c_w)$$

$$Q = tr(\mathbf{H}^T\mathbf{B}\mathbf{H}), \quad s.t. \quad tr(\mathbf{H}^T\mathbf{H}) = n,$$

Modeling microscopic structure：

First-order proximity $\mathbf{S}^{(1)} = [S_{ij}^{(1)}]$ the adjacency matrix $\mathbf{A}$

Second-order Proximity $\mathbf{S}^{(2)} = [S_{ij}^{(2)}]$ $S_{ij}^{(2)} = \frac{\mathcal{N}_i \mathcal{N}_j}{\|\mathcal{N}_i\| \|\mathcal{N}_j\|}$,

$$\mathbf{S} = \mathbf{S}^{(1)} + \eta \mathbf{S}^{(2)} \quad \eta > 0$$

$$\min \|\mathbf{S} - \mathbf{M}\mathbf{U}^T\|_F^2 \quad s.t. \quad \mathbf{M} \geq 0, \quad \mathbf{U} \geq 0. \quad (3)$$

nonnegative basis matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$

nonnegative representation matrix $\mathbf{U} \in \mathbb{R}^{n \times m}$,

The unified network embedding model：

$$\min_{\mathbf{M},\mathbf{U},\mathbf{H},\mathbf{C}} \|\mathbf{S}-\mathbf{M}\mathbf{U}^T\|_F^2 + \alpha\|\mathbf{H}-\mathbf{U}\mathbf{C}^T\|_F^2 - \beta tr(\mathbf{H}^T\mathbf{B}\mathbf{H})$$
$$s.t., \mathbf{M} \geqslant 0, \mathbf{U} \geqslant 0, \mathbf{H} \geqslant 0, \mathbf{C} \geqslant 0, tr(\mathbf{H}^T\mathbf{H}) = n, \quad (4)$$

an auxiliary nonnegative matrix $\mathbf{C} \in \mathbb{R}^{k\times m}$, named community representation matrix, where the r-th row ($\mathbf{C}_r$) is the representation of community r.

Optimization：

The objective function (4) is not convex, and we separate the optimization of (4) to four subproblems and iteratively optimize them, which guarantees each subproblem converges to the local minima.

M-subproblem: Updating $\mathbf{M}$ with other parameters in (4) fixed leads to a standard NMF formulation (Lee and Seung 2001), so the updating rule for $\mathbf{M}$ is

$$\mathbf{M} \leftarrow \mathbf{M} \odot \frac{\mathbf{S}\mathbf{U}}{\mathbf{M}\mathbf{U}^T\mathbf{U}}. \quad (5)$$

C-subproblem: Updating $\mathbf{C}$ with other parameters in (4) fixed also leads to a standard NMF formulation, so the updating rule of $\mathbf{C}$ is

$$\mathbf{C} \leftarrow \mathbf{C} \odot \frac{\mathbf{H}^T\mathbf{U}}{\mathbf{C}\mathbf{U}^T\mathbf{U}}. \tag{7}$$

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*

https://link.springer.com/content/pdf/10.1007%2F978-1-4471-5571-3.pdf

$$V \approx WH$$

Theorem 1 *The Euclidean distance* $||V - WH||$ *is nonincreasing under the update rules*

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{4}$$

U-subproblem: Updating $\mathbf{U}$ with other parameters in (4) fixed leads to a joint NMF problem (Akata, Thurau, and Bauckhage 2011), whose updating rule is

$$\mathbf{U} \leftarrow \mathbf{U} \odot \frac{\mathbf{S}^T\mathbf{M} + \alpha\mathbf{H}\mathbf{C}}{\mathbf{U}(\mathbf{M}^T\mathbf{M} + \alpha\mathbf{C}^T\mathbf{C})}. \tag{6}$$

Akata, Z.; Thurau, C.; and Bauckhage, C. 2011. Non-negative matrix factorization in multimodality data for segmentation and label prediction. In 16th Computer vision winter workshop.

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \quad \text{and} \quad \boldsymbol{Y} \approx \boldsymbol{V}\boldsymbol{H} \tag{10}$$

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{H}} (1-\lambda)\left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\right\|^2 + \lambda\left\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{H}\right\|^2 \\ &\quad \text{s.t. } \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{H} \succeq \mathbf{0} \end{aligned} \tag{11}$$

$$\boldsymbol{H} = \boldsymbol{H} \odot \frac{(1-\lambda)\boldsymbol{W}^T\boldsymbol{X} + \lambda\boldsymbol{V}^T\boldsymbol{Y}}{((1-\lambda)\boldsymbol{W}^T\boldsymbol{W} + \lambda\boldsymbol{V}^T\boldsymbol{V})\boldsymbol{H}}. \tag{13}$$

H-subproblem: when update $\mathbf{H}$ with other parameters in (4) fixed, we need to solve the following function:

$$\begin{aligned}\min_{\mathbf{H}\geq 0} L(\mathbf{H}) =& \alpha\|\mathbf{H}-\mathbf{U}\mathbf{C}^T\|_F^2 - \beta tr(\mathbf{H}^T(\mathbf{A}-\mathbf{B_1})\mathbf{H}),\\ & s.t. \quad tr(\mathbf{H}^T\mathbf{H}) = n,\end{aligned} \tag{8}$$

where the element in $\mathbf{B}_1$ is $\frac{k_i k_j}{2e}$. Recall that $\mathbf{H}$ is the community indicator matrix, and the constraint makes the optimization of (8) an NP-hard problem. Instead, we relax the constraint to $\mathbf{H}^T\mathbf{H} = \mathbf{I}$. Finally, by introducing a regularization coefficient λ for $\mathbf{H}^T\mathbf{H} = \mathbf{I}$, we transform (8) to the following function:

$$\begin{aligned}\min_{\mathbf{H}\geq 0} L(\mathbf{H}) =& -\beta tr(\mathbf{H}^T\mathbf{A}\mathbf{H}) + \beta tr(\mathbf{H}^T\mathbf{B_1}\mathbf{H})\\ & + \alpha\|\mathbf{H}-\mathbf{U}\mathbf{C}^T\|_F^2 + \lambda\|\mathbf{H}^T\mathbf{H}-\mathbf{I}\|_F^2,\end{aligned} \tag{9}$$

where $\lambda > 0$ should be large enough

Lagrange multiplier matrix $\boldsymbol{\Theta} = [\Theta_{ij}]$

$$\|\mathbf{X}\|_F^2 = tr(\mathbf{X}^T\mathbf{X})$$

$$\begin{aligned}
L'(\mathbf{H}) = &- \beta tr(\mathbf{H}^T\mathbf{A}\mathbf{H}) + \beta tr(\mathbf{H}^T\mathbf{B_1}\mathbf{H}) \\
&+ \alpha tr(\mathbf{H}\mathbf{H}^T - 2\mathbf{H}\mathbf{C}\mathbf{U}^T + \mathbf{U}\mathbf{C}^T\mathbf{C}\mathbf{U}^T) \\
&+ \lambda tr(\mathbf{H}^T\mathbf{H}\mathbf{H}^T\mathbf{H} - 2\mathbf{H}^T\mathbf{H} + \mathbf{I}) + tr(\mathbf{\Theta}\mathbf{H}^T).
\end{aligned} \tag{10}$$

Set derivative of $L'(\mathbf{H})$ with respect to $\mathbf{H}$ to 0, we have:

$$\begin{aligned}
\mathbf{\Theta} =& 2\beta\mathbf{A}\mathbf{H} - 2\beta\mathbf{B_1}\mathbf{H} - 2\alpha\mathbf{H} + 2\alpha\mathbf{U}\mathbf{C}^T \\
&- 4\lambda\mathbf{H}\mathbf{H}^T\mathbf{H} + 4\lambda\mathbf{H}.
\end{aligned} \tag{11}$$

Following the Karush-Kuhn-Tucker (KKT) condition for the nonnegativity of $\mathbf{H}$, we have the following equation:

$$\begin{aligned}
&(2\beta\mathbf{A}\mathbf{H} - 2\beta\mathbf{B_1}\mathbf{H} - 2\alpha\mathbf{H} + 2\alpha\mathbf{U}\mathbf{C}^T \\
&- 4\lambda\mathbf{H}\mathbf{H}^T\mathbf{H} + 4\lambda\mathbf{H})_{ij}H_{ij} = \Theta_{ij}H_{ij} = 0.
\end{aligned} \tag{12}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \sqrt{\frac{-2\beta\mathbf{B_1}\mathbf{H} + \sqrt{\Delta}}{8\lambda\mathbf{H}\mathbf{H}^T\mathbf{H}}}, \tag{13}$$

where $\Delta = 2\beta(\mathbf{B_1}\mathbf{H}) \odot 2\beta(\mathbf{B_1}\mathbf{H}) + 16\lambda(\mathbf{H}\mathbf{H}^T\mathbf{H}) \odot (2\beta\mathbf{A}\mathbf{H} + 2\alpha\mathbf{U}\mathbf{C}^T + (4\lambda - 2\alpha)\mathbf{H})$.

Complexity analysis：

$\mathcal{O}(n^2m + nm^2), \mathcal{O}(nm^2 + n^2m + m^2k), \mathcal{O}(kmn)$ and $\mathcal{O}(n^2k + k^2n + mnk)$,

$$\mathcal{O}(n^2m + n^2k)$$

Experimental evaluations

Dataset: WebKB network consists of 4 subnetworks with 877 webpages and 1608 edges.

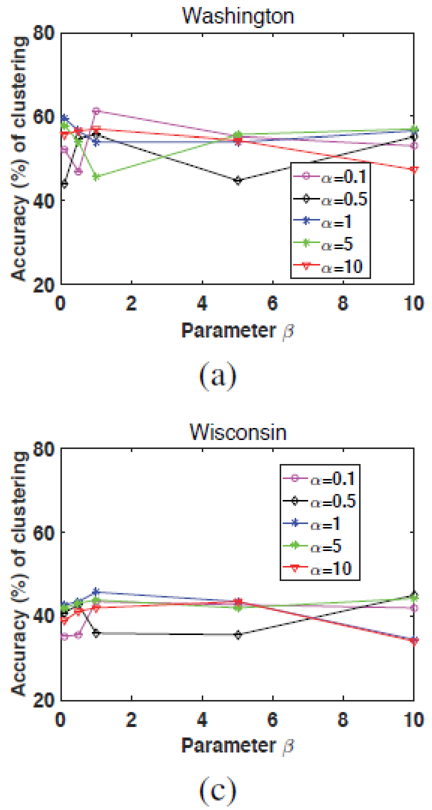
http://linqs.cs.umd.edu/projects/projects/lbc/

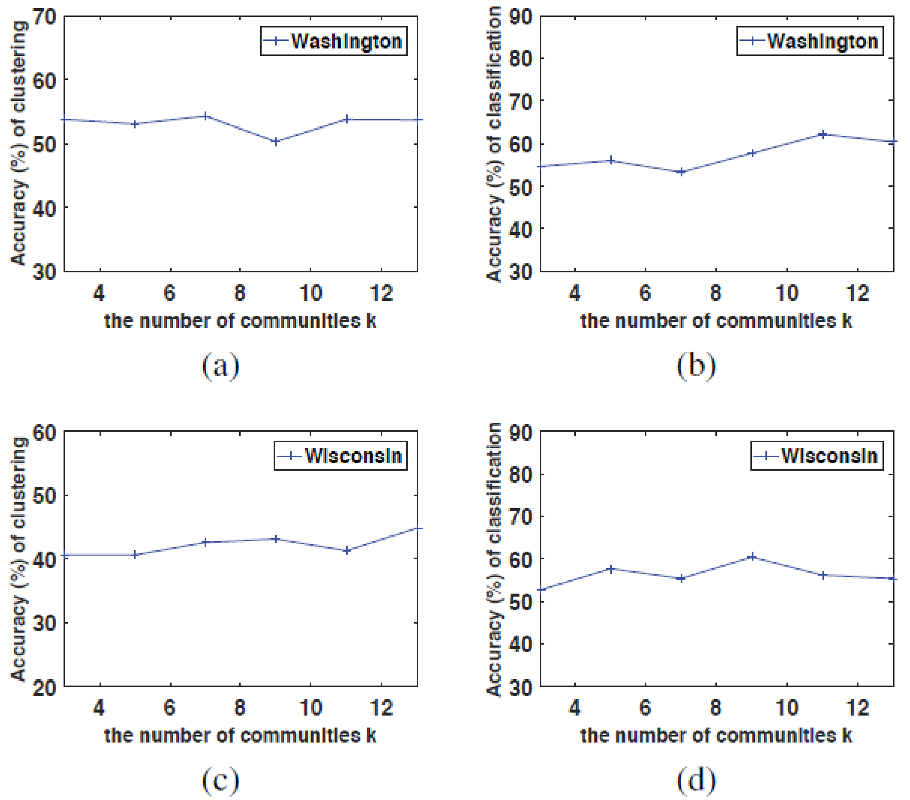
Table 1: Accuracy (%) of node clustering (bold numbers represent the best results).

Methods	DeepWalk	LINE1	LINE2	GraRep	Node2Vec	M-NMF0	M-NMF
Cornell	32.82	35.38	42.56	33.85	34.36	40.00	**43.05**
Texas	37.97	40.64	55.61	35.29	50.27	47.06	**63.10**
Washington	35.65	38.70	53.48	36.52	41.74	55.65	**59.57**
Wisconsin	34.34	35.09	43.77	36.60	35.47	42.64	**45.66**
Polblogs	52.68	57.38	63.88	53.42	**84.83**	72.75	82.82
Amherst	10.34	42.36	44.38	46.41	41.66	43.54	**47.25**
Hamilton	10.15	33.47	31.30	38.81	35.41	38.34	**42.49**
Mich	11.66	15.58	14.63	**35.12**	14.05	29.66	31.50
Rochester	7.94	17.88	16.86	33.80	18.00	30.35	**38.09**

Table 2: Accuracy (%) of node classification (bold numbers represent the best results).

Methods	DeepWalk	LINE1	LINE2	GraRep	Node2Vec	M-NMF0	M-NMF
Cornell	24.10	27.69	44.62	45.38	38.46	27.69	**47.18**
Texas	22.63	34.21	**73.16**	68.42	51.05	47.89	70.00
Washington	24.44	25.33	50.22	52.00	53.78	54.67	**63.56**
Wisconsin	26.15	28.46	51.54	59.62	44.62	39.62	**61.15**
Polblogs	64.77	83.02	80.87	89.60	84.03	80.20	**90.67**
Amherst	41.59	91.51	87.99	91.46	89.73	87.74	**92.00**
Hamilton	39.95	91.64	87.27	91.64	91.45	89.36	**92.92**
Mich	25.44	62.09	60.75	60.79	61.98	58.15	**62.26**
Rochester	34.78	87.04	84.23	85.47	83.65	84.28	**87.18**





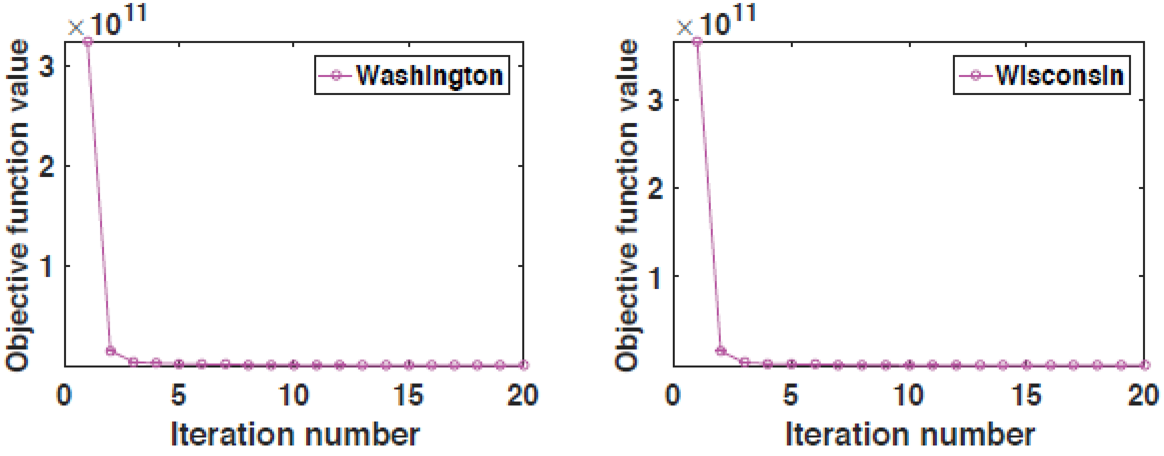


Figure 3: Convergence analysis.