mathematics

MDPI

Article

Community Detection of Multi-Layer Attributed Networks via Penalized Alternating Factorization

Jun Liu, Jiangzhou Wang and Binghui Liu *

KLAS of MOE & School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China; liuj292@nenu.edu.cn (J.L.); wangjz695@nenu.edu.cn (J.W.)

* Correspondence: liubh100@nenu.edu.cn

Received: 1 January 2020; Accepted: 8 February 2020; Published: 13 February 2020

Abstract: Communities are often associated with important structural characteristics of a complex network system, therefore detecting communities is considered to be a fundamental problem in network analysis. With the development of data collection technology and platform, more and more sources of network data are acquired, which makes the form of network as well as the related data more complex. To achieve integrative community detection of a multi-layer attributed network that involves multiple network layers together with their attribute data, effectively utilizing the information from the multiple networks and the attributes may greatly enhance the accuracy of community detection. To this end, in this article, we study the integrative community detection problem of a multi-layer attributed network from the perspective of matrix factorization, and propose a penalized alternative factorization (PAF) algorithm to resolve the corresponding optimization problem, followed by the convergence analysis of the PAF algorithm. Results of the numerical study, as well as an empirical analysis, demonstrate the advantages of the PAF algorithm in community discovery accuracy and compatibility with multiple types of network-related data.

Keywords: community detection; matrix factorization; multi-layer attributed network; penalized alternating factorization

1. Introduction

Network science is one of the most active research fields in recent years [1], which has been successfully applied in many fields, including the social science to study social relationships among individuals [2], biology to study interactions among genes and proteins [3], neuroscience to study the structure and function of the brain [4], and so on. Networks can represent and analyze the relational structure among interacting units of a complex system, and in many cases, the units of a network can be divided into groups with the property that there are many edges between units in the same group, but relatively few edges between units in different groups. Such groups are known as communities, which are often associated with important structural characteristics of a complex system. [5,6].

For example, in social networks, communities can correspond to groups with common interests [7,8]. In World Wide Web networks, communities can correspond to webpages with related topics [9]; in brain networks, they can correspond to specialized functional components [10]; and in protein–protein interaction networks, they can correspond to groups of proteins that contribute to the same cellular function [11]. Communities are often useful for understanding the essential functionality and organizational principles of networks. Therefore, community detection is considered a fundamental problem in understanding and analyzing networks [6].

Community detection has been widely studied in many application fields since the 1980s. Various models and algorithms have been developed in different fields, such as machine learning, network science, social science, and statistical physics. Community detection is a computationally

challenging problem because the number of possible partitions of nodes into nonoverlapping groups is non-polynomial in the size of a network, especially in large networks. To deal with this challenging problem, a large number of algorithmic approaches have been proposed [12–16], including various greedy algorithms, such as hierarchical clustering [17], graph partitioning [18], and the methods based on optimizing a global criterion over all possible partitions, such as normalized cuts [19] and modularity [20]. Other algorithmic approaches include spectral methods [21–24], semi-definite programming [25,26], low-rank approximation [27], and non-negative matrix factorization [28].

Recently the quantities and types of network-related data are rising very fast, as data collection technologies or platforms rapidly evolve. Consequently, a large number of studies on community detection for various types of network-related data have been conducted, which will be introduced as follows according to the type of networks they are targeting.

First, for a single network, a number of approaches to community detection have been proposed based on probabilistic models for networks with communities, such as the stochastic block model [29], the degree-corrected stochastic block model [30], and the latent factor model [31]. Other approaches by optimizing a criterion measuring the strength of community structure in some sense also have appeared, often through non-negative matrix factorization [28] and spectral approximations, such as normalized cuts [19], modularity [20,32], and many variants of spectral clustering [33,34].

For attributed network, i.e., a network together with its attribute data, several generative models for jointly modeling the edges and the attributes have been proposed, including the network random effects model [35], the embedding feature model [36], the latent variable model [37], the discriminative approach [38], the latent multi-group membership graph model [39], the social circles model for ego networks [40], the communities from edge structure and node attributes model [41], the Bayesian graph clustering model [42], the topical communities and personal interest model [43], the modified stochastic block model [44], and a criterion-based method [45].

Multi-layer network involves networks from interdependent but distinct sources [46,47], which can be simultaneously collected for a certain group of units [48]. Community detection of this type of network has been applied to a variety of problems [49,50], including clustering of temporal networks through a dynamic stochastic block model [51], modeling and analysis of air transportation routes [52], studying individuals with multiple sociometric relations [53,54], and analyzing relationships between social interactions and economic exchange [55].

In addition, in many real network data analysis, there will be a more complex or general type of network, named multi-layer attributed network, which involves multiple network layers together with their attribute data. If the multiple networks share a common community structure and the distribution of unit attributes is also correlated with this community structure, then an integrative community detection approach that can integrate information from the multiple networks as well as the attributes may make better use of all these network-related data, therefore increase the accuracy of community detection as much as possible. Unfortunately, the research on multi-layer attributed network is still in its infancy, which leads us to further explore the problem description and corresponding solution of its community detection.

To this end, in this article, we employ the framework of integrative matrix factorization to formulate and achieve community detection of a multi-layer attributed network, which can be compatible with all the special cases of a multi-layer attributed network: the single network, attributed network, and the multi-layer network. In pursuit of community discovery accuracy and compatibility with multiple types of network-related data, we propose to use the penalized alternative factorization, named the PAF algorithm, to resolve the corresponding optimization problem.

The rest of this article is organized as follows. We elaborate the community detection problem of multi-layer attributed network in Section 2, and present the PAF algorithm to learn communities in Section 3, followed by the theoretical analysis of PAF in Section 4. The numerical performance of PAF is demonstrated in Section 5, and an empirical analysis is demonstrated in Section 6. Finally, we conclude this article in Section 7 and relegate the technical proofs to Appendies A and B.

2. Problem Formulation

In this section, we will describe in detail the problems of community detection based on matrix factorization from the single network to the multi-layer attributed networks.

2.1. Single Network

Let $G = (N, E)$ denote a single network, where $N = \{1, \ldots, n\}$ is the node set that represents the units of the modeled system, and $E \subseteq N \times N$ is the edge set containing all pairs of nodes (u, v) such that nodes u and v share a social, physical, or functional relationship, where $N \times N$ denotes the Cartesian product of N and N. A network G can be characterized by an $n \times n$ adjacency matrix $A = (A_{ij})$ with each $A_{ij} \in \{0, 1\}$, where $A_{ij} = 1$ means that there exists an edge from nodes i to j in network G; otherwise, is not. The purpose of community detection is to identify a partition of N with community structure via the observed adjacency matrix A. Due to numerous definitions of communities, there are numerous approaches to implementing community detection. In view of the simplicity and effectiveness of a matrix factorization approach, in this article we consider the problem of community detection based on the framework of matrix factorization.

In the framework of matrix factorization, the problem of community detection, given a predetermined number of communities k^*, can be formulated as the following optimization problem,

$$\min_{\substack{C \in \mathcal{R}^{n \times k^*} \\ S \in \mathcal{R}^{k^* \times k^*}}} \|A - CSC^T\|_F^2, \tag{1}$$

where C is the unknown $n \times k^*$ matrix used to find k^* communities, S is the unknown $k^* \times k^*$ weight matrix, and $\|\cdot\|_F$ denotes the Frobenius norm. This optimization problem is the same as that studied in [28], except that in our optimization problem the non-negative constraints on matrix elements of C and S are removed to improve computational efficiency. The matrix C can be viewed as the community label matrix of A. By treating each row of C as a point in $\mathbb{R}^{k^*}$, we divide these points into k^* clusters via k-means or any other clustering algorithm. Then, we assign the network node $i \in N$ to community $k \in \{1, \ldots, k^*\}$ if and only if row i of matrix C is assigned to cluster k.

From a statistical point of view, we find that the above optimization problem (1) is closely related to the well-known stochastic block model (SBM) [29]. Specifically, under the k^*-community SBM with the $n \times k^*$ ground truth label matrix $C^{(0)}$ and the $k^* \times k^*$ connectivity probability matrix $S^{(0)}$, once the diagonal elements of the adjacency matrix A are also considered as random terms, not fixed to be zero, then the conditional expectation of A given $C^{(0)}$ is

$$\mathbb{E}(A|C^{(0)}) = C^{(0)} S^{(0)} {C^{(0)}}^T. \tag{2}$$

Similar to the least-squares method, the ground truth labels $C^{(0)}$ can be predicted by minimizing the sum of squares of the observations A_{ij}'s and their conditional expectation $\mathbb{E}(A_{ij}|C)$'s:

$$\min_{\substack{C \in \{0,1\}^{n \times k^*} \\ S \in [0,1]^{k^* \times k^*}}} \|A - \mathbb{E}(A|C)\|_F^2 = \min_{\substack{C \in \{0,1\}^{n \times k^*} \\ S \in [0,1]^{k^* \times k^*}}} \|A - CSC^T\|_F^2, \tag{3}$$

subject to $\sum_{j=1}^{k^*} C_{ij} = 1$ for each $i \in \{1, \ldots, n\}$. This minimization problem is very hard to achieve, as the range of C, $\{C \in \{0,1\}^{n \times k^*} : \sum_{j=1}^{k^*} C_{ij} = 1\}$, includes ${k^*}^n$ values. Consequently, to make the corresponding calculation feasible, (3) may be relaxed into (1), if the accuracy of community recovery can be guaranteed. Note that in (1), the ranges of C and S are relaxed into the Euclidean spaces $\mathbb{R}^{n \times k^*}$ and $\mathbb{R}^{k^* \times k^*}$, respectively, whereas in other methods, such as the non-negative matrix factorization methods [28], the ranges can be relaxed into $\mathbb{R}^{+^{n \times k^*}}$ and $\mathbb{R}^{+^{k^* \times k^*}}$. Here, we remove the non-negative constraints to improve computational efficiency and compatibility of the proposed method, which will be explained in the following section.

2.2. Multi-Layer Attributed Network

Once the structural information from multiple sources and the attribution information of the network nodes can be collected together, we will consider the so-called multi-layer attributed network, which is written as $G_{Att}^{Mul} = (N, E^{(1)}, \ldots, E^{(m^*)}, X)$ and characterized by m^* $n \times n$ adjacent matrices $\{A^{(1)}, \ldots, A^{(m^*)}\}$ as well as an $n \times p$ attribution matrix X. This is a unified framework, which can include the single network, multi-layer network, and attributed network. To achieve community detection of G_{Att}^{Mul}, we study the following integrative matrix factorization problem,

$$\min_{\substack{C \in R^{n \times k^*} \\ S^{(1)}, \ldots, S^{(M)} \in \mathbb{R}^{k^* \times k^*} \\ V \in \mathbb{R}^{p \times k^*}}} \sum_{m=1}^{m^*} \omega_m \| A^{(m)} - C S^{(m)} C^T \|_F^2 + \omega_0 \| X - C V^T \|_F^2, \tag{4}$$

where $\{\omega_m\}_{m=0}^{m^*}$ with $\sum_{m=0}^{m^*} \omega_m = 1$ are the weight parameters specified beforehand and V is a $p \times k^*$ matrix as the right part of the matrix factorization of X. Similarly, to solve (4), we consider the following approximate minimization problem,

$$\begin{aligned} \min_{\substack{C \in R^{n \times k^*} \\ S^{(1)}, \ldots, S^{(m^*)} \in \mathbb{R}^{k^* \times k^*} \\ V \in \mathbb{R}^{p \times k^*}}} & \sum_{m=1}^{m^*} \omega_m \| A^{(m)} - C^{(1)} S^{(m)} C^{(2)} \|_F^2 + \frac{\omega_0}{2} \sum_{t=1}^{2} \| X - C^{(t)} V^T \|_F^2 + \lambda \| C^{(1)} - C^{(2)} \|_F^2 \\ & + \nu \{ \| C^{(1)} \|_F^2 + \| C^{(2)} \|_F^2 + \| V \|_F^2 + \sum_{m=1}^{m^*} \| S^{(m)} \|_F^2 \}. \end{aligned} \tag{5}$$

Note that throughout this section, the weights $\{\omega_m\}_{m=0}^{m^*}$ need to be given beforehand by the users according to background knowledge. To determine these weights, one user may have to take into account the importance and scale of data from each source. If no additional information is available, for simplicity, the weights can be equally distributed.

3. Learning Algorithm

We present a penalized alternating factorization (PAF) scheme to minimize (5). In particular, the objective function is minimized step by step by fixing any $m^* + 2$ matrices in $\{C^{(1)}, C^{(2)}, S^{(1)}, \ldots, S^{(m^*)}, V\}$ and then optimizing the objective function with respect to the remaining one. The algorithm is described in details as follows.

Algorithm 1 Penalized Alternating Factorization (PAF) Algorithm.

Input: m^* $n \times n$ adjacent matrices $\{A^{(1)}, \ldots, A^{(m^*)}\}$, an $n \times p$ attribution matrix X, the number of communities k^*.

Output: a length-n community label vector $L = (L_1, \ldots, L_n)$.

1: Initialization:

2: (a) $t = 0$; $C^{(1,-1)}$ and $C^{(2,-1)}$ are both $n \times k^*$ zero matrix.

3: (b) apply SCP, the spectral clustering with perturbations [33], to $A^* = \sum_{m=1}^{m^*} \omega_m A^{(m)}$ and find k^* initial communities, transform the resulting length-n community label vector into the $n \times k^*$ community label matrix, then make $C^{(1,0)}$ and $C^{(2,0)}$ equal to this initial community label matrix, where $C^{(1,0)}, C^{(2,0)} \in R^{n \times k^*}$ are initial chooses of $C^{(1)}, C^{(2)}$ respectively.

4: (c) let

$$\begin{aligned} V^{(0)} &= X^T \left[C^{(1,0)} + C^{(2,0)} \right] \left[{C^{(1,0)}}^T C^{(1,0)} + {C^{(2,0)}}^T C^{(2,0)} \right]^{-1}, \\ S^{(m,0)} &= \left[{C^{(1,0)}}^T C^{(1,0)} \right]^{-1} {C^{(1,0)}}^T A^{(m)} C^{(2,0)} \left[{C^{(2,0)}}^T C^{(2,0)} \right]^{-1}. \end{aligned}$$

5: **while** $C^{(1,t-1)}$, $C^{(2,t-1)}$, $C^{(1,t)}$, $C^{(2,t)}$ are not equal **do**

6: (a) given $\Big(C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}\Big)$, update $C^{(1,t+1)}$ by

$$\Big[\sum_{m=1}^{m^*} \omega_m A^{(m)} C^{(2,t)} S^{(m,t)^T} + \frac{\omega_0}{2} X V^{(t)} + \lambda\Big]$$
$$\Big[\sum_{m=1}^{m^*} \omega_m \big[C^{(2,t)} S^{(m,t)^T}\big]^T \big[C^{(2,t)} S^{(m,t)^T}\big] + \frac{\omega_0}{2} V^{(t)^T} V^{(t)} + (\lambda+\nu) I_{k^*}\Big]^{-1}; \quad (6)$$

7: (b) given $\Big(C^{(1,t+1)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}\Big)$, update $C^{(2,t+1)}$ by

$$\Big[\sum_{m=1}^{m^*} \omega_m A^{(m)^T} C^{(1,t+1)} S^{(m,t)} + \frac{\omega_0}{2} X V^{(t)} + \lambda\Big]$$
$$\Big[\sum_{m=1}^{m^*} \omega_m \big[C^{(1,t+1)} S^{(m,t)}\big]^T \big[C^{(1,t+1)} S^{(m,t)}\big] + \frac{\omega_0}{2} V^{(t)^T} V^{(t)} + (\lambda+\nu) I_{k^*}\Big]^{-1}; \quad (7)$$

8: (c) given $\Big(C^{(1,t+1)}, C^{(2,t+1)}, V^{(t)}\Big)$, update $S^{(m,k+1)}$ for each $m \in \{1, \ldots, m^*\}$ as follows,

$$vec(S^{(m,k+1)}) = \big(\omega_m B^{(2,k+1)^T} \otimes B^{(1,k+1)} + (\alpha+\nu) I_{k^*}\big)^{-1} vec(U^{(m,k+1)}), \quad (8)$$

where $B^{(1,k+1)} = C^{(1,t+1)^T} C^{(1,t+1)}$, $B^{(2,k+1)} = C^{(2,t+1)^T} C^{(2,t+1)}$, $U^{(m,k+1)} = \omega_m C^{(1,t+1)^T} A^{(m)} C^{(2,t+1)} - \alpha S^{(m,t)}$, and $vec(\cdot)$ denotes the vectorization of a matrix by stacking its columns and $\otimes$ is the Kronecker product of two matrices;

9: (d) given $\Big(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,k+1)}\}_{m=1}^{m^*}\Big)$, update $V^{(t+1)}$ by

$$\big[X^T (C^{(1,t+1)} + C^{(2,t+1)}) + 2\beta/\omega_0 V^{(t)}\big] \big[C^{(1,t+1)^T} C^{(1,t+1)} + C^{(2,t+1)^T} C^{(2,t+1)} + 2(\beta+\nu)/\omega_0\big]^{-1}; \quad (9)$$

10: (e) $t = t + 1$.

11: **end while**

12: **return** the community label vector L by applying k-means to cluster the rows of $C^{(1,t)}$.

Here, α, β, and ν are set to be three small positive numbers, used to ensure convergence of the algorithm. Note that in the update step of the above algorithm, all the update formulas have explicit expressions. Specifically, given $\Big(C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}\Big)$, we update $C^{(1,t+1)}$ by

$$C^{(1,t+1)} = \underset{C^{(1)} \in \mathbb{R}^{n \times k^*}}{\arg\min} \Big\{ \sum_{m=1}^{m^*} \omega_m \|A^{(m)} - C^{(1)} S^{(m,t)} C^{(2,t)}\|_F^2 + \frac{\omega_0}{2} \|X - C^{(1)} V^{(t)^T}\|_F^2 + \lambda \|C^{(1)} - C^{(2,t)}\|_F^2 + \nu \|C^{(1)}\|_F^2 \Big\},$$

which has the explicit expression in in (6)

$$\Big[\sum_{m=1}^{m^*} \omega_m A^{(m)} C^{(2,t)} S^{(m,t)^T} + \frac{\omega_0}{2} X V^{(t)} + \lambda\Big]$$
$$\Big[\sum_{m=1}^{m^*} \omega_m \big[C^{(2,t)} S^{(m,t)^T}\big]^T \big[C^{(2,t)} S^{(m,t)^T}\big] + \frac{\omega_0}{2} V^{(t)^T} V^{(t)} + (\lambda+\nu) I_{k^*}\Big]^{-1}.$$

Similarly, we update $C^{(1,t+1)}$ in (7). Then, in (8), given $\left(C^{(1,t+1)}, C^{(2,t+1)}, V^{(t)}\right)$, we update $S^{(m,t+1)}$ by

$$S^{(m,t+1)} = \underset{S^{(m)} \in \mathbb{R}^{k^* \times k^*}}{\arg\min} \ \omega_m \|A^{(m)} - C^{(1,t+1)} S^{(m)} C^{(2,t+1)}\|_F^2 + \alpha \|S^{(m)} - S^{(m,t)}\|_F^2 + \nu \|S^{(m)}\|_F^2.$$

Finally, given $\left(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}\right)$, we update $V^{(t+1)}$ by

$$V^{(t+1)} = \underset{V \in \mathbb{R}^{p \times k^*}}{\arg\min} \ \frac{\omega_0}{2} \sum_{l=1}^{2} \|X - C^{(l,t+1)} V^T\|_F^2 + \beta \|V - V^{(t)}\|_F^2 + \nu \|V\|_F^2,$$

which has the explicit expression in (9)

$$\left[X^T \left(C^{(1,t+1)} + C^{(2,t+1)}\right) + 2\beta/\omega_0 V^{(t)}\right] \left[{C^{(1,t+1)}}^T C^{(1,t+1)} + {C^{(2,t+1)}}^T C^{(2,t+1)} + 2(\beta+\nu)/\omega_0\right]^{-1}.$$

4. Theoretical Analysis

Next, we consider the convergence theory of the PAF algorithm. We will present that the iteration sequence $\{\Theta^{(t)} = (C^{(1,t)}, C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^t)\}_{t=1}^{\infty}$ generated by the PAF algorithm converges to a critical point of (5).

Proposition 1. *There exist a constant $\delta > 0$ such that for each $t \in \{1, 2, \dots\}$, $\|\Theta^{(t)}\|_F \leq \delta$.*

Proof. Please see Appendix A.1. □

Proposition 2. *For each $t \in \{1, 2, \dots\}$,*

$$\rho_1 \|\Theta^{(t+1)} - \Theta^{(t)}\|_F^2 \leq [\mathcal{H}(\Theta^{(t)}) - \mathcal{H}(\Theta^{(t+1)})], \tag{10}$$

where $\rho_1 = \min\{2(\lambda+\nu)^2, \alpha, \beta\}$.

Proof. Please see Appendix A.2. □

Proposition 3. *For each $t \in \{1, 2, \dots\}$,*

$$\forall\, \kappa^{(t+1)} \in \partial H(\Theta^{(t+1)}),\ \|\kappa^{(t+1)}\|_F \leq \rho_2 \|\Theta^{(t+1)} - \Theta^{(t)}\|_F, \tag{11}$$

where $\rho_2 = \max\{4\delta(2\delta^3 + \tau) + 2\alpha, 2(\delta^4 + \tau\delta M + \lambda), 2(2\delta^2 + \|X\|_F + \beta)\}$.

Proof. Please see Appendix A.3. □

Theorem 1. *$\{\Theta^{(t)}\}$ converges to a critical point of $\mathcal{H}(\Theta)$.*

Proof. Please see Appendix A.4. □

5. Numerical Study

We now present the results of some numerical study to demonstrate the performance of the PAF algorithm, and the comparison with some existing methods, abbreviated as SCP, ANMF, and NMF, respectively. SCP is the spectral clustering with perturbations [33]. ANMF and NMF are the non-negative matrix factorization methods proposed in [28] for directed and undirected networks respectively. All the network data are generated from SBM or multi-layer SBM, and the attribution

data are generated from multivariate normal distributions, where the distribution parameters will be specified in each following setting. We will use the normalized mutual information (NMI) to measure the consistency between the predicted labels and the true community labels.

First, we consider the following two simulation settings for single networks and attributed networks:

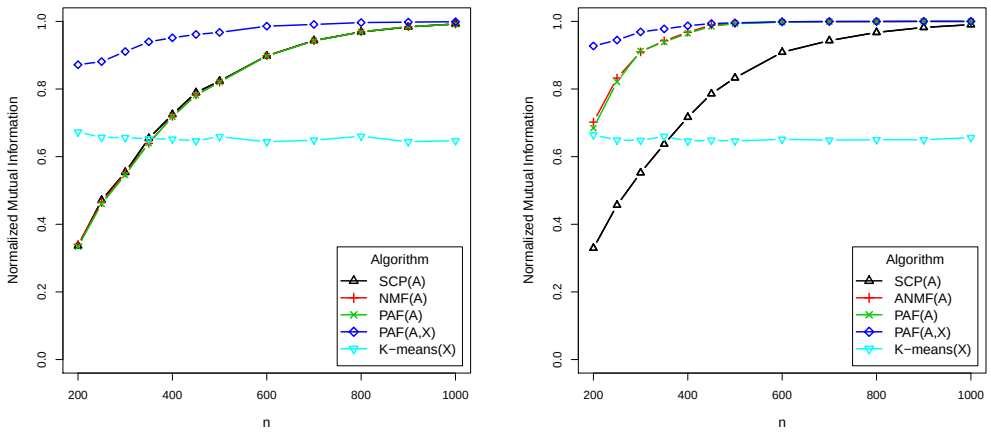
I The $n \times n$ adjacency matrix A is generated from the undirected SBM with the parameters

$$P = \begin{pmatrix} 0.20 & 0.12 \\ 0.12 & 0.20 \end{pmatrix}, \ \pi = \begin{pmatrix} 0.50 \\ 0.50 \end{pmatrix}. \tag{12}$$

Each row of the $n \times 2$ attribution matrix is independently generated from the multivariate normal distribution $N_2(\mu_k, \sigma^2 I_2)$, where the kth element of μ_k is 1 and the remaining element is 0, and $\sigma^2 = 0.15$.

II The same as Setting I, except that the undirected SBM is replaced by directed SBM.

The simulation results for Settings I and II are summarized in Figure 1, where SCP(A), NMF(A), ANMF(A), and PAF(A) denote applying SCP, NMF, ANMF, and PAF to A, respectively, k-means(X) denotes applying k-means to X and PAF(A, X) denotes applying PAF to (A, X). The results of SCP(A), NMF(A), ANMF(A), and PAF(A) in Figure 1 suggest that (1) PAF is a very good alternative to NMF and ANMF in terms of accuracy of community detection, and (2) NMF, ANMF, and PAF outperform SCP in situation where directed networks are studied.



(**a**) NMI, Setting I (**b**) NMI, Setting II

Figure 1. The two panels present the NMI results for Setting I (undirected SBM) and Setting II (directed SBM), respectively.

On the other hand, the comparison between PAF(A, X) and the other methods in Figure 1 suggests that applying k-means to the attribution data alone fails to achieve community detection; however, once the attribution data and the network data are combined, much better results can be obtained than using the network and attribution data separately.

Next, we consider the following two simulation settings for multi-layer networks and multi-layer attributed networks:

III The $m^* = 3$ $n \times n$ adjacent matrices $\{A^{(1)}, A^{(2)}, A^{(3)}\}$ are generated independently from the undirected multi-layer SBM with common community labels, where the parameters are set as follows,

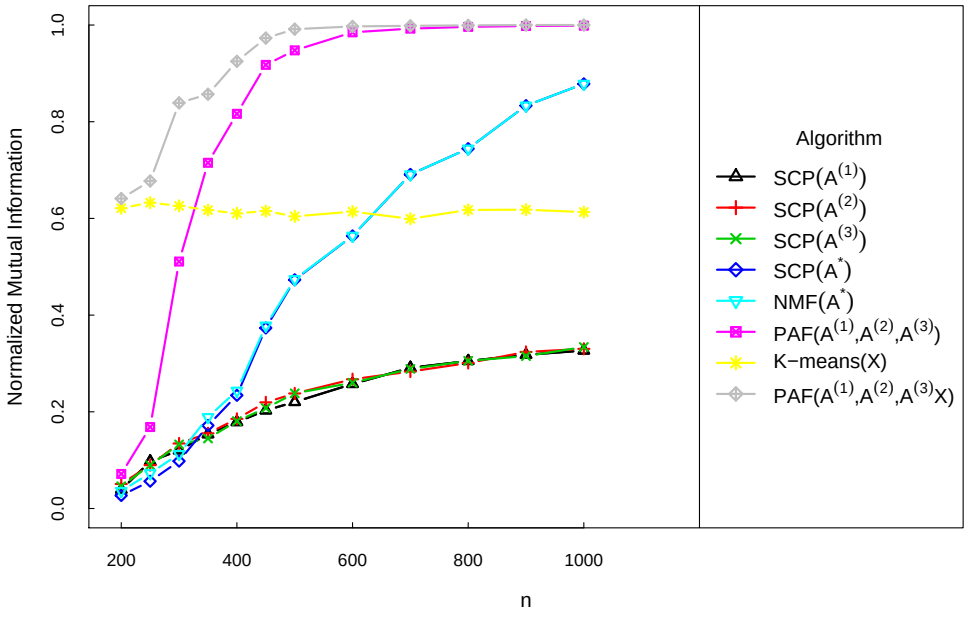
$$P_1 = \begin{pmatrix} 0.2 & 0.2 & 0.13 \\ 0.2 & 0.2 & 0.13 \\ 0.13 & 0.13 & 0.2 \end{pmatrix}, P_2 = \begin{pmatrix} 0.2 & 0.13 & 0.13 \\ 0.13 & 0.2 & 0.2 \\ 0.13 & 0.2 & 0.2 \end{pmatrix}, \quad (13)$$

$$P_3 = \begin{pmatrix} 0.2 & 0.13 & 0.2 \\ 0.13 & 0.2 & 0.13 \\ 0.2 & 0.13 & 0.2 \end{pmatrix}, \pi = \begin{pmatrix} 0.3\dot{3} \\ 0.3\dot{3} \\ 0.3\dot{3} \end{pmatrix}. \quad (14)$$

Each row of the $n \times 3$ attribution matrix is independently generated from the multivariate normal distribution $N_3(\mu_k, \sigma^2 I_3)$, where the kth element of μ_k is 1 and the remaining elements are 0, and $\sigma^2 = 0.15$.

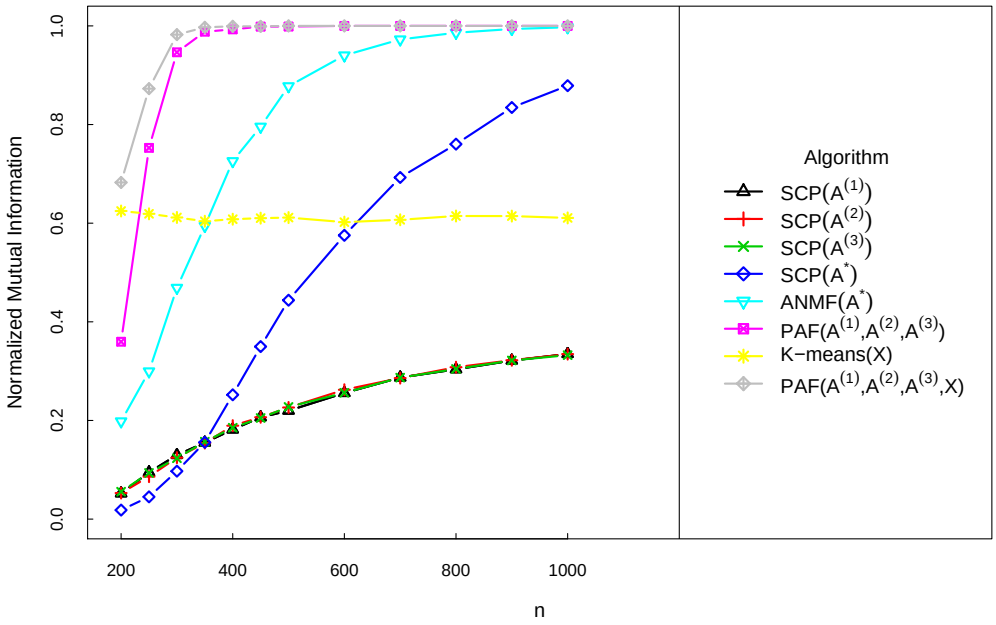
IV The same as Setting III, except that the undirected multi-layer SBM model is replaced by directed multi-layer SBM.

The simulation results for Settings III and IV are summarized in Figure 2, where PAF$(A^{(1)}, A^{(2)}, A^{(3)}, X)$ denotes applying PAF to $(A^{(1)}, A^{(2)}, A^{(3)}, X)$ and $A^* = \frac{1}{m^*} \sum_{m=1}^{m^*} A^{(m)}$. The comparison between PAF$(A^{(1)}, A^{(2)}, A^{(3)}, X)$ and NMF(A^*), ANMF(A^*), SCP(A^*), SCP$(A^{(1)})$, SCP$(A^{(2)})$, and SCP$(A^{(3)})$ suggests that (1) integrating community information from the multiple adjacent matrices of the network layers may perform better than using each network layer separately, and (2) using the PAF algorithm to achieve integrative community detection for the multi-layer attributed network can make appropriate use of the network-related data from multiple sources.



(a) NMI, Setting III

Figure 2. *Cont.*



(**b**) NMI, Setting IV

Figure 2. The two panels present the NMI results for Setting III (undirected multi-layer SBM) and Setting IV (directed multi-layer SBM), respectively.

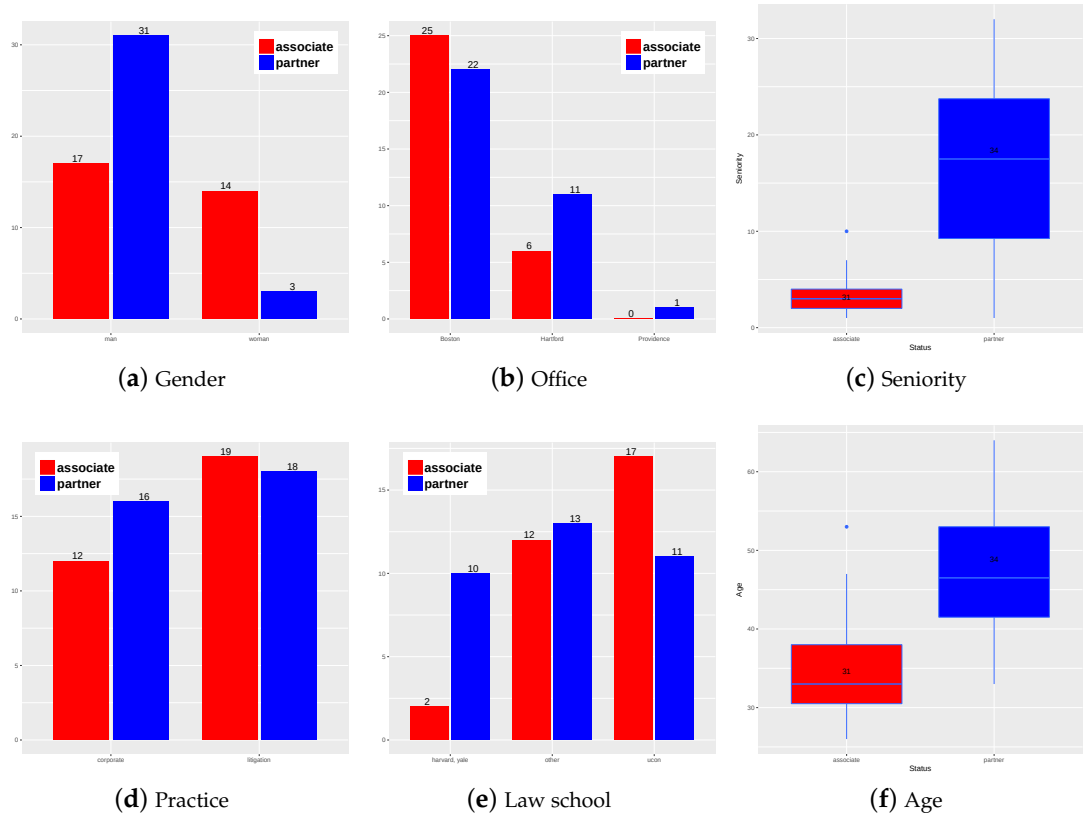
6. Empirical Analysis

In this section, we apply the proposed PAF method to a dataset that comes from a network study of a corporate law partnership, which was carried out in a Northeastern US corporate law firm, referred to as SG&R, 1988–1991 in New England and previously studied in [45,56,57]. The dataset includes 71 attorneys of this firm and three network layers, co-work layer, advice layer, friendship layer, as well as some attributes of the attorneys, such as status (1 = partner; = associate), gender (1 = man; 2 = woman), office (1 = Boston; 2 = Hartford; 3 = Providence), years with the firm, age, practice (1 = litigation; 2 = corporate), and law school (1: Harvard, Yale; 2: UCON; 3: other). We treat the attribute "status" as the ground truth community label as in [45]. In fact, after eliminating six isolated nodes, the heatmap plots of the adjacency matrices with nodes sorted by each attribute variable indicate that the partition by "status" can present a strong assortative structure. Then, the data of the remaining six attributes together with the three network layers form a multi-layer attributed network to be studied, with $m^* = 3$ network layers and $p = 6$ attribute variables, which falls right into the scope of application of the proposed method.

Intuitively, all these three network layers and six attributes can contribute to the community detection task with the ground truth label "status". Specifically, the descriptive analysis results in Figure 3 present that all these six attributes can provide useful information to distinguish the two values of "status"; the top three panels of Figure 4, i.e., the heatmap plots of the three adjacent matrices partitioned by the ground truth labels, partly present block structure according to the two values of "status".

The authors of [45] offered a comparison of seven methods for community detection of this dataset, we recall the NMI results of these methods in [45] by Table 1, together with the NMI result obtained by applying the proposed PAF method to the multi-layer attributed network with $m^* = 3$ network layers and $p = 6$ attribute variables. Table 1 indicates that the NMI performance of the PAF method is almost the same as the best existing one. Intuitively, the heatmap plots of the three adjacent matrices partitioned by the predicted labels of the PAF method are given in the bottom three panels of Figure 4, which are quite similar to those partitioned by the ground truth labels. Viewed from another perspective, we present the plots of the three network layers, colored by the ground truth labels and the predicted labels by PAF, respectively, in Figure 5, which indicate that the partition by both the ground

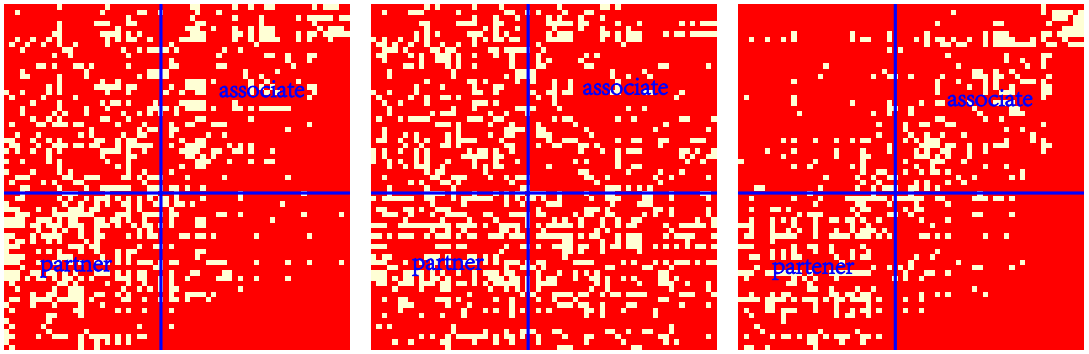
truth labels and the predicted labels by the proposed PAF method can present a strong assortative structure for the three network layers, especially for the friendship layer. These results demonstrate the practicability of the PAF method in community detection of multi-layer attributed networks.



(**a**) Gender (**b**) Office (**c**) Seniority

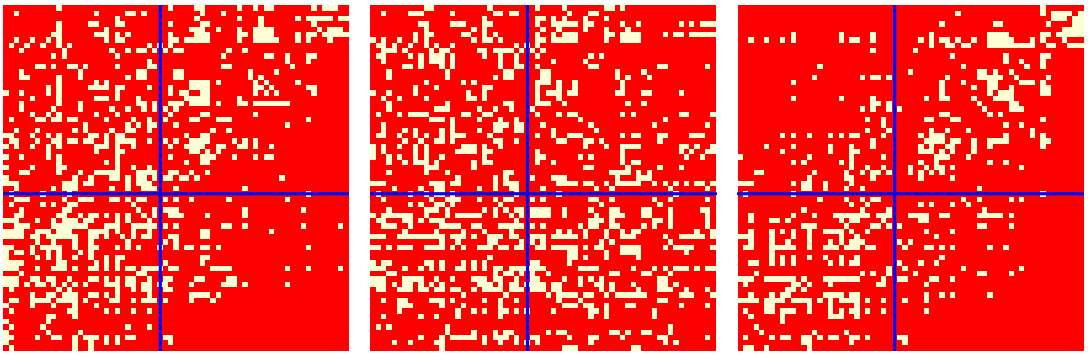
(**d**) Practice (**e**) Law school (**f**) Age

Figure 3. The left four panels are the grouped bar charts of status versus the four categorical features: "gender", "office", "practice", and "law school". The right two panels are the box-plots of "status" versus the two count variables "seniority" and "age".



(**a**) Advice network, partitioned by the ground truth labels (**b**) Cowork network, partitioned by the ground truth labels (**c**) Friendship network, partitioned by the ground truth labels

Figure 4. *Cont.*

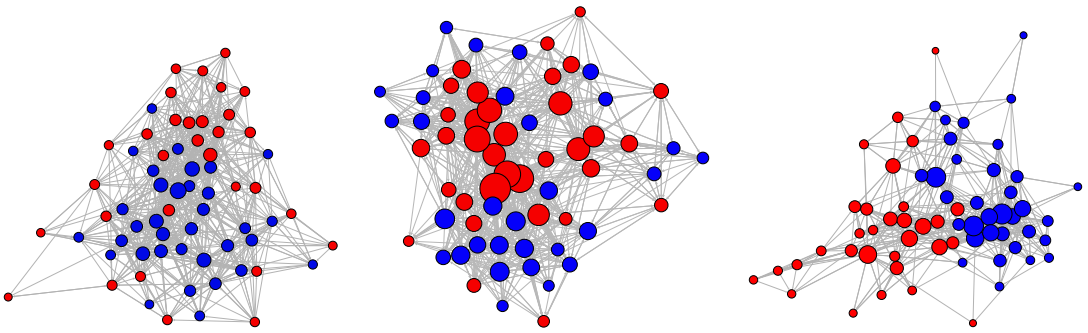


(**d**) Advice network, partitioned by the predicted labels by PAF

(**e**) Cowork network, partitioned by the predicted labels by PAF

(**f**) Friendship network, partitioned by the predicted labels by PAF

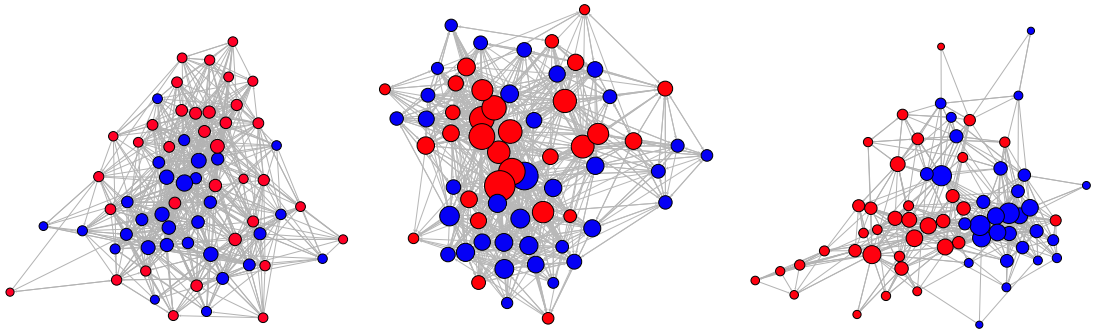
Figure 4. Heatmap plots of the adjacent matrices of the three network layers, ordered by the ground truth labels and the predicted labels by PAF, respectively.



(**a**) Advice network, colored according to the ground truth labels

(**b**) Cowork network, colored according to the ground truth labels

(**c**) Friendship network, colored according to the ground truth labels



(**d**) Advice network, colored according to the predicted labels by PAF

(**e**) Cowork network, colored according to the predicted labels by PAF

(**f**) Friendship network, colored according to the predicted labels by PAF

Figure 5. Plots of the three network layers, colored by the ground truth labels and the predicted labels by PAF, respectively.

Table 1. The NMI results of eight community detection methods for the multi-layer attributed network with $m^* = 3$ network layers and $p = 6$ attribute variables.

Method	NMI
JCDC, $\omega_n = 5$	0.54
JCDC, $\omega_n = 1.5$	0.50
SCP	0.44
k-means	0.44
CASC	0.49
CESNA	0.07
BAGC	0.20
PAF	0.58

7. Conclusions

We have proposed PAF—a unified framework and algorithm that is applicable to community detection of multi-layer attributed networks—as well as its special cases, such as single networks, attributed networks, and multi-layer networks. The main idea of PAF is replacing the community label matrix at two different positions in the original objective function with two different substitution matrices, penalizing the gap between the two substitution matrices, and then alternately optimizing each of the substitution matrices as well as some other variable matrices. The results of the simulation study and empirical analysis demonstrate the advantages of the PAF algorithm in community discovery accuracy and compatibility with multiple types of network-related data.

In our future work, we will study community detection of multi-layer attributed networks in statistical ways, where likelihood functions under some statistical models of multi-layer attributed networks will be considered.

Author Contributions: Conceptualization, B.L.; methodology, B.L.; software, J.W.; validation, J.W.; formal analysis, J.L.; investigation, J.W.; resources, J.L.; data curation, J.W.; writing—original draft preparation, J.W.; writing—review and editing, B.L.; visualization, J.W.; supervision, J.L.; project administration, B.L.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge Special Fund for Key Laboratories of Jilin Province (20190201285JC), Jilin Provincial Department of Education (JJKH20190293KJ) and Jilin Provincial Science and Technology Development Plan funded Project (20180520026JH).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANMF	asymmetric non-negative matrix factorization
KL	Kurdyka-Łojasiewicz
NMF	non-negative matrix factorization
NMI	normalized mutual information
PAF	penalized alternative factorization
SBM	stochastic block model
SCP	spectral clustering with perturbations

Appendix A. Proof of Some Theoretical Results

Appendix A.1. Proof of Proposition 1

Proof. As $\mathcal{H}(\Theta) \geq 0$ and $\{\mathcal{H}(\Theta^{(t)})\}_{t=1}^{\infty}$ is a monotone decreasing sequence, there exists a constant $\delta > 0$ such that $\|\Theta^{(t)}\|_F \leq \delta$ for each $t \in \{1, 2, \dots\}$. □

Appendix A.2. Proof of Proposition 2

Proof. Let $x^{(t)} = vec(C^{(1,t)} - C^{(1,t+1)})$. According to Step 2(a) in the PAF algorithm, we obtain

$$\begin{aligned}
&\mathcal{H}(C^{(1,t)}, C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) - \mathcal{H}(C^{(1,t+1)}, C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) \\
=&\langle \partial_{C^{(1)}} \mathcal{H}(C^{(1,t+1)}, C^{(2,t)}, S^k, V^{(t)}), C^{(1,t)} - C^{(1,t+1)} \rangle \\
&+ 2 \left\| x^{(t)^T} \left(\sum_{m=1}^{m^*} \omega_m (S^{(m,t)} C^{(2,t)T})^T S^{(m,t)} C^{(2,t)T} \otimes I_{k^* \times k^*} + (\lambda+\nu) I_{nk^* \times nk^*} \right) x^{(t)} \right\|_F^2 \qquad \text{(A1)} \\
\geq& 2(\lambda+\nu)^2 \|x^{(t)}\|_F^2 = 2(\lambda+\nu)^2 \|C^{(1,t)} - C^{(1,t+1)}\|_F^2.
\end{aligned}$$

Similarly, let $y^{(t)} = vec(C^{(2,t+1)} - C^{(2,t)})$, then according Step 2(b) in the PAF algorithm, we have

$$\begin{aligned}
&\mathcal{H}(C^{(1,t+1)}, C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) - \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) \\
=&\langle \partial_{C^{(2)}} \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}), C^{(2,t)} - C^{(2,t+1)} \rangle \\
&+ 2 \left\| y^{(t)^T} \left(\sum_{m=1}^{m^*} \omega_m (C^{(1,t+1)} S^{(m,t)})^T C^{(1,t+1)} S^{(m,t)} \otimes I_{k^* \times k^*} + (\lambda+\nu) I_{nk^* \times nk^*} \right) y^{(t)} \right\|_F^2 \qquad \text{(A2)} \\
\geq& 2(\lambda+\nu)^2 \|y^{(t)}\|_F^2 = 2(\lambda+\nu)^2 \|C^{(2,t)} - C^{(2,t+1)}\|_F^2.
\end{aligned}$$

From Steps 2(c,d) in the PAF algorithm, we obtain

$$\begin{aligned}
&\mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) - \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t)}) \\
\geq& \alpha \sum_{m=1}^{m^*} \|S^{(m,t+1)} - S^{(m,t)}\|_F^2, \qquad \text{(A3)} \\
&\mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t)}) - \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t+1)}) \\
\geq& \beta \|V^{(t+1)} - V^{(t)}\|_F^2. \qquad \text{(A4)}
\end{aligned}$$

Combining the inequalities (A1)–(A4), we have

$$\begin{aligned}
&\mathcal{H}(\Theta^{(t)}) - \mathcal{H}(\Theta^{(t+1)}) \\
\geq& 2(\lambda+\nu)^2 (\|C^{(1,t+1)} - C^{(1,t)}\|_F^2 + \|C^{(2,t+1)} - C^{(2,t)}\|_F^2) \\
&+ \alpha \sum_{m=1}^{m^*} \|S^{(m,t+1)} - S^{(m,t)}\|_F^2 + \beta \|V^{(t+1)} - V^{(t)}\|_F^2,
\end{aligned}$$

which implies that

$$\rho_1 \|\Theta^{(t)} - \Theta^{(t+1)}\|_F^2 \leq \mathcal{H}(\Theta^{(t)}) - \mathcal{H}(\Theta^{(t+1)}), \qquad \text{(A5)}$$

where $\rho_1 = \min\{2(\lambda+\nu)^2, \alpha, \beta\}$. □

Appendix A.3. Proof of Proposition 3

Proof. We will first analyze the boundness of $\partial_{C^{(1)}} \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t+1)})$. It is easy to check that

$$\begin{aligned}
&\partial_{C^{(1)}} \mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t+1)}) \\
=& 2 \sum_{m=1}^{m^*} \omega_m (C^{(1,t+1)} S^{(m,t+1)} C^{(2,t+1)T} - A^{(m)}) (S^{(m,t+1)} C^{(2,t+1)T})^T \qquad \text{(A6)} \\
&+ 2\lambda (C^{(1,t+1)} - C^{(2,t+1)}) + \omega_0 (C^{(1,t+1)} V^{(t+1)^T} - X) V^{(t+1)} + 2\nu C^{(1,t+1)}.
\end{aligned}$$

According to Step 2(a) in the PAF algorithm, we know

$$\begin{aligned}
&\partial_{C^{(1)}}\mathcal{H}(C^{(1,t+1)}, C^{(2,t)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) \\
&= 2\sum_{m=1}^{m^*}\omega_m(C^{(1,t+1)}S^{(m,t)}C^{(2,t)T} - A^{(m)})(S^{(m,t)}C^{(2,t)T})^T \\
&\quad + 2\lambda(C^{(1,t+1)} - C^{(2,t)}) + \omega_0(C^{(1,t+1)}{V^{(t)}}^T - X)V^{(t)} + 2\nu C^{(1,t+1)} = 0.
\end{aligned} \tag{A7}$$

Together with (A6) and (A7), we have

$$\begin{aligned}
&\|\partial_{C^{(1)}}\mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t+1)})\|_F \\
&\le 2\sum_{m=1}^{m^*}\omega_m\|C^{(1,t+1)}S^{(m,t+1)}C^{(2,t+1)T})(S^{(m,t+1)}C^{(2,t)T})^T - C^{(1,t+1)}S^{(m,t)}C^{(2,t)T}(S^{(m,t)}C^{(2,t)T})^T\|_F \\
&\quad + 2\sum_{m=1}^{m^*}\omega_m\|A^{(m)}((S^{(m,t+1)}C^{(2,t+1)T})^T - (S^{(m,t)}C^{(2,t)T})^T)\|^F + 2\lambda\|C^{(2,t+1)} - C^{(2,t)}\|_F \\
&\quad + \omega_0\|C^{(1,t+1)}({V^{(t+1)}}^T V^{(t+1)} - {V^{(t)}}^T V^{(t)})\|_F + 2\omega_0\|X(V^{(t+1)} - V^{(t)})\|_F \\
&\le 2\|C^{(1,t+1)}\|_F\sum_{m=1}^{m^*}\omega_m\Big(\|(S^{(m,t+1)} - S^{(m,t)})C^{(2,t+1)T}(S^{(m,t+1)}C^{(2,t+1)T})^T\|_F \\
&\quad + \|S^{(m,t)}(C^{(2,t+1)} - C^{(2,t)})^T(S^{(m,t+1)}C^{(2,t+1)T})^T\|_F + \|S^{(m,t)}C^{(2,t)T}C^{(2,t+1)}(S^{(m,t+1)} - S^{(m,t)})^T\|_F \\
&\quad + \|S^{(m,t)}C^{(2,t)T}(C^{(2,t+1)} - C^{(2,t)}){S^{(m,t)}}^T\|_F\Big) + 2\sum_{m=1}^{m^*}\omega_m\Big[\|A^{(m)}\|_F\|(C^{(2,t+1)} - C^{(2,t)}){S^{(m,t+1)}}^T\|_F \\
&\quad + \|A^{(m)}\|_F\|C^{(2,t)}(S^{(m,t+1)} - S^{(m,t)})^T\|_F\Big] + 2\lambda\|C^{(2,t+1)} - C^{(2,t)}\|_F \\
&\quad + \omega_0\|C^{(1,t+1)}({V^{(t+1)}}^T V^{(t+1)} - {V^{(t)}}^T V^{(t)})\|_F + \omega_0\|X(V^{(t+1)} - V^{(t)})\|_F \\
&\le 2\delta(2\delta^3 + \tau)\sum_{m=1}^{m^*}\|S^{(m,t+1)} - S^{(m,t)}\|_F + (2\delta^4 + 2\tau\delta m^* + 2\lambda)\|C^{(2,t+1)} - C^{(2,t)}\|_F \\
&\quad + (2\delta^2 + \|X\|_F)\|V^{(t+1)} - V^{(t)}\|_F,
\end{aligned} \tag{A8}$$

where $\tau = \max\{\|A^{(1)}\|_F, \|A^{(2)}\|_F, \ldots, \|A^{(m^*)}\|_F\}$.

Next, we see that

$$\begin{aligned}
&\partial_{C^{(2)}}\mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t+1)}\}_{m=1}^{m^*}, V^{(t+1)}) \\
&= 2\sum_{m=1}^{m^*}\omega_m(C^{(1,t+1)}S^{(m,t+1)})^T(C^{(1,t+1)}S^{(m,t+1)}C^{(2,t+1)T} - A) + 2\lambda(C^{(2,t+1)} - C^{(1,t+1)}) \\
&\quad + \omega_0(C^{(2,t+1)}{V^{(t+1)}}^T - X)V^{(t+1)} + 2\nu C^{(2,t+1)},
\end{aligned} \tag{A9}$$

and from Step 2(b) in the PAF algorithm,

$$\begin{aligned}
&\partial_{C^{(2)}}\mathcal{H}(C^{(1,t+1)}, C^{(2,t+1)}, \{S^{(m,t)}\}_{m=1}^{m^*}, V^{(t)}) \\
&= 2\sum_{m=1}^{m^*}\omega_m(C^{(1,t+1)}S^{(m,t)})^T(C^{(1,t+1)}S^{(m,t)}C^{(2,t+1)T} - A) + 2\lambda(C^{(2,t+1)} - C^{(1,t+1)}) \\
&\quad + \omega_0(C^{(2,t+1)}{V^{(t)}}^T - X)V^{(t)} + 2\nu C^{(2,t+1)} = 0.
\end{aligned} \tag{A10}$$

As a result,

$$\begin{aligned}
&\|\partial_{C^{(2)}}\mathcal{H}(C^{(1,t+1)},C^{(2,t+1)},\{S^{(m,t+1)}\}_{m=1}^{m^*},V^{(t+1)})\|_F\\
&\le 2\|C^{(2,t+1)}\|_F\sum_{m=1}^{m^*}\omega_m\|(C^{(1,t+1)}S^{(m,t+1)})^TC^{(1,t+1)}S^{(m,t+1)}-(C^{(1,t+1)}S^{(m,t)})^TC^{(1,t+1)}S^{(m,t)}\|_F\\
&\quad+2\sum_{m=1}^{m^*}\omega_m\|(S^{(m,t+1)}C^{(1,t+1)}-S^{(m,t)}C^{(1,t+1)})A^{(m)^T}\|_F+\omega_0\|X\|_F\|V^{(t+1)}-V^{(t)}\|_F\\
&\quad+\omega_0\|C^{(2,t+1)}\|_F\|V^{(t+1)}V^{(t+1)^T}-V^{(t)}V^{(t)^T}\|_F\\
&\le 2\sum_{m=1}^{m^*}\omega_m\big(\|C^{(2,t+1)}\|_F\|S^{(m,t+1)}C^{(1,t+1)^T}C^{(1,t+1)}\|_F\|S^{(m,t+1)}-S^{(m,t)}\|_F\\
&\quad+\|S^{(m,t+1)}-S^{(m,t)}\|_F\|C^{(2,t+1)}\|_F\|C^{(1,t+1)^T}C^{(1,t+1)}S^{(m),k}\|_F\big)\\
&\quad+2\sum_{m=1}^{m^*}\omega_m\|(S^{(m,t+1)}-S^{(m,t)})\|_F\|C^{(1,t+1)}\|_F\|A^{(m)}\|_F\\
&\quad+\omega_0\|C^{(1,t+1)}\|_F\|V^{(t+1)}\|_F\|V^{(t+1)}-V^{(t)}\|_F+\omega_0\|C^{(2,t+1)}\|_F\|V^{(t)}\|_F\|V^{(t+1)}-V^{(t)}\|_F\\
&\quad+\omega_0\|X\|_F\|V^{(t+1)}-V^{(t)}\|_F\\
&\le 2\delta(2\delta^3+\tau)\sum_{m=1}^{m^*}\|S^{(m,t+1)}-S^{(m,t)}\|_F+(2\delta^2+\|X\|_F)\|V^{(t+1)}-V^{(t)}\|_F.
\end{aligned}\tag{A11}$$

Similarly, we have

$$\|\partial_{S^{(m)}}\mathcal{H}(C^{(1,t+1)},C^{(2,t+1)},\{S^{(m,t+1)}\}_{m=1}^{m^*},V^{(t+1)})\|_F\le 2\alpha\|(S^{(m,t+1)}-S^{(m,t)})\|_F\tag{A12}$$

and

$$\|\partial_V\mathcal{H}(C^{(1,t+1)},C^{(2,t+1)},\{S^{(m,t+1)}\}_{m=1}^{m^*},V^{(t+1)})\|_F\le 2\beta\|(V^{(t+1)}-V^{(t)})\|_F.\tag{A13}$$

According to (A8) and (A11)–(A13), we finally obtain that

$$\|\partial H(\Theta^{(t+1)})\|_F\le\rho_2\|\Theta^{(t+1)}-\Theta^{(t)}\|_F,\tag{A14}$$

where $\rho_2=\max\{4\delta(2\delta^3+\tau)+2\alpha,2(\delta^4+\tau\delta m^*+\lambda),2(2\delta^2+\|X\|_F+\beta)\}$.

□

Appendix A.4. Proof of Theorem 1

To establish the proof of Theorem 1, we need to recall the definition of Kurdyka-Łojasiewicz (KL) property and prove the following two properties.

1. Sufficient decrease property:

$$\exists\,\rho_1>0,\ \text{such that}\ \rho_1\|\Theta^{(t+1)}-\Theta^{(t)}\|_F^2\le[\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\Theta^{(t+1)})];\tag{A15}$$

2. A subgradient lower bound for the iteration gap:

$$\exists\,\rho_2>0,\ \forall t\in\{0,1,\dots\}\ \text{and}\ \forall\kappa^{(t+1)}\in\partial\mathcal{H}(\Theta^{(t+1)}):\ \|\kappa^{(t+1)}\|_F\le\rho_2\|\Theta^{(t+1)}-\Theta^{(t)}\|_F,\tag{A16}$$

where $\partial\mathcal{H}$ is the subderivative of the function $\mathcal{H}$ and

$$\begin{aligned}
\Theta &= (C^{(1)},C^{(2)},\{S^{(m)}\}_{m=1}^{m^*},V),\\
\mathcal{H}(\Theta) &= \sum_{m=1}^{m^*}\omega_m\|A^{(m)}-C^{(1)}S^{(m)}C^{(2)}\|_F^2+\frac{\omega_0}{2}\sum_{l=1}^{2}\|X-C^{(l)}V^T\|_F^2
\end{aligned}$$

$$+\lambda\|C^{(1)}-C^{(2)}\|_F^2+\nu\{\|C^{(1)}\|_F^2+\|C^{(2)}\|_F^2+\|V\|_F^2+\sum_{m=1}^{m^*}\|S^{(m)}\|_F^2\}, \quad \text{(A17)}$$

$$\partial\mathcal{H}(\Theta) = \big(\partial_{C^{(1)}}\mathcal{H}(\Theta),\partial_{C^{(2)}}\mathcal{H}(\Theta),\partial_{S^{(1)}}\mathcal{H}(\Theta),\ldots,\partial_{S^{(m^*)}}\mathcal{H}(\Theta),\partial_V\mathcal{H}(\Theta)\big).$$

Definition A1. (KL property [58,59]) *Let $\sigma:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper lower semicontinuous function. For $\bar{x}\in dom\,\partial\sigma\doteq\{x\in\mathbb{R}^d:\partial\sigma(x)\neq\emptyset\}$ if there exists an $\eta\in(0,+\infty]$, a neighborhood Γ of $\bar{x}$ and a function $\xi\in\Phi_\eta$, such that for all $x\in X\cap\{y\in\mathbb{R}^d:\sigma(\bar{x})<\sigma(y)<\sigma(\bar{x})+\eta\}$ the following inequality holds*

$$\xi'\big(\sigma(u)-\sigma(\bar{x})\big)\mathrm{dist}\big(0,\partial\sigma(x)\big)\geq 1,$$

then σ is said to have the KL property at $\bar{x}$. σ is called a KL function, if σ satisfies the KL property at each point in dom $\partial\sigma$.

Here, $\mathrm{dist}(x,\mathcal{X})$ denotes the shortest distance between the point x and the point set $\mathcal{X}$, i.e., $\mathrm{dist}(x,\mathcal{X})=\min_{y\in\mathcal{X}}\mathrm{dist}(x,y)$, and Φ_η denotes the class of all concave and continuous functions $\xi:[0,\eta)\to\mathbb{R}_+$, $\eta\in\mathbb{R}_+$, which satisfies: $a)$ $\xi(0)=0$; $b)$ ξ is continuous differentiable on $(0,\eta)$; $c)$ for all $s\in(0,\eta)$, $\xi'(s)>0$.

Now, we are ready to present the proof of Theorem 1, based on the fact that the function $\mathcal{H}$ is a KL function.

Proof. Suppose that $\bar{\Theta}$ is the limit point of a sub-sequence $\{\Theta^{(t_i)}\}$, i.e., $\lim_{i\to+\infty}\Theta^{(t_i)}=\bar{\Theta}$. It is clear that the function $\mathcal{H}$ is continuous w.r.t. Θ, therefore

$$\lim_{i\to+\infty}\mathcal{H}(\Theta^{(t_i)})=\mathcal{H}(\bar{\Theta}). \quad \text{(A18)}$$

Note that $\mathcal{H}(\Theta^{(t)})$ is monotonically non-increasing and then converges to $\mathcal{H}(\bar{\Theta})$.

From Propositions 2 and 3, we obtain that $\partial H(\Theta^{(t)})\to 0$ as $t\to\infty$. Then we obtain $0\in\partial H(\bar{\Theta})$, i.e., $\bar{\Theta}$ is a critical point of Φ. Let Ω be the set of all limit points of subsequences of $\{\Theta^{(t)}\}_{t=1}^{\infty}$, we then know that

$$\mathrm{dist}(\Theta^{(t)},\Omega)\to 0,\ \text{as}\ t\to\infty. \quad \text{(A19)}$$

Accordingly, from (A18) and (A19), we obtain that for any $\gamma>0,\epsilon>0$, there exists an integer $t_u>0$ such that for all $t>t_u$,

$$\mathcal{H}(\Theta^{(t)})<\mathcal{H}(\bar{\Theta})+\gamma\ \text{and}\ \mathrm{dist}(\Theta^{(t)},\Omega)<\epsilon. \quad \text{(A20)}$$

It is known that the function $\mathcal{H}$ is a KL function [59], then by using the KL inequality in Definition (A1) and (A20), there exist ξ, such that

$$\xi'(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\bar{\Theta}))\mathrm{dist}(0,\partial\mathcal{H}(\Theta^{(t)}))\geq 1. \quad \text{(A21)}$$

According to Proposition 3, we have

$$\xi'(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\bar{\Theta}))\geq\frac{1}{\rho_2\|\Theta^{(t)}-\Theta^{(t-1)}\|_F}. \quad \text{(A22)}$$

Besides, ξ is concave and we have that

$$\xi(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\bar{\Theta}))-\xi(\mathcal{H}(\Theta^{(t+1)})-\mathcal{H}(\bar{\Theta}))\geq\xi'(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\bar{\Theta}))(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\Theta^{(t+1)})). \quad \text{(A23)}$$

For convenience, let

$$\Delta_{t,t+1}=\xi(\mathcal{H}(\Theta^{(t)})-\mathcal{H}(\bar{\Theta}))-\xi(\mathcal{H}(\Theta^{(t+1)})-\mathcal{H}(\bar{\Theta}))$$

Then according to propositions 2 and 3, we have

$$\Delta_{t,t+1} \geq \frac{\rho_1 \|\Theta^{(t+1)} - \Theta^{(t)}\|_F^2}{\rho_2 \|\Theta^{(t)} - \Theta^{(t-1)}\|_F}$$

i.e.,

$$4\|\Theta^{(t+1)} - \Theta^{(t)}\|_F^2 \leq 4\frac{\rho_2}{\rho_1}\Delta_{t,t+1}\|\Theta^{(t)} - \Theta^{(t-1)}\|_F \leq (\frac{\rho_2}{\rho_1}\Delta_{t,t+1} + \|\Theta^{(t)} - \Theta^{(t-1)}\|_F)^2,$$

which indicates that

$$2\|\Theta^{(t+1)} - \Theta^{(t)}\|_F \leq \|\Theta^{(t)} - \Theta^{(t-1)}\|_F + \frac{\rho_2}{\rho_1}\Delta_{t,t+1}. \tag{A24}$$

Summing up (A24) over t from 1 to z and then yields the following inequality,

$$\begin{aligned} 2\sum_{t=1}^{z}\|\Theta^{(t+1)} - \Theta^{(t)}\|_F &\leq \sum_{t=1}^{z}\|\Theta^{(t)} - \Theta^{(t-1)}\|_F + \frac{\rho_2}{\rho_1}\sum_{t=1}^{z}\Delta_{t,t+1} \\ &\leq \sum_{t=1}^{z}\|\Theta^{(t+1)} - \Theta^{(t)}\|_F + \|\Theta^{(1)} - \Theta^{(0)}\|_F + \frac{\rho_2}{\rho_1}\Delta_{1,z+1} \end{aligned} \tag{A25}$$

i.e.,

$$\sum_{t=1}^{z}\|\Theta^{(t+1)} - \Theta^{(t)}\|_F < \|\Theta^{(1)} - \Theta^{(0)}\|_F + \frac{\rho_2}{\rho_1}\xi(\mathcal{H}(\Theta^{(1)}) - \mathcal{H}(\bar{\Theta})). \tag{A26}$$

We take limits on the left side of inequality of (A26) as $z \to \infty$ and get

$$\sum_{t=1}^{+\infty}\|\Theta^{(t)} - \Theta^{(t+1)}\|_F \leq \|\Theta^{(1)} - \Theta^{(0)}\|_F + \frac{\rho_2}{\rho_1}\xi(\mathcal{H}(\Theta^{(1)}) - \mathcal{H}(\bar{\Theta})) < +\infty. \tag{A27}$$

It is easy to know that (A27) implies that $\{\Theta^{(t)}\}_{t=1}^{\infty}$ is a Cauchy sequence, and thus is a convergent sequence and this completes the proof, i.e., the sequence $\{\Theta^{(t)}\}_{t=1}^{\infty}$ converges to a critical point of $\mathcal{H}$ in (A17). □

Appendix B. Some Additional Numerical Results

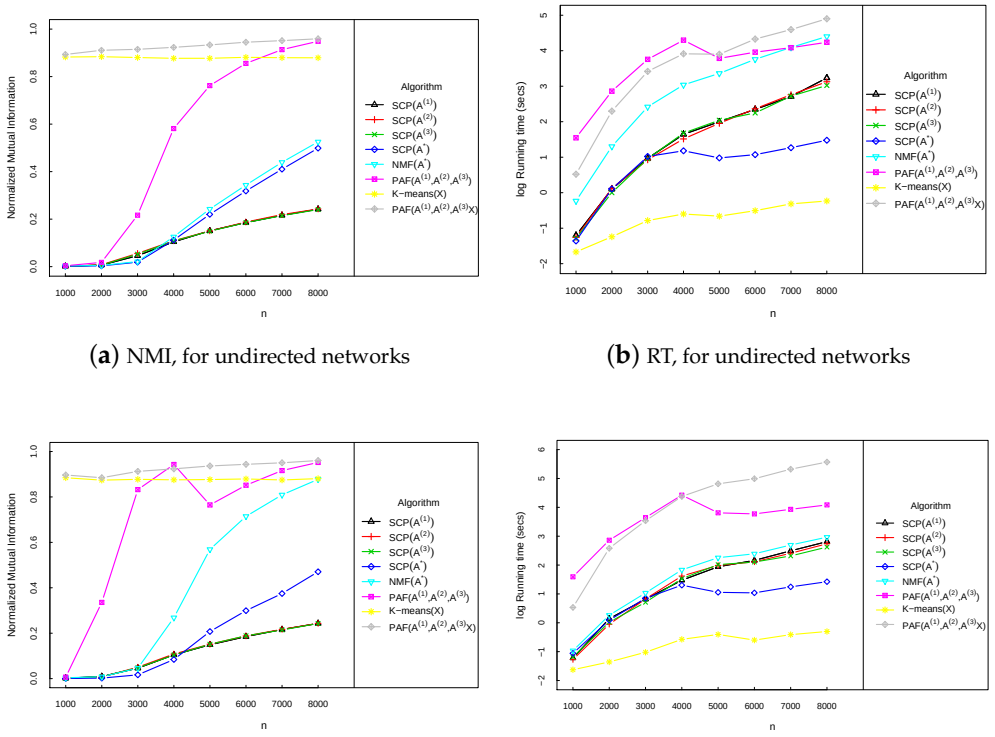
In this section, we mainly investigate the computational efficiency of the proposed algorithm for relatively large scale multi-layer attributed networks via some additional numerical results. We consider the following simulation setting for multi-layer networks and multi-layer attributed networks.

V The $m^* = 3$ $n \times n$ adjacent matrices $\{A^{(1)}, A^{(2)}, A^{(3)}\}$ are generated independently from the undirected and directed multi-layer SBMs, respectively, with common community labels, where the parameters are set as follows,

$$\begin{aligned} P_1 &= \begin{pmatrix} 0.2 & 0.2 & 0.13 \\ 0.2 & 0.2 & 0.13 \\ 0.13 & 0.13 & 0.2 \end{pmatrix}, \; P_2 = 0.8\begin{pmatrix} 0.2 & 0.13 & 0.13 \\ 0.13 & 0.2 & 0.2 \\ 0.13 & 0.2 & 0.2 \end{pmatrix}, \\ P_3 &= 0.8\begin{pmatrix} 0.2 & 0.13 & 0.2 \\ 0.13 & 0.2 & 0.13 \\ 0.2 & 0.13 & 0.2 \end{pmatrix}, \; \pi = \begin{pmatrix} 0.3\dot{3} \\ 0.3\dot{3} \\ 0.3\dot{3} \end{pmatrix}. \end{aligned}$$

Each row of the $n \times 3$ attribution matrix is independently generated from the multivariate normal distribution $N_3(\mu_k, \sigma^2 I_3)$, where the kth element of μ_k is 1 and the remaining elements are 0, and $\sigma^2 = 0.08$.

As suggested in Figure A1, the proposed algorithm framework is compatible with a variety of network related data, has relatively good accuracy of community discovery and acceptable computational efficiency. Compared with the algorithm SCP, which provides initial value for the proposed algorithm, the proposed algorithm does not excessively reduce the computational efficiency.



(**a**) NMI, for undirected networks

(**b**) RT, for undirected networks

(**c**) NMI, for directed networks

(**d**) RT, for directed networks

Figure A1. The left two panels present the NMI results for Setting V and the right two panels present the RT (running time in log-second) results for Setting V.

References

1. Newman, M.E.J. *Networks*; Oxford University Press: Oxford, UK, 2018.
2. Wasserman, S. *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*; Sage: Thousand Oaks, CA, USA, 1994.
3. Bader, G.D.; Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef]
4. Sporns, O. *Networks of the Brain*; MIT Press: Cambridge, MA, USA, 2010.
5. Rogers, E.M.; Kincaid, D.L. *Communication Networks: Toward a New Paradigm for Research*; Free Press: New York, NY, USA, 1981, Volume 11, p. 2.
6. Schlitt, T.; Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinform.* **2007**, *8*, S9. [CrossRef]
7. Mcpherson, M.; Smithlovin, L.; Cook, J.M. Birds of a feather: Homophily in social networks. *Rev. Sociol.* **2001**, *27*, 415–444. [CrossRef]
8. Moody, J.; White, D.R. Structural cohesion and embeddedness: A hierarchical concept of social groups. *Am. Sociol. Rev.* **2003**, 103–127. [CrossRef]
9. Flake, G.W.; Lawrence, S.; Giles, C.L. Efficient identification of web communities. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 20–23 August 2000; pp. 150–160.
10. Sporns, O.; Betzel, R.F. Modular brain networks. *Annu. Rev. Psychol.* **2016**, *67*, 613–640. [CrossRef] [PubMed]

11. Spirin, V.; Mirny L.A. Protein complexes and functional modules in molecular networks. *Proc. Natil. Acad. Sci. USA* **2003**, *100*, 12123–12128. [CrossRef]
12. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *10*, 75–174. [CrossRef]
13. Fortunato, S.; Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **2016**, *659*, 1–44. [CrossRef]
14. Khan, B.S.; Niazi, M.A. Network community detection: A review and visual survey. *arXiv* **2017**, arXiv:1708.00977.
15. Porter, M.A.; Onnela, J.-P.; Mucha P.J. Communities in networks. *Not. AMS* **2009**, *56*, 1082–1097.
16. Schaub, M.T.; Delvenne, J.-C.; Rosvall, M.; Lambiotte, R. The many facets of community detection in complex networks *Appl. Netw. Sci.* **2017**, *2*, 4. [CrossRef] [PubMed]
17. Newman, M.E.J. Detecting community structure in networks. *Eur. Phys. J. B* **2004**, *38*, 321–330. [CrossRef]
18. Hespanha, J.P. *An Efficient Matlab Algorithm for Graph Partitioning*; University of California: Santa Barbara, CA, USA, 2004.
19. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
20. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. [CrossRef] [PubMed]
21. Jin, J. Fast community detection by score. *Ann. Stat.* **2015**, *43*, 57–89. [CrossRef]
22. Lei, J.; Rinaldo, A. Consistency of spectral clustering in stochastic block models. *Ann. Stat.* **2015**, *43*, 215–237. [CrossRef]
23. McSherry, F. Spectral partitioning of random graphs. In Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, Newport Beach, CA, USA, 8–11 October 2001; pp. 529–537.
24. Rohe, K.; Chatterjee, S.; Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Stat.* **2011**, *39*, 1878–1915. [CrossRef]
25. Cai, T.T.; Li, X. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Stat.* **2015**, *43*, 1027–1059. [CrossRef]
26. Hajek, B.; Wu, Y.; Xu, J. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inf. Theory* **2016**, *62*, 5918–5937. [CrossRef]
27. Le, C.M.; Levina, E.; Vershynin, R. Optimization via low-rank approximation for community detection in networks. *Ann. Stat.* **2016**, *44*, 373–400. [CrossRef]
28. Wang, F.; Li, T.; Wang, X.; Zhu, S.; Ding, C. Community discovery using non-negative matrix factorization. *Data Min. Knowl. Discov.* **2011**, *22*, 493–521. [CrossRef]
29. Holland, P.W.; Laskey, K.B.; Leinhardt, S. Stochastic block models: First steps. *Soc. Netw.* **1983**, *5*, 109–137. [CrossRef]
30. Karrer, B.; Newman, M.E.J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **2011**, *83*, 016107. [CrossRef]
31. Hoff, P.D. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2008; pp. 657–664.
32. Newman, M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef] [PubMed]
33. Amini, A.A.; Chen, A.; Bickel, P.J.; Levina, E. Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Stat.* **2013**, *41*, 2097–2122. [CrossRef]
34. Qin, T.; Rohe, K. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2013; pp. 3120–3128.
35. Hoff, P.D. Random effects models for network data. In *Dynamic Social Network Modeling and Analysis Workshop Summary and Papers*; National Academies Press: Washington, DC, USA,2003.
36. Zanghi, H.; Volant, S.; Ambroise, C. Clustering based on random graph model embedding vertex features. *Pattern Recogn. Lett.* **2010**, *31*, 830–836. [CrossRef]
37. Handcock, M.S.; Raftery, A.E.; Tantrum, J.M. Model-based clustering for social networks. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2007**, *170*, 301–354. [CrossRef]

38. Yang, T.; Jin, R.; Chi, Y.; Zhu, S. Combining link and content for community detection: A discriminative approach. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 927–936.
39. Kim, M.; Leskovec, L.J. Latent multi-group membership graph model. *arXiv* **2012**, arXiv:1205.4546.
40. Leskovec, J.; Mcauley, J.J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2012; pp. 539–547.
41. Yang, J.; McAuley, J.; Leskovec, J. Community detection in networks with node attributes. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; pp. 1151–1156.
42. Xu, Z.; Ke, Y.; Wang, Y.; Cheng, H.; Cheng, J. A model-based approach to attributed graph clustering. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 505–516.
43. Hoang, T.-A.; Lim, E.-P. On joint modeling of topical communities and personal interest in microblogs. In *International Conference on Social Informatics*; Springer: Cham, Switzerland, 2014; pp. 1–16.
44. Newman, M.E.J.; Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **2016**, *7*, 11863. [CrossRef]
45. Zhang, Y.; Levina, E.; Zhu, J. Community detection in networks with node features. *Electron. J. Stat.* **2016**, *10*, 3153–3178. [CrossRef]
46. Boorman, S.A.; White, H.C. Social structure from multiple networks. ii. role structures. *Am. J. Sociol.* **1976**, *81*, 1384–1446. [CrossRef]
47. Breiger, R.L. Social structure from multiple networks. *Am. J. Sociol.* **1976**, *81*, 730–780.
48. Cheng, W.; Zhang, X.; Guo, Z.; Wu, Y.; Sullivan, P.F.; Wang, W. Flexible and robust co-regularized multi-domain graph clustering. *Knowl. Discov. Data Min.* **2013**, 320–328.
49. Boccaletti, S.; Bianconi, G.; Criado, R.; DelGenio, C.I.; Gómez-Gardenes, J.; Romance, M.; Sendina-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.* **2014**, *544*, 1–122. [CrossRef]
50. Kivelä; M; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. *J. Complex Netw.* **2014**, *2*, 203–271.
51. Matias, C.; Miele, V. Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2017**, *79*, 1119–1141. [CrossRef]
52. Cardillo, A.; Gómez-Gardenes, J.; Zanin, M.; Romance, M.; Papo, D.; DelPozo, F.; Boccaletti, S. Emergence of network features from multiplexity. *Sci. Rep.* **2013**, *3*, 1344. [CrossRef]
53. Fienberg, S.E.; Meyer, M.M.; Wasserman, S.S. *Analyzing Data from Multivariate Directed Graphs: An Application to Social Networks*; Technical Report; Department of Statistics, Carnegie Mellon University: Pittsburgh, PA, USA, 1980.
54. Fienberg, S.E.; Meyer, M.M.; Wasserman, S.S. Statistical analysis of multiple sociometric relations. *J. Am. Stat. Assoc.* **1985**, *80*, 51–67. [CrossRef]
55. Ferriani, S.; Fonti, F.; Corrado, R. The social and economic bases of network multiplexity: Exploring the emergence of multiplex ties. *Strateg. Organ.* **2013**, *11*, 7–34. [CrossRef]
56. Yan, T.; Jiang, B.; Fienberg, E.S.; Leng, C. Statistical inference in a directed network model with covariates. *J. Am. Stat. Assoc.* **2019**, *114*, 857–868. [CrossRef]
57. Lazega, E. *The Collegial Phenomenon: The Social Mechanisms of Cooperation among Peers in a Corporate Law Partnership*; Oxford University Press on Demand: Oxfor, UK, 2001.
58. Attouch, H.; Bolte, J.; Svaiter, B.F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Math. Programm.* **2013**, *137*, 91–129. [CrossRef]
59. Bolte, J.; Sabach, S.; Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Programm.* **2014**, *146*, 459–494. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).