Received December 6, 2019, accepted December 20, 2019, date of publication December 30, 2019, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962634

Semi-Supervised Community Detection via Constraint Matrix Construction and Active Node Selection

SUQI ZHANG[1], JUNYAN WU[2], JIANXIN LI[3], JUNHUA GU[2], XIANCHAO TANG[4], AND XINYUN XU[2]

[1]School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China
[2]School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
[3]School of Info Technology, Deakin University, Melbourne, VIC 3000, Australia
[4]China Academy of Electronics and Information Technology, Beijing 100089, China

Corresponding author: Suqi Zhang (zhangsuqie@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC0820106, and in part by the National Natural Science Foundation of China under Grant 61802282 and Grant 61972442.

ABSTRACT Identification of community structures is essential for characterizing and analyzing complex networks. Having focusing primarily on network topological structures, most existing methods for community detection ignore two types of non-topological relationships among nodes, i.e., pairwise ''must-link'' constraints among pairs of nodes and labels of nodes, such as functions they may have. Here, we present a novel semi-supervised and active learning method for community detection to integrate these two types of information of a network so as to increase the accuracy of community identification. Our new method will honor the ''must-link'' relationship without introducing new parameters and is efficient with a guaranteed convergence. An essential component of the method is a linear representation that is particularly suited to an active learning to help select the most critical nodes that impact community discovery. We present results from extensive experiments on synthetic and real networks to show the superior performance of the new methods over the existing approaches.

INDEX TERMS Community detection, non-negative matrix factorization, semi-supervised learning, active learning.

I. INTRODUCTION

Networks in real world are not random, but rather contain groups or community structures, which manifest organizational structures and functional components of the underlying systems. The nodes within a community are densely connected, while nodes in different communities are sparsely connected [1]. Community structures are essential characteristics of complex systems [2]; the nodes in the same community tend to share the same or similar attributes and may have the same function. Identification of communities is an important step toward characterization of a complex system as a whole and understanding of the functional roles of individual components of the system. Much effort has been devoted to community identification for network analysis. Most exiting methods for community detection focus on network topological structures alone without taking into consideration non-topological information on nodes or links [3]–[5], and thus can be considered to fall into the category of unsupervised learning. Since network topology is merely one aspect of a network, ignoring other description may prevent accurate community identification.

Furthermore, the community detection method is introduced to reveal the networks clustering characteristics [6]. To be specific, there are two main non-topological information: given labels for a small set of nodes and the pairwise must-link constraints to restrain two nodes from being put into two communities. For instance, in World Wide Web network, it is not difficult to see which webpages belong to sport categories, and which webpages share common features on movies. In social network, it is also easy to determine who must belong to the same university or company.

Non-topological information can be utilized for community detection through semi-supervised learning [7], [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

Indeed, semi-supervised learning algorithms have been proposed recently for community identification. In particular, Ma *et al.* [9] incorporates pairwise constraints on nodes in symmetric nonnegative matrix factorization (NMF); Eaton and Mansbach [10] develops a semi-supervised spin-glass model from statistical physics to combine both pairwise constraints among nodes and community labels on nodes, with the existing modularity-based community detection method; Zhang [11] extends pairwise constraints to other community detection methods, such as spectral clustering; and Zhang [12] designs an enhanced semi-supervised learning approach to utilize pairwise constraints through logical inference. However, all these methods introduce additional parameters, generally more than two, which are difficult to tune in practice, to make tradeoffs between pairwise node constraints with network topological constraints. As a result, none of them is able to guarantee that two nodes having a pairwise must-link constraint are indeed assigned to the same community.

Considering the non-topological must-link constraints as discussed above, it is not practical to treat the network's nodes as equal players in identifying communities of the network. It pays to identify those nodes that contribute the most to improve community detection if they are assigned to the right communities they belong to. A feasible alternative is active learning method. Intuitively, if a node has a small impact on other nodes, its information may have little value. On the other hand, if a node is highly related to and/or has a large influence on other nodes, using its additional information can help produce better community structures. The problem of identifying such informative nodes and adopting such labeling information in a learning process has been cast as an active learning problem [13], [14]. Using the Hamming distance, Zhang [11] adds the must-link and cannot-link constraints to nodes with the largest and the smallest distances, respectively; despite using pairwise constraints and active learning, the results are not competitive comparing with a random scheme. Leng *et al.* [15] proposes to select critical nodes through a complex and costly graph algorithm. Nevertheless, this method only aims at selecting critical nodes that can cover as many communities as possible, which does not consider the nodes lying in the boundary of communities that may also have impact on community structures. Yang *et al.* [16] proposes an active link selection framework. According to the entropy of nodes, they denoted the hub and boundary nodes, and then decided to add or delete links between them. So, they efficiently sharpened the block structure of the network's adjacency matrix. But this method is designed for selecting links rather than nodes and it is an iterative procedure with high computation complexity. Yafang *et al.* [17] proposes an active semi-supervised community detection method. This method actively selects a small amount of links as side information to "sharpen" the boundaries; meanwhile, refers the number of communities automatically. This method transfers the network topology into similarity space for node representation, therefore,the feature dimensionality will be too high as an increase of nodes in real networks. All of these methods rely on prior information excessively.

To this end, in this paper we propose a novel, NMF-based, semi-supervised and active learning method, to exploit non-topological information for community detection. It guarantees to abide by the additional constraints on links and nodes. An efficient updating rule is devised to facilitate fast convergence for the optimization problem of NMF. A salient feature of our new basic semi-supervised method is parameter free, except the number of communities that is typically required, making it easily applicable to large and complex real networks. As an extension to this method, we adopt a linear representation scheme to help identify critical vertices that have the most impact on community discovery and exploit such information in active learning. Our main idea is making the active learning process less random and more interpretable. The new methods are then compared with several existing methods on several synthetic and real networks.

We summarize the main contribution as follows.

1) We present a new Semi-supervised community detection which utilizes the must-link priors by constructing a constraint matrix.

2) We introduce a linear combination of the topological structure to active learning. The most critical nodes that impact community discovery can be selected by this linear representation. Furthermore, we can give them labels or must-link constraints and then use the proposed method in 1).

3) Extensive experiments on both artificial and real dynamic networks based on two evaluation indexes demonstrate demonstrate the superiority of the proposed Semi-supervised community detection in comparison with state-of-art methods.

The paper is organized as follows. We introduce the related work in Section II. The new methods is illustrated in Section III, including the semi-supervised and active learning methods. Experimental results of our algorithms and their comparison with other algorithms are reported in Section IV. We close with some conclusions in the last Section V.

II. RELATED WORK

In this section, we briefly review related work on semi-supervised community detection.

Community detection can capture community structures in complex networks by analyzing the associations between nodes in the network.In the past few years, a large number of community detection algorithms have been proposed and some of them have achieved good performance in many fields [2], [3] [18]–[20]. Most of these methods, however, only take into account the topology information but ignore some prior information. that is important for community detection.In recent years, many semi-supervised community detection algorithms are proposed.By making use of the supervised information or background information, they significantly improved the performance of traditional

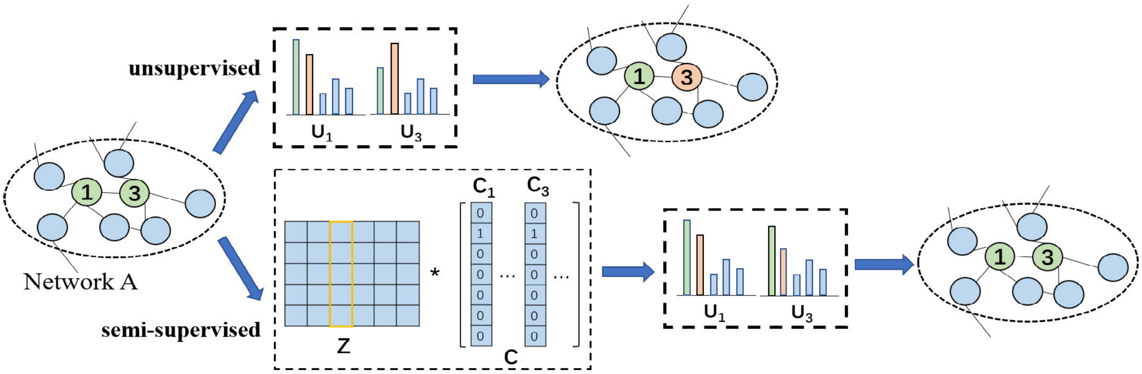


FIGURE 1. An illustration of unsupervised and semi-supervised methods. Given a network A and the must-link constraint between nodes 1 and 3 (marked in green), the unsupervised method obtains different community memberships U1 and U3 for the two nodes, so there is no guarantee that they must be in the same community. The semi-supervised method constructs a constraint matrix **C** whose 1-th and 3-th columns (i.e., **C**1 and **C**3) are the same and U1 = U3 so that nodes 1 and 3 will be in the same community.

topology-based method [16].Yang *et al.* [21] interprets community detection as the clustering problem in the latent space, and then adds a graph regularization term to penalize the latent space dissimilarity of nodes which have "must-link" constraints. Fan *et al.* [22] propose a fast and semi-supervised community detection method that integrates the prior information into the distance dynamics models. Yang *et al.* [23] propose a framework that transforms the original network into an equivalent but much smaller Super-Network by constructing the indivisible super-nodes and by forming the weighted super-edge based on network topology and cannot-link constraints. They, however, often ignore the problem that which supervised link information is the most important and useful information for performance improvement, and they only add the randomly selected supervised link information.To address the former problem, many semi-supervised community detection methods based on active link or node selection for pair-wise constraints are proposed [15]–[17].But all of them are too subjective,that is, rely on prior information immoderately. In recent three years, the node attribute is proposed as priori information [24], [25] and merged with the network topology, so as to improve the accuracy of semi-supervised community detection, especially in using deep learning to mine node content [26]–[28]. This approach is to improve the accuracy of semi-supervised community detection from another perspective. In theory, it is better to incorporate constraints into topology structure.

III. THE PROPOSED FRAMEWORK

In this section, we introduce the notations and our new proposed semi-supervised community detection framework.

A. PRELIMINARIES

Firstly, we introduce some important notations used throughout the paper. We are interested in finding k communities in an un-directed network $G(V, E)$ with n nodes V and m edges E, represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Let U_{ij} be the propensity of node j belonging to community i. The community membership of all the nodes of the network is then $\mathbf{U} = (U_{ij})$, for $i = 1, 2, \ldots, n$, where the jth column,$\mathbf{U}_j \in \mathbb{R}^{k\times 1}$,is the community membership of node j. The overall objective of the community finding problem is to compute community identities for the nodes in G. If network topology denoted by the adjacency matrix $\mathbf{A}$ offers the only guidance for community finding, the topological structure and the node community membership $\mathbf{U}$ to be determined should be as closely consistent as possible, which gives rise to the following optimization function [29]:

$$\min_{\mathbf{U}\geq\mathbf{0}} L(\mathbf{U}) = \left\|\mathbf{A} - \mathbf{U}^T\mathbf{U}\right\|_F^2, \tag{1}$$

where $\|\mathbf{X}\|_F$ is the Frobenius norm of matrix $\mathbf{X}$. In particularly, this can be viewed as a low-rank dimension reduction, where the rank corresponds to the number k of communities. We employ k probabilistic communities to describe the network. Then we can use $U_{ij}U_{ir}$ to present the expected number of links lies between nodes j and r in community i. Summing over communities j, the expected number of links between j and r in the network is $\sum_{i=1}^{k} U_{ij}U_{ir}$. Using squared loss, the problem of fitting the model to the given network is the symmetric-NMF formulation (1). The above function then leads to an unsupervised method using only network topology [30].

B. MODEL OVERVIEW

In this section, we will have an overall view of the proposed new semi-supervised community detection method. Our framework includes two aspects:constraint matrix construction and active node selection,named **SSNMF** and **SSNMF_AL**,respectively.

In SSNMF, must-link constraints are transformed to the form of constraint matrix.In fact, the matrix reflects the relationship between node and constraint. A new non-negative auxiliary matrix are also proposed to map constraint to community, and then, node community membership matrix can be find by optimizing the objective function. In this way, constraint information can be fully used. We illustrate in Fig.1 the

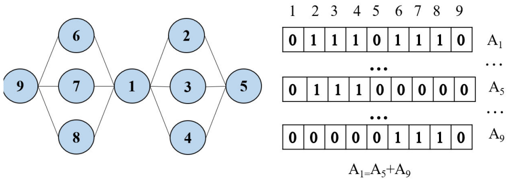


FIGURE 2. **An illustration of linear representation. $\mathbf{A}_1$, $\mathbf{A}_5$ and $\mathbf{A}_9$ are the topological structures of nodes 1, 5 and 9 in the left network, respectively.**

effects of unsupervised and semi-supervised methods on a given network **A** with nodes 1 and 3 that in fact belong to the same community (marked in green), because of a must-link constraint. An unsupervised method (with no must-link constraint) may produce different community memberships for nodes 1 and 3 and consequently assign the two nodes to different communities, marked in green and red. In contrast, the must-link constraint used by the semi-supervised method will result in the same community membership for the two nodes, guaranteed that nodes 1 and 3 to belong to the same community, marked in green.

In SSNMF_AL, linear representation is critical for mining active nodes. For simplicity, we assume a undirected and unweighted graph which contains 9 vertices as shown in Fig.2. To see the rationale of this representation, consider a social network, if two people share common friends, they may share similar topological structures as well. Therefore, it is viable to represent one node by the other nodes. To be concrete, consider a toy example in Figure 2. Although node 1 does not directly connect to nodes 5 and 9, its topological structure can be represented by a linear combination of that of $\mathbf{A}_5$ and $\mathbf{A}_9$, because node 1 shares common nodes 2, 3 and 4 with node 5 and common neighbors 6, 7 and 8 with node 9.

IV. CONSTRAINT MATRIX CONSTRUCTION IN SSNMF

Consider must-link constraint in addition to the adjacency matrix. Assume for a given collection of sets of nodes, all nodes in one set must appear in the same community. Specifically, we are given $P = P_1 \cup P_2 \cup \ldots \cup P_q$, where P_i is a set of nodes that must be assigned to the same community as specified by the must-link constraints. As the must-link relationship is transitive, we have $P_i \cap P_j = \emptyset$. Apparently, P defines the initial, incomplete q communities. The remaining nodes not in P may be assigned to these q communities or form communities of their own. To start looking for communities using the must-link constraints and network adjacency matrix, we first put each of the unassigned nodes (singletons) into its own set. In other words, the network is temporally partitioned into $q+n-p$ subsets, represented as $\left\{\bigcup_{i=1}^{q+n-p} P_i\right\}$, where p is the total number of nodes in P. For example, in a network of 6 nodes, suppose that nodes 1 and 2 must be in one community and nodes 4 and 5 in another, the remaining nodes 3 and 6 are singletons whose community identities are to be determined. That is, $P_1 = \{1, 2\}$, $P_2 = \{4, 5\}$, $P_3 = \{3\}$ and $P_4 = \{6\}$.

We now introduce a non-topological constraint matrix $(C_{ij}) \in \mathbb{R}^{(q+n-p)\times n}$, where $C_{ij} = 1$ if node j is in the subset P_i, or $C_{ij} = 0$, otherwise. This means that if nodes i and j belong to the same community, the i-th and j-th columns of **C** are equal. For the above simple 6-node network, the non-topological constraint matrix **C** is

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

In essence, the constraint matrix **C** captures the relationship between nodes and constraints. To facilitate finding node-community membership, we introduce a new nonnegative auxiliary matrix $\mathbf{Z} = (Z_{ij}) \in \mathbb{R}^{k\times(q+n-p)}$ to map constraints to communities. More importantly, we turn the problem of finding node community memberships **U** into the problem of finding **Z** such that $\mathbf{U} = \mathbf{ZC}$. Consequently, the original objective function in (1) can then be rewritten as:

$$\min_{\mathbf{Z}\geq \mathbf{0}} L(\mathbf{Z}) = \left\| \mathbf{A} - \mathbf{C}^T\mathbf{Z}^T\mathbf{ZC} \right\|_F^2. \tag{2}$$

Unlike the other methods that introduce extra parameter to balance additional information with the network structures, our objective function is parameter free, which greatly simplifies the process of incorporating additional information. Further, if there is no additional information, the constraint matrix **C** becomes an identity matrix $\mathbf{I} \in \mathbb{R}^{n\times n}$, and the nonnegative auxiliary matrix **Z** is equal to the community membership matrix **U**. Therefore, the objective function for unsupervised community detection in (1) is a special case of the new objective function in (2).

A. PARAMETERS EVALUATION

In order to derive the updating rule for optimizing the objective function in (2), we introduce a Lagrange multiplier matrix $\Theta = (\Theta_{ij})$ for the nonnegative constraints on **Z** to (2), resulting in the following equivalent objective function:

$$L(\mathbf{Z}) = tr(\mathbf{C}^T\mathbf{Z}^T\mathbf{ZCC}^T\mathbf{Z}^T\mathbf{ZC}) - 2tr(\mathbf{AC}^T\mathbf{Z}^T\mathbf{ZC}) + tr(\mathbf{AA}) + tr(\Theta\mathbf{Z}^T).$$

For any stationary state, we have

$$\frac{\partial L(\mathbf{Z})}{\partial \mathbf{Z}} = 4\mathbf{ZCC}^T\mathbf{Z}^T\mathbf{ZCC}^T - 4\mathbf{ZCAC}^T + \Theta. \tag{3}$$

By setting (3) to 0 and using the Karush-Kuhn-Tucker condition $\Theta_{ij}Z_{ij}^4 = 0$, we then have:

$$(\mathbf{ZCC}^T\mathbf{Z}^T\mathbf{ZCC}^T)_{ij}Z_{ij}^4 - (\mathbf{ZCAC}^T)_{ij}Z_{ij}^4 = 0,$$

which gives rise to the following updating rule for **Z**:

$$Z_{ij} = Z_{ij}\left(\frac{(\mathbf{ZCAC}^T)_{ij}}{\mathbf{ZCC}^T\mathbf{Z}^T\mathbf{ZCC}^T)_{ij}}\right)^{\frac{1}{4}}. \tag{4}$$

To prove the convergence of the updating rule, we have the following results.

Theorem 1 : The objective function of (2) is non-increasing under the iterative updating rule (4).

Proof : The proof follows the auxiliary function method in [31]. If a function satisfies

$$L(\mathbf{Z}) \leq F(\mathbf{Z}, \mathbf{Z}'), \forall \mathbf{Z}', F(\mathbf{Z}, \mathbf{Z}) = L(\mathbf{Z}),$$

$F(\mathbf{Z}, \mathbf{Z}')$ is then an auxiliary function of $L(\mathbf{Z})$. We can define $\mathbf{Z}^{t+1} = argmin_{\mathbf{Z}} F(\mathbf{Z}, \mathbf{Z}^t)$. And then have

$$L(\mathbf{Z}^{t+1}) = F(\mathbf{Z}^{t+1}, \mathbf{Z}^{t+1}) \leq F(\mathbf{Z}^{t+1}, \mathbf{Z}^t) \leq L(\mathbf{Z}^t).$$

This proves that $L(\mathbf{Z})$ is monotonically non-increasing.

Next we need to find the specific form of the auxiliary function $F(\mathbf{Z}, \mathbf{Z}')$ for the objective function (2) as followed,

$$\begin{aligned} L(\mathbf{Z}) &= \left\| \mathbf{A} - \mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C} \right\|_F^2 \\ &= tr(\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) - 2tr(\mathbf{A}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) + tr(\mathbf{A}\mathbf{A}) \\ &= tr(\mathbf{C}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}) - 2tr(\mathbf{A}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) + tr(\mathbf{A}\mathbf{A}) \end{aligned}$$

(by Lemma 1)

$$\begin{aligned} &\leq \frac{1}{2} tr(\mathbf{C}\mathbf{C}^T\mathbf{P}\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}' + \mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}'\mathbf{C}\mathbf{C}^T\mathbf{P}) \\ &\quad - 2tr(\mathbf{A}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) + tr(\mathbf{A}\mathbf{A}) \\ &\leq tr(\mathbf{P}\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}'\mathbf{C}\mathbf{C}^T) - 2tr(\mathbf{A}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) + tr(\mathbf{A}\mathbf{A}) \end{aligned}$$

(by Lemma 2)

$$\leq tr(\mathbf{R}\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}'\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T) - 2tr(\mathbf{A}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}) + tr(\mathbf{A}\mathbf{A})$$

(by Lemma 3)

$$\begin{aligned} &\leq tr(\mathbf{R}\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}'\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T) - 2tr(\mathbf{Z}'\mathbf{C}\mathbf{A}\mathbf{C}^T\mathbf{Q}^T) \\ &\quad - 2tr(\mathbf{Q}\mathbf{C}\mathbf{A}\mathbf{C}^T\mathbf{Z}'^T) - 2tr(\mathbf{Z}'\mathbf{C}\mathbf{A}\mathbf{C}^T\mathbf{Z}'^T) + tr(\mathbf{A}\mathbf{A}) \\ &= F(\mathbf{Z}, \mathbf{Z}'), \end{aligned}$$

where $P_{ij} = \frac{(\mathbf{Z}^T\mathbf{Z})_{ij}^2}{(\mathbf{Z}'^T\mathbf{Z}')_{ij}}$, $R_{ij} = \frac{Z_{ij}^4}{Z_{ij}'^3}$, and$Q_{ij} = Z_{ij}' ln(\frac{Z_{ij}}{Z_{ij}'})$.This provides the specific form $F(\mathbf{Z}, \mathbf{Z}')$of the auxiliary function for objective function (2). We can then have the solution for $min_{\mathbf{Z}} F(\mathbf{Z}, \mathbf{Z}')$ by the following Karush-Kuhn-Tucker condition

$$\begin{aligned} \frac{\partial F(\mathbf{Z}, \mathbf{Z}')}{\partial \mathbf{Z}_{ij}} &= 4\frac{Z_{ij}^3}{Z_{ij}'^3}(\mathbf{Z}'\mathbf{C}\mathbf{C}^T\mathbf{Z}'^T\mathbf{Z}\mathbf{C}\mathbf{C}^T)_{ij} \\ &\quad - 2\frac{Z'_{ij}}{Z_{ij}}(\mathbf{Z}'\mathbf{C}\mathbf{A}^T\mathbf{C}^T + \mathbf{Z}'\mathbf{C}\mathbf{A}\mathbf{C}^T)_{ij} = 0, \end{aligned}$$

which gives rise to the updating rule in (4). Therefore, under this updating rule the objective function $L(\mathbf{Z})$ of (2) will monotonically decrease, and converge to a local minimum.

The proofs to Lemmas 1 to 3 are in Appendix.

The detailed algorithm is described in Algorithm 1.

Algorithm 1 Proposed Semi-Supervised Method

Input: The adjacency matrix **A**, the number of communities k, and the set of labeled nodes P

Output: The community membership matrix **U**

```
1: Initialize auxiliary matrix Z;
2: According to P, constructing the constraint matrix C;
3: while not converge do
4:     Update Z via (4);
5: end while
6: U = ZC;
7: return U.
```

B. ACTIVE NODE SELECTION IN SSNMF_AL

Not every node in a network equally affects community discovery. A node that is highly related to others may have a large influence on community detection and knowing its community identity may affect the community membership of the other nodes. Therefore, we are interested in identifying such critical nodes, and using such "labeled" nodes in active learning. To this end, we consider to represent the topological structure of a node by a linear combination of the topological structures of the other nodes.

Formally, the topological structure $\mathbf{A}_i$ of node i can be rep-resented by a linear combination of the topological structure of other nodes, i.e.,$\mathbf{A}_i = \sum_{j=1, j\neq i}^{n} B_{ij}\mathbf{A}_j$. The coefficient B_{ij} denotes the correlation between nodes i and j.

Different from the node connectivities, this new representation captures deep structural similarities among nodes beyond the information in the adjacency matrix. Even though two nodes do not directly link with one another, they may still be highly related to or similar with each other. This new representation enables us to discover the most critical nodes that influence community structures.

If the topological structures of nodes can be obtained by a linear representation of the topological structures of the other nodes, the obtained network structure and original structure A should be as closely consistent as possible, which gives rise to the following objective function:

$$\min_{\mathbf{B}\geq\mathbf{0}} L(\mathbf{B}) = \|\mathbf{A} - \mathbf{A}\mathbf{B}\|_F^2 ,$$

where **B** holds the nonnegative coefficients. Furthermore, in order to have a better discriminative power and find the most critical nodes, we add to **B** a $l_{2,1}$ norm constraint, $\|\mathbf{B}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{N} B_{i,j}}$. Minimizing $\|\mathbf{B}\|_{2,1}$ results in a group sparse solution to **B**, i.e., some rows of B are all zeros, giving a better discriminative power to distinguish the most critical nodes. We thus have the following objective function for finding **B**:

$$\min_{\mathbf{B}\geq\mathbf{0}} L(\mathbf{B}) = \|\mathbf{A} - \mathbf{A}\mathbf{B}\|_F^2 + \beta \|\mathbf{B}\|_{2,1} , \tag{5}$$

where the diagonal elements B_{ii} are forced to 0, making a node to be represented by other nodes except for itself, otherwise, there may be a trivial solution, i.e., each node can be represented by themselves perfectly. β is the weight of $l_{2,1}$ norm's contribution to the overall objective.

The final resultant $\mathbf{B}$ provides information of critical or representative nodes of the network. Specifically, these nodes must have many non-zero entries in the final $\mathbf{B}$ as they are highly related to the other nodes. We can sort the rows of $\mathbf{B}$ in a decreasing order by the row-sum values. The higher a value is, the more critical a node will be. For active learning, we take the first t nodes that have the highest row-sum values of the final $\mathbf{B}$.

C. INFER PARAMETERS OF ACTIVE LEARNING

To perform active learning, we rewrite the objective function(8) as:

$$L(\mathbf{B}) = tr(\mathbf{A}\mathbf{A}^T - \mathbf{A}\mathbf{B}^T\mathbf{A}^T - \mathbf{A}\mathbf{B}\mathbf{A}^T + \mathbf{A}\mathbf{B}\mathbf{B}^T\mathbf{A}^T) + \beta \left\|\mathbf{B}\right\|_{2,1}. \tag{6}$$

To take care of the constraints $B_{ij} \geq 0$, we introduce Largrange multipliers $\boldsymbol{\gamma} = (\gamma_{ij})$ and revise the objective function using a Largrange function $\mathcal{L}(\mathbf{B})$ as follows:

$$\mathcal{L}(\mathbf{B}) = L(\mathbf{B}) + Tr(\boldsymbol{\gamma}\mathbf{B}^T). \tag{7}$$

Taking the derivative of (7) w.r.t $\mathbf{B}$, we have

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \mathbf{B}} = -2\mathbf{A}^T\mathbf{A} + 2\mathbf{A}^T\mathbf{A}\mathbf{B} + 2\beta\mathbf{D}\mathbf{B} + \boldsymbol{\gamma}, \tag{8}$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element $D_{ii} = \frac{1}{\|\mathbf{B}_i\|_F}$ and $\mathbf{B}_i$ is the i-th row of $\mathbf{B}$. By setting (8) to 0 and using the Karush-Kuhn-Tucker condition $\gamma_{ij}B_{ij} = 0$, we then have:

$$(\mathbf{A}^T\mathbf{A})_{ij}B_{ij} - (\mathbf{A}^T\mathbf{A}\mathbf{B} + \beta\mathbf{D}\mathbf{B})_{ij}B_{ij} = 0,$$

which gives rise to the following updating rule:

$$B_{ij} = B_{ij}\frac{(\mathbf{A}^T\mathbf{A})_{ij}}{(\mathbf{A}^T\mathbf{A}\mathbf{B} + \beta\mathbf{D}\mathbf{B})_{ij}}. \tag{9}$$

Here, we can set $B_{ii} = 0$ at each iteration to simplify the computation.

The detailed algorithm is described in Algorithm 2. After we obtain a set of selected nodes by proposed active learning method, we can give them labels or must-link constraints, and then we can build the constraint matrix and use the proposed semi-supervised method above to get the community memberships.

D. COMPLEXITY

The adjacency matrix $\mathbf{A}$ of most real networks and their community constraint matrices $\mathbf{C}$ are sparse. In the semi-supervised learning, bulk of the computation spends on the updating rule for optimizing the objective function in (4). The computation of $\mathbf{Z}\mathbf{C}\mathbf{A}\mathbf{C}^T$ and $\mathbf{Z}\mathbf{C}\mathbf{C}^T\mathbf{Z}^T\mathbf{Z}\mathbf{C}\mathbf{C}^T$ in (4) runs in $\mathcal{O}(kn(q+n-p)+k(q+n-p)^2)$ and $\mathcal{O}((q+n-p)^2+k(q+n-p)+k^2(q+n-p))$, respectively. Since q, p and $k \ll n$, the total computational cost of (4) is $\mathcal{O}(n^2)$. Consequently, the time complexity for semi-supervised learning is $\mathcal{O}(T(n^2))$ for T iterations to convergence. For active learning, the main computation is also for the updating rule in (12). The time to compute $\mathbf{A}^T\mathbf{A}$, $\mathbf{A}^T\mathbf{A}\mathbf{B}$ and $\beta\mathbf{D}\mathbf{B}$ is $\mathcal{O}(n^2)$, so the cost to update $\mathbf{B}$ once is $\mathcal{O}(n^2)$ as well, and the time to find critical nodes for active learning is $\mathcal{O}(T(n^2))$ for a total of T iterations to update $\mathbf{B}$.

Algorithm 2 Proposed Active Learning Method

Input: The adjacency matrix $\mathbf{A}$, the number of selected nodes t, and β

Output: The set of selected nodes P

```
Initialize auxiliary matrix $\mathbf{B}$;
while not converge do
    Update $D_{ii} = \frac{1}{\|\mathbf{B}_i\|_F}$;
    Update $\mathbf{B}$ via (9);
    $B_{ii} = 0$;
end while
Compute the row-sum values of $\mathbf{B}$.
Sort the rows of $\mathbf{B}$ in a decreasing order by the row-sum values;
Select the first t nodes as $P$;
return $P$.
```

Note that most computation is by matrix multiplication; to gain efficiency, we can adopt sparse matrix multiplication. Moreover, the updating rule for each element at each iteration is independent, therefore, updating all elements can be carried out in parallel. We may consider CUDA [32], cuBLAS or other parallel computing platforms with matrix computation currently available to further speed up the computation.

V. EXPERIMENTS

A. BASELINES FOR COMPARISON

We experimentally evaluated our methods, named as SSNMF for the basic algorithm and SSNMF_AL for the method with active learning, by comparing them with four state-of-the-art community detection methods. The algorithms that we evaluated are listed below:

- Our proposed semi-supervised symmetric nonnegative matrix factorization method (SSNMF).
- Symmetric nonnegative matrix factorization (SNMF) [29]. This is a popular method based on symmetric nonnegative matrix factorization for community detection. It is also an unsupervised method, which only takes topology structures into account. Notice that, it is also the basic model for our semi-supervised model when there is no prior knowledge.
- Zhang's method [11]. This is a semi-supervised community detection framework which integrates both of the must-link and cannot-link constraints by modifying the adjacency matrix. Here we chose standard NMF with least square error, which often shows the best performance compared with the other NMF-based methods under this framework. Besides, as suggested in [10],

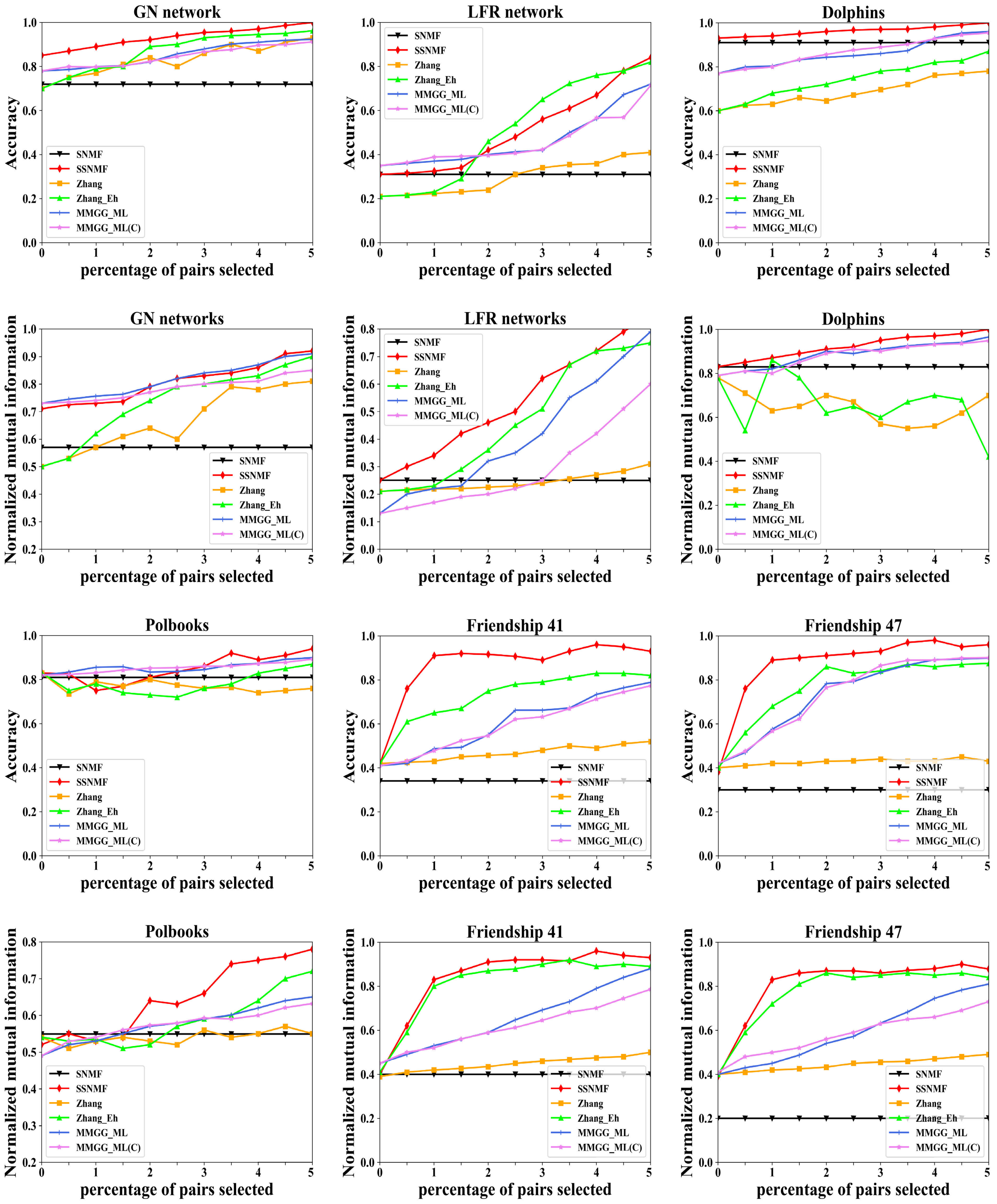


FIGURE 3. Comparison of SSNMF and five competing methods on the six networks, measured by accuracy and normalized mutual information.

the parameter which measures the weight of constraints is specified as 2.

- Zhang's Enhancement method (Zhang_Eh) [12]. This is an enhanced semi-supervised community detection method. Different from the previous one, in order to fully utilize the must-link and cannot-link constraints, the method adds a logical inference step to infer more must-link and cannot-link constraints. We also chose NMF with least square error and specified the parameter measuring the weight of constraints as 2.

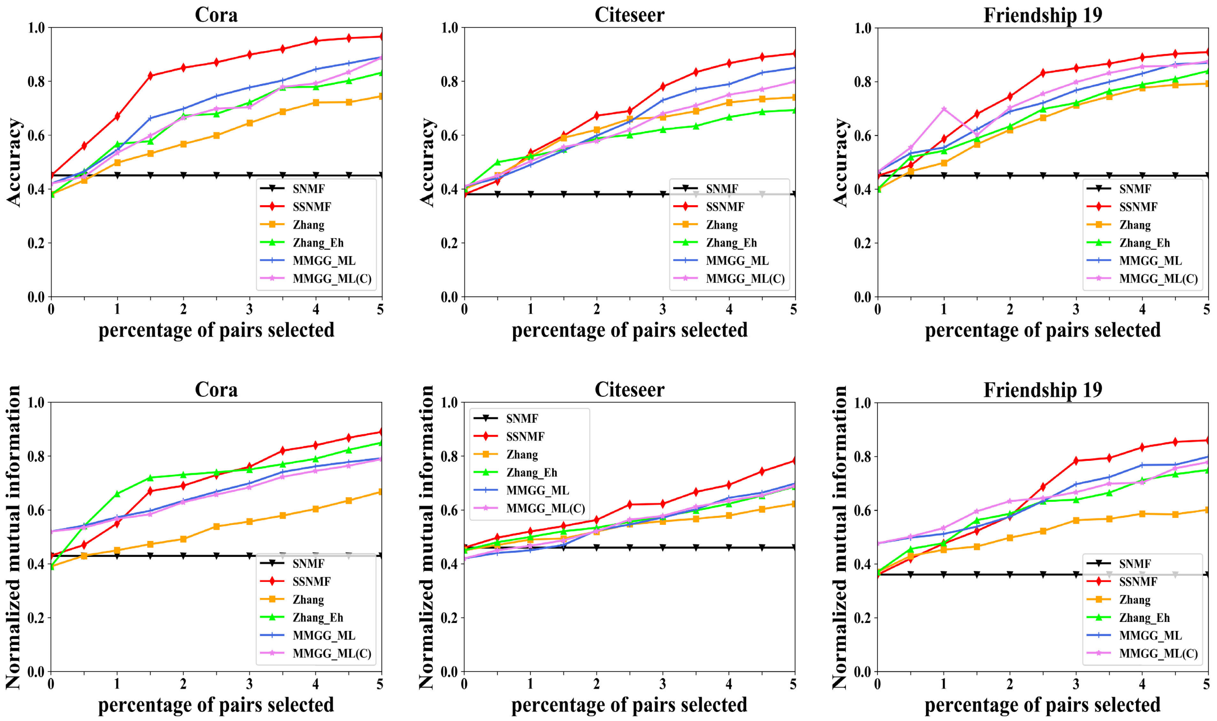


FIGURE 4. Comparison of SSNMF and five competing methods on the three networks, measured by accuracy and normalized mutual information.

- Multi-variance Mixed Gaussian Generative method (MMGG) [23]. This model considers the generation process as a Mixed Gaussian Model with Multi-variance of the network topology, must-link and cannot-link constraints together. Then semi-supervised community detection can be solved via a weighted nonnegative matrix factorization method. We use three methods in the must-link subgraph for comparison: MMGG-ML and MMGG-ML(C).

B. DATA USED

We considered synthetic and real networks in our experimental study.

- Girvan-Newman (GN) network [33]. It is a synthetic network which consists of 128 nodes, divided into four communities of 32 nodes each. Each node has an expected degree of 16, including an average z_{in} links connecting to nodes within the same community and z_{out} links to nodes in the other communities. As z_{out} increases, the community structure becomes weaker and more difficult to identify. In our experiment, we set $z_{out} = 8$, to make the number of between-community links per node equal to that of within-community links, so that there is no clear community structure, making the problem difficult.
- Lancichinetti-Fortunato-Radicchi (LFR) network [34]. This is also a synthetic network. It has a few parameters to tune, which is thus more sophisticated than the GN network, and resembles more closely real networks. The LFR network has 1000 nodes; the average node degree is 20, and the maximum node degree is 50; the exponent of the degree and the community size distributions are -2 and -1, respectively; and the mixing parameter μ, i.e., the fraction of the links of a node connecting with nodes in the other communities, is 0.8. The communities in the LFR network are non-overlapping.
- Dolphins social network (Dolphins) [35]. In this real network, dolphins are represented as nodes and each link with two dolphins represents that they are observed together more often than expected by chance over a period of seven years from 1994 to 2001. According to their genders, the network is divided into the male and female dolphin communities, respectively.
- Political books network (Polbooks) [36] In this real net-work, the nodes represent the books about US politics sold by Amazon.com. Links represent frequent co-purchasing of books by the same buyers. Generally, according to their political viewpoints, these books are divided into three communities: liberal, neutral, and conservative communities.
- Core networks(Cora) [33].The Cora dataset consists of 2708 scientific publications classified into one of seven classes. The citation network consists of 5429 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from

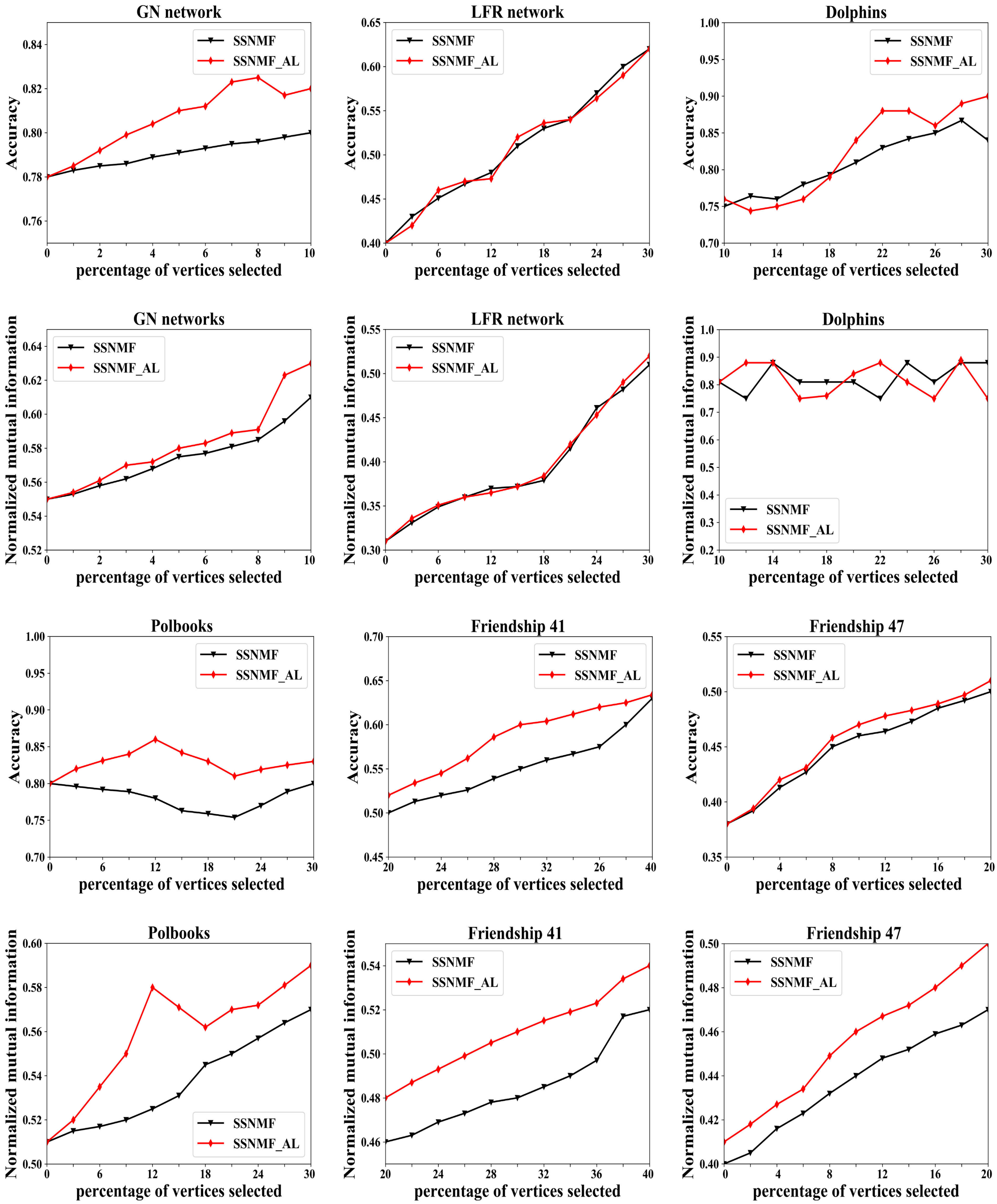


FIGURE 5. **Comparison of SSNMF with additional information from random selection scheme and SSNMF_AL with information from active learning on the six networks tested.**

the dictionary. The dictionary consists of 1433 unique words.

- CiteSeer networks(CiteSeer) [36].The CiteSeer dataset consists of 3312 scientific publications classified into one of six classes. The citation network consists of 4732 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 3703 unique words.
- Friendship networks [37]. This is a real Facebook social network in the U.S. The friendships are undirected,

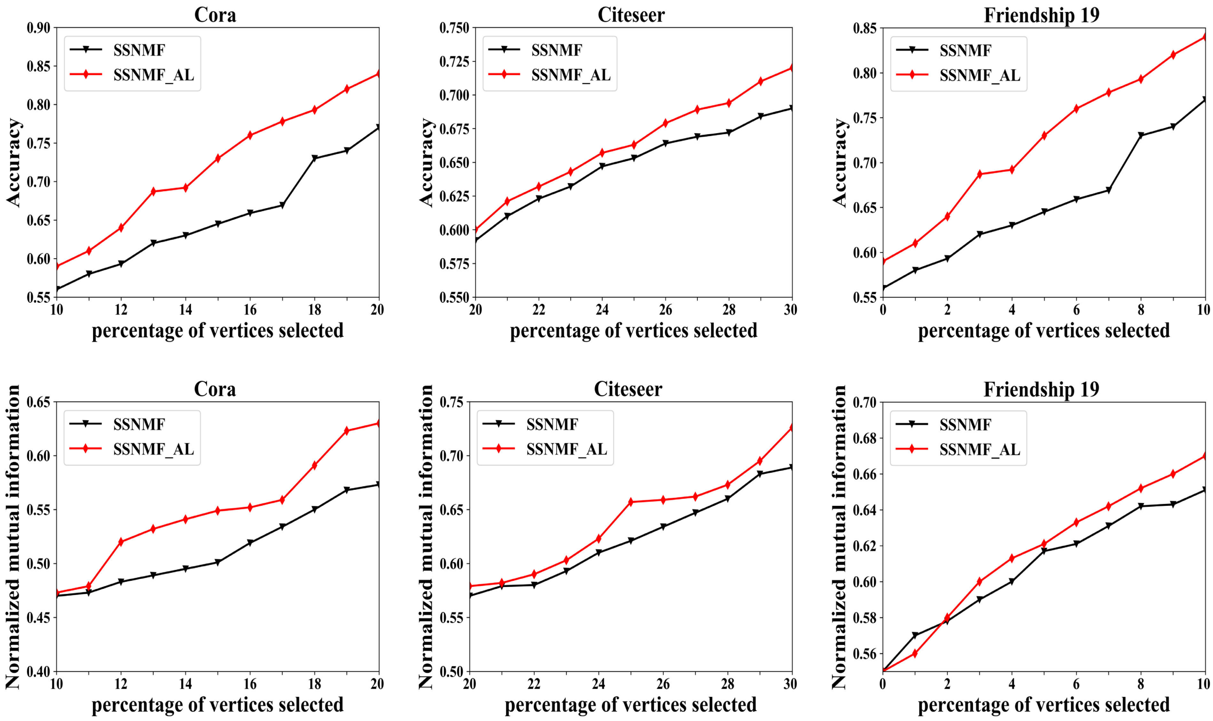


FIGURE 6. Comparison of SSNMF with additional information from random selection scheme and SSNMF_AL with information from active learning on the three networks tested.

TABLE 1. Four real networks with known community structures which were tested. Here, *n*, *m* and *k* are the number of nodes, links and communities, respectively.

Datasets	n	m	k
Dolphins social network	62	159	2
Political books network	105	441	3
Cora	2,708	5,429	7
Citeseer	3,312	4,732	6
Friendship 41	2,235	90,954	16
Friendship 47	2,252	84,387	19
Friendship 19	13,882	381,935	7

and there are six pieces of person's metadata: residence hall, major, second major, class year, former high school and gender, respectively. According to [37], the class year is used as the ground-truth of community structure.

The specific statistical features of these four real networks can be found in Table. 1.

C. EVALUATION METRICS

We adopted two metrics to assess the quality of community detection results.

- Accuracy (AC) [7]. This is used to measure the percentage of nodes that are correctly assigned to the right communities. In particular, for each node, we use l_i as the community of node i provided by algorithms, and r_i as its actual community. Then the accuracy is defined as

$$AC = \frac{\sum_{i=1}^{N} \delta(r_i, map(l_i))}{N},$$

where $\delta(x, y)$ is the delta function whose value equals 1 if $x = y$ or 0, otherwise. The function $map(l_i)$ maps each community l_i to the equivalent community from the network. We use the same mapping method suggested by [7]. The higher the AC, the better the result is.

- Normalized mutual information (NMI) [38]. Given the community detection result, the NMI is estimated by

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} log \frac{n_{ij}\cdot N}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{k} n_i log \frac{n_i}{N})((\sum_{j=1}^{k} \hat{n}_j log \frac{\hat{n}_j}{N})}}$$

where k is the number of communities, n_i denotes the number of nodes in ith computed community, $\hat{n}_j$ denotes the number of nodes in jth ground-truth community, N denotes the number of nodes in the network, n_{ij} is the number of nodes that lies in the intersection between the ith calculated community and the jth ground-truth community. The NMI index measures how similar two sets of communities are. In general, a higher value of NMI represents a better result.

D. SEMI-SUPERVISED LEARNING

To facility the five semi-supervised learning methods we compared, we embedded must-link constraints into each of the synthetic and real networks based on the underlying true node-community identities. We randomly selected a percentage of all possible node pairs; if the two nodes of a pair belong to the same community, as their community identifies are known, a must-link constraint was then added; otherwise a cannot-link constraint was introduced. In method SSNMF, we were given **P** and $P_i \cap P_j = \emptyset$ before constructing a constraint matrix **C**. Therefore, we used logical inference step to ensure it. For each network, we varied the percent of chosen node pairs out of the total node pairs in the network from 0% to 5% with an increment of 0.5%.

FIGURE 3,4 show the results of the six methods on these nine networks. As shown, the performance of SSNMF, as well as that of the other methods that also used additional information, increased with the amount of additional non-topological information provided, as they exploited must-link constraints. SSNMF also outperformed the other methods on most of the network settings tested. Importantly, SSNMF beat the two Zhang methods on nearly all network instances, demonstrating its effectiveness in utilizing the additional non-topological semi-supervised information.

E. SEMI-SUPERVISED AND ACTIVE LEARNING

In order to assess the effectiveness of the active learning method, we considered the non-topological information of two sets of nodes: the ones selected randomly; and the nodes chosen by active learning method. We then directly compared our methods SSNMF using additional information from the random selection scheme and SSNMF_AL with information from active learning.

To make a fair comparison, the same number of nodes was chosen by the active learning and random selection schemes; must-link constraints were introduced to pairs of nodes that have the same community label, based on their true community memberships. For each network, we varied the percent of chosen nodes out of the total nodes in the network from 0% to 100% with an increment of 10%. Note that the parameter β in active learning needs to be fine tuned to get a better result. Specifically, in our experiments, for GN, LFR, dolphins and polbooks networks, we varied β from 0 to 5 with an increment of 0.5; and for Amherst and Bowdoin networks, we varied β from 10 to 40 with an increment of 5. Then we chose the best AC and NMI results. As shown in FIGURE 5 and 6, the performance of both these two methods improved as more additional information was included. More importantly, SSNMF_AL with active learning outperformed SSNMF with random selection on most network instances, showing the effectiveness of the active learning.

F. CASE STUDY

Here we use the real network, i.e., Dolphins network (FIGURE 7) for case study to illustrate our semi-supervised framework. According to their gender, the network is divided into the male and female dolphin communities, respectively. In figure 7, we use the shape to represent the true gender of the dolphins and the color to represent the results of two semi-supervised method, i.e. SSNMF and Zhang_Eh, with prior information. In each plot,the shapes "circle" and "hexagon" represent the ground-truth communities, and the colors "yellow" and "red" represent the results of two semi-supervised method. If the color of one node does not match its ground-truth shape, the node is not classified correctly by the method. We use the number of misclassified nodes to judge the result. Figure 7(a) and Figure 7(c) are the results of SSNMF with 2% and 5% priors. Figure 7(b) and Figure 7(d) are the results of Zhang_Eh with 2% and 5% priors similarly. As shown, misclassified nodes decrease with the percentage of priors provided. We can also find that the performance of SSNMF is better than Zhang_Eh either with 2% or 5% pairwise constraints. Furthermore, all nodes are divided into accurate community in Figure 7(c). In conclusion, SSNMF has better effectiveness.

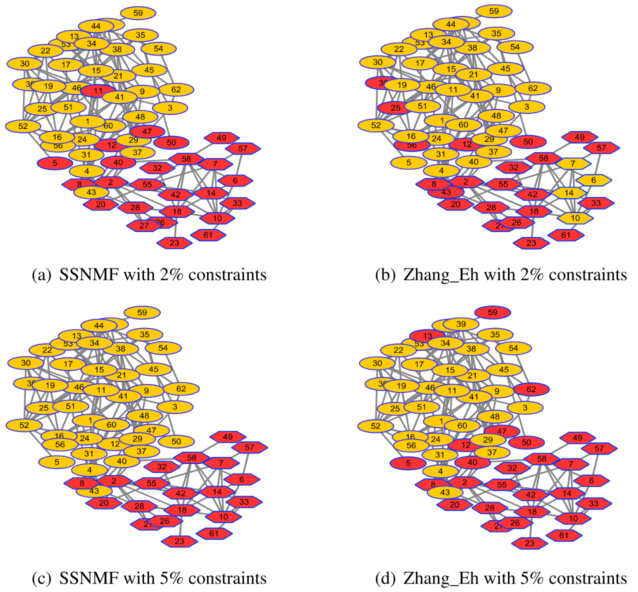

(a) SSNMF with 2% constraints
(b) Zhang_Eh with 2% constraints
(c) SSNMF with 5% constraints
(d) Zhang_Eh with 5% constraints

FIGURE 7. The detected communities by SSNMF and Zhang_Eh.(a)SSNMF with 2% priors, misclassification: 6/62. (b)Zhang_Eh with 2% priors, misclassification: 10/62. (c)SSNMF with 5% priors, misclassification: 0/62.SSNMF with 5% priors, misclassification: 8/62.

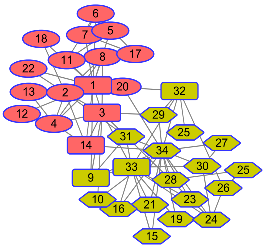


FIGURE 8. An illustration of active learning. Here different colors represent the real community membership, and the rectangle represents the selected nodes.

Here we use the well-known real network, i.e., Karate Club network [23](FIGURE 8), as an example to show which nodes will be selected by our active learning method. In this network, a disagreement developed between the administrator (node33) and the clubs instructor (node 1) of the club, ultimately resulted in the instructor's leave and starting a new club. Thus, the whole network is divided into two communities (represented by green and red colors). The active learning method selected 7 nodes (20 percentages of all the nodes), showed in squares in Figure 6. As shown, the method selected the most representative nodes, such as nodes 1 and 33, so that after we labeled them, they can influent more other nodes. In addition, the method selected the nodes on the boundary of the two communities, such as nodes 3, 9, 14, and 32. Because these nodes connect with both communities, they are typically difficult to assign to their right communities. Intuitively, after selecting them, we can label them directly and thus avoid misclassification.

VI. CONCLUSION

We proposed a novel semi-supervised community detection method to integrate non-topological information, particularly pairwise must-links among pairs of nodes and labels of nodes. Our method guarantees convergence to local optimization, abides by in the final results the additional constraints, and is parameter free, a desirable feature for practical applications. To further improve the semi-supervised method, we chose and labeled some nodes by solving an optimization problem and exploited the labeled information in an active learning. Extensive experimental evaluation on synthetic and real networks showed the superior performance of the new semi-supervised and active learning method. Overall, our results evidently demonstrated the value of non-topological information in community detection.

APPENDIX-LEMMAS

The lemmas and their proofs used in the proof of Theorem 1 follow [25]:

Lemma 1 : If matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}$ are nonnegative, and matrix $\mathbf{Z}'$ is positive. We have

$$tr(\mathbf{B}\mathbf{Z}^T\mathbf{A}\mathbf{Z}) \leq \frac{1}{2}tr(\mathbf{B}\mathbf{P}^T\mathbf{A}\mathbf{Z}' + \mathbf{B}\mathbf{Z}'^T\mathbf{A}\mathbf{P}),$$

where $P_{ij} = \frac{Z_{ij}^2}{Z'_{ij}}$.The equality holds when $\mathbf{Z}' = \mathbf{Z}$.

Lemma 2 : If matrix $\mathbf{A}$ is nonnegative symmetric, matrix $\mathbf{Z}$ is nonnegative, and matrix $\mathbb{Z}'$ is positive. We have

$$tr(\mathbf{PA}) \leq tr(\mathbf{R}\mathbf{A}\mathbf{Z}'^T),$$

where $P_{ij} = \frac{(\mathbf{Z}^T\mathbf{Z})_{ij}^2}{(\mathbf{Z}^T\mathbf{Z})_{ij}}$and $R_{ij} = \frac{Z_{ij}^4}{Z_{ij}'^3}$. The equality holds when $\mathbf{Z}' = \mathbf{Z}$.

Lemma 3 : If matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}$ are nonnegative, and matrix $\mathbf{Z}'$ is positive. We have

$$-tr(\mathbf{B}\mathbf{Z}^T\mathbf{A}\mathbf{Z}) \leq -tr(\mathbf{B}\mathbf{Z}'^T\mathbf{A}\mathbf{Q}) - tr(\mathbf{B}\mathbf{Q}^T\mathbf{A}\mathbf{Z}') - tr(\mathbf{B}\mathbf{Z}'^T\mathbf{A}\mathbf{Z}'),$$

where $Q_{ij} = Z'_{ij}ln\frac{Z_{ij}}{Z'_{ij}}$. The equality holds when $\mathbf{Z}' = \mathbf{Z}$.

REFERENCES

[1] B. Huang, C. Wang, and B. Wang, "NMLPA: Uncovering overlapping communities in attributed networks via a multi-label propagation approach," *Sensors*, vol. 19, no. 2, p. 260, Jan. 2019, doi: 10.3390/s19020260.

[2] C. Liu, J. Liu, and Z. Jiang, "A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2274–2287, Dec. 2014, doi: 10.1109/tcyb.2014.2305974.

[3] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *CoRR*, vol. abs/1308.0971, pp. 95–142, Aug. 2013. [Online]. Available: http://arxiv.org/abs/1308.0971

[4] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using neighborhood-inflated seed expansion," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1272–1284, May 2016, doi: 10.1109/tkde.2016.2518687.

[5] H.-J. Li, Z. Bu, A. Li, Z. Liu, and Y. Shi, "Fast and accurate mining the community structure: Integrating center locating and membership optimization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2349–2362, Sep. 2016, doi: 10.1109/tkde.2016.2563425.

[6] V. Karyotis, K. Tsitseklis, K. Sotiropoulos, and S. Papavassiliou, "Big data clustering via community detection and hyperbolic network embedding in IoT applications," *Sensors*, vol. 18, no. 4, p. 1205, Apr. 2018, doi: 10.3390/s18041205.

[7] H. Liu, Z. Wu, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012, doi: 10.1109/tpami.2011.217.

[8] H. Li, "Detecting fuzzy network communities based on semi-supervised label propagation," *IFS*, vol. 31, no. 6, pp. 2887–2893, Dec. 2016, doi: 10.3233/jifs-169171.

[9] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 389, no. 1, pp. 187–197, Jan. 2010.

[10] E. Eaton and R. Mansbach, "A spin-glass model for semi-supervised community detection," in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, OnN, Canada, Jul. 2012, pp. 900–906. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5160

[11] Z.-Y. Zhang, "Community structure detection in complex networks with partial background information," *EPL*, vol. 101, no. 4, Feb. 2013, Art. no. 48005.

[12] Z.-Y. Zhang, K.-D. Sun, and S.-Q. Wang, "Enhanced community structure detection in complex networks with partial background information," *Sci. Rep.*, vol. 3, no. 1, p. 3241, Nov. 2013.

[13] F. Nie, H. Wang, H. Huang, and C. H. Q. Ding, "Early active learning via robust representation and structured sparsity," in *Proc. IJCAI 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, Aug. 2013, pp. 1572–1578. [Online]. Available: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6984

[14] C. Li, H. Liu, and D. Cai, "Active learning on manifolds," *Neurocomputing*, vol. 123, pp. 398–405, Jan. 2014, doi: 10.1016/j.neucom.2013.08.002.

[15] M. Leng, Y. Yao, J. Cheng, W. Lv, and X. Chen, "Active semi-supervised community detection algorithm with label propagation," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, Wuhan, China, Apr. 2013, pp. 324–338, doi: 10.1007/978-3-642-37450-0_25.

[16] L. Yang, D. Jin, X. Wang, and X. Cao, "Active link selection for efficient semi-supervised community detection," *Sci. Rep.*, vol. 5, no. 1, Mar. 2015, Art. no. 9039.

[17] Y. Li, C. Jia, J. Li, X. Wang, and J. Yu, "Enhanced semi-supervised community detection with active node and link selection," *Phys. A, Stat. Mech. Appl.*, vol. 510, pp. 219–232, Nov. 2018.

[18] S. Fortunato, "Community detection in graphs," *CoRR*, vol. abs/0906.0612, pp. 75–174, Jun. 2009. [Online]. Available: https://arxiv.org/abs/0906.0612

[19] B. Yang, J. Liu, and D. Liu, "Characterizing and extracting multiplex patterns in complex networks," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 2, pp. 469–481, Apr. 2012, doi: 10.1109/tsmcb.2011.2167751.

[20] T. P. Leal, A. C. Gonçalves, V. D. F. Vieira, and C. R. Xavier, "DECoDe—Differential evolution algorithm for community detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 4635–4640.

[21] L. Yang, X. Cao, D. Jin, X. Wang, and D. Meng, "A unified semi-supervised community detection framework using latent space graph regularization," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2585–2598, Nov. 2015.

[22] L. Fan, S. Xu, D. Liu, and Y. Ru, "Semi-supervised community detection based on distance dynamics," *IEEE Access*, vol. 6, pp. 37261–37271, 2018.

[23] L. Yang, M. Ge, D. Jin, D. He, H. Fu, J. Wang, and X. Cao, "Exploring the roles of cannot-link constraint in community detection via multi-variance mixed Gaussian generative model," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0178029.

[24] S. Liu, C. Ding, F. Jiang, Y. Wang, and B. Yin, "Auto-weighted multi-view learning for semi-supervised graph clustering," *Neurocomputing*, vol. 362, pp. 19–32, Oct. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231219309397

[25] Z. Li, J. Liu, and K. Wu, "A multiobjective evolutionary algorithm based on structural and attribute similarities for community detection in attributed networks," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 1963–1976, Jul. 2018.

[26] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 265–271. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11964

[27] L. Yang, X. Cao, D. He, C. Wang, X. Wang, and W. Zhang, "Modularity based community detection with deep learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, Jul. 2016, pp. 2252–2258. [Online]. Available: http://www.ijcai.org/Abstract/16/321

[28] J. Cao, D. Jin, L. Yang, and J. Dang, "Incorporating network structure with node contents for community detection on large networks using deep learning," *Neurocomputing*, vol. 297, pp. 71–81, Jul. 2018, doi: 10.1016/j.neucom.2018.01.065.

[29] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, "Community discovery using nonnegative matrix factorization," *Data Mining Knowl. Discovery*, vol. 22, no. 3, pp. 493–521, May 2011, doi: 10.1007/s10618-010-0181-y.

[30] L. Yang, Y. Wang, J. Gu, X. Cao, X. Wang, D. Jin, G. Ding, J. Han, and W. Zhang, "Autonomous semantic community detection via adaptively weighted low-rank approximation," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 3s, pp. 1–22, Nov. 2019, doi: 10.1145/3355393.

[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Denver, CO, USA, 2000, pp. 556–562. [Online]. Available: http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization

[32] R. Hochberg, *Matrix Multiplication With CUDA—A Basic Introductionto the CUDA Programming Model*, document 44, 2012.

[33] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002, doi: 10.1073/pnas.122653799.

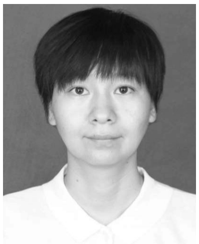
[34] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, 2008, Art. no. 046110.

[35] D. Lusseau and M. E. J. Newman, "Identifying the role that animals play in their social networks," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 271, pp. S477–S481, Dec. 2004.

[36] L. Yang, D. Jin, D. He, H. Fu, X. Cao, and F. Fogelman-Soulie, "Improving the efficiency and effectiveness of community detection via prior-induced equivalent super-network," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 634.

[37] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of Facebook networks," *CoRR*, vol. abs/1102.2166, pp. 583–617, Feb. 2011. [Online]. Available: http://arxiv.org/abs/1102.2166

[38] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.



SUQI ZHANG received the Ph.D. degree from the School of Electronic Information Engineering, Tianjin University, Tianjin, China, in 2014. She is currently working with the School of Information Engineering, Tianjin University of Commerce, Tianjin. Her research interests include intelligent recommendation, complex network analysis, and data mining.

JUNYAN WU received the B.S. degree in software engineering from the Tianjin University of Commerce, Tianjin, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with the Hebei University of Technology, Tianjin. Her current research interests include machine learning and data mining.

JIANXIN LI received the Ph.D. degree in computer science from the Swinburne University of Technology, Australia, in 2009. He is currently an Associate Professor with the School of Info Technology, Deakin University. His research interests include database query processing and optimization, social network analytics, and traffic network data processing.

JUNHUA GU was born in 1966. He is currently working at the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. His main research interests include data mining, intelligent information processing, information acquisition and integration, intelligent computing and optimization, function and information display, and software engineering and project management.

XIANCHAO TANG received the Ph.D. degree from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2017. He is currently working with the China Academy of Electronics and Information Technology, Beijing, China. He received the China Scholarship Council Fellowship, in 2015, and visited the Université Catholique de Louvain, Belgium, as a joint training student from September 2015 to September 2016. His research interests include complex network analysis, machine learning, and data mining.

XINYUN XU received the B.S. degree in communication engineering from Shanghai University, Shanghai, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with the Hebei University of Technology, Tianjin, China. Her current research interests include machine learning and natural language processing.