Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8236157, 12 pages
https://doi.org/10.1155/2022/8236157

Research Article

General Community Detection in Attributed Networks with Consistent-Module Constrained Nonnegative Matrix Factorization

Yafang Li, Yuanda Liu, Jianwen Wei, Baokai Zu, and Hongyuan Wang

Faculty of Information Technology, Beijing University of Technology, Beijing, China

Correspondence should be addressed to Baokai Zu; bzu@bjut.edu.cn

Received 29 April 2022; Revised 10 October 2022; Accepted 17 October 2022; Published 14 November 2022

Academic Editor: Xiangjie Kong

Nonnegative matrix factorization (NMF) model has been successfully applied to discover latent community structures due to its good performance and interpretability advantages in extracting hidden patterns. However, most previous studies explore only the structural information of the network while ignoring the rich attributes. Besides, they aim at detecting densely connected communities (also called community structures) and fail to identify general structures, such as bipartite structures and mixture structures. In this paper, we research on general structure discovery and propose a new method GCDNMF (General Community Detection based on Nonnegative Matrix Factorization), which integrates structural information and node attributes through consistency module constraint to capture the community interactions. It can discover the general community structures of nodes by iteratively updating the community-interaction matrix and the node-membership matrix. We also introduce matrix initialization based on centrality and dispersion of nodes for center selection to reduce the sensitivity of random initialization. Experimental results on real-world networks with a variety of characteristics validate the performance of our approach, especially on networks with general structures. In addition, the associated initialization evaluations demonstrate the effectiveness of our method in obtaining stable results.

1. Introduction

Many complex systems in the real world can be described as networks, such as social networks, transportation networks, and citation networks. Community structure is an essential and common topological property in these networks. The identification of community structure is a fundamental issue in understanding network topology and functional modules, and it has attracted the attention of many researchers [1–9]. A comprehensive review of existing community detection methods can be found in the literature [10]. Furthermore, with the rapid emergence of user-generated media(e.g., Microblog, WeChat, and Twitter), while structural connections between nodes indicate various interdependencies between individuals or organizations [6], real-world networks also contain rich attribute information that characterizes nodes and are referred to as attribute networks. As revealed in previous work, informative node attributes can help to find meaningful groups of users with similar interests, backgrounds, or purposes, which can further effectively support applications in recommendation, sentiment analysis, and user profiling [11]. Moreover, realistic complex networks often contain multiple structures, in addition to the traditional community structure, also known as assortative mixing, i.e., defined as a structure with tight intra-community node links and sparse inter-community links, such as the classical citation network Cora dataset; they also contain multiple complex network structures, such as the bipartite network [12] generated by the English lexical link network Adjnoun, and mixture structures containing both structures, also called disassortative mixing [13]. Mining the various underlying structures and interaction patterns between communities in a network is of great theoretical and practical significance for understanding the function of networks, discovering hidden patterns and predicting the behavior of individuals in the network.

In the past decades, several methods have been proposed to detect communities in attributed networks. They are mainly classified into modularity based methods [14, 15], clustering based methods [16–20], random walk based methods [21, 22], statistical inference models [13, 23, 24], and matrix factorization based methods [3, 25–27]. Among them, nonnegative matrix factorization (NMF) based methods have attracted much interest due to their good performance and strong interpretability. For example, Jia et al. [28] developed a modularized trifactor matrix factorization model Mtrinmf to exploit the topological and the modularity information of the network. Zhang et al. [3] used the NMF method to improve density peak clustering in community detection. However, node attributes are not considered in these models. To introduce node attributes, Jin et al. [29] used the node attribute matrix to construct a NMF framework for underlying community membership. Chen et al. [27] proposed CDCN by combining node attribute information and community structure information using the NMF framework to identify communities with semantic annotation. However, above attributed methods commonly assumed that the community structure obtained from the link structure is consistent with the community structure obtained from node attribute mining. Hence, they embed the structure and attribute into the same space and obtain the common node community matrix. In this way, they typically aim to extract traditional communities that are assortative, i.e. nodes are mostly connected with others in the community. They may overlook intercommunity relationship, making it difficult to exploit the generalized community structures, including assortative communities, disassortative communities, i.e. most connections are from different communities(such as bipartite networks), or the mixed community structures.

To address these issues, in this paper, we research on discovering general structures and propose a new nonnegative matrix factorization model named GCDNMF. It integrates the structural information and node attributes of networks through consistency module constraints to capture the interactions between communities. By iteratively updating the community-interaction matrix and node-membership matrix, it captures the general community structures of nodes. In addition, we initialize the initial matrix by the centrality and dispersion of nodes to reduce the sensitivity caused by random initialization. In summary, the innovation of this paper is threefold:

(1) We propose a novel NMF-based model to detect general structures in attribute networks, which naturally combines structural connections and node attributes into a joint decomposition model. To the best of our knowledge, we are the first to model the general structures using the NMF-based model.

(2) We propose consensus factorization to exploit general communities by studying the consistency between nodes and communities in terms of structural connections and node attributes. It is addressed by alternately updating the community-interaction matrix in the link structure and node-membership matrix in the node attributes.

(3) Extensive experiments are conducted on benchmark networks to demonstrate the effectiveness of our proposed method by comparing it with the state-of-the-art methods. The experimental results show the superior performance of our model in detecting general structures.

The remainder of this paper is organized as follows. Section 2 introduces the related work on community detection based on nonnegative matrix factorization. Section 3 presents our proposed GCDNMF model which integrates the topological information and node attributes based on consistent-model constraints. To verify the performance of our method, several experiments are carried out in Section 4. Section 5 draws the conclusion and gives further consideration.

2. Related Work

Nonnegative matrix factorization (NMF) has good capability in extracting hidden patterns and structures from high-dimensional data. Kumari et al. [30] pioneered a standard community detection method based on NMF, which can effectively mine the structural characteristics of the community by quantifying the link relationships between nodes. Owing to its advantages of simple implementation, innate interpretability, and outstanding performance, it has become a vital technique for community detection [31] and has attracted much attention by researchers to improve the performance of NMF-based community detection.

For directed and undirected networks classified according to the directionality of edges, Kuncheva et al. [32] proposed SNMF and ANMF to extract the intrinsic community structures, respectively. Considering the modularity information of the network, Jia et al. [28] presented a trifactor NMF model that combines the modularity information as a regularization term. To further capture the complex underlying network structure effectively and preserve the global and local structures, Li et al. [33] proposed a multilayer model based on NMF, which consists of an encoder module and a decoder module. Li et al. [34] explored the implicit association between nodes and presented a community detection method based on SNMF. However, the above conventional methods mainly explore the topology of the network to obtain communities.

Some of the most recently developed state-of-the-art methods use both topology and attribute to extract communities. Jin et al. [29] utilized the NMF technique to combine the observed network structure and node attributes, but the model does not focus on factorizing the node attributes matrix and ignores the various implications of edges in forming the community structure. Li and Liu [35] proposed a trifactor nonnegative matrix factorization clustering framework NMTF to combine three types of graph regularization in social networks. This approach utilizes additional content information to detect communities, and fails to

Table 1: Notations utilized in the paper.

Notations	Descriptions
A	$A_{ij} = 1$ if there is an edge between nodes v_i and v_j; 0 otherwise
C	$C_{ij} = 1$ if node v_i has attribute j; 0 otherwise
H	H_{ic} is the propensity of node v_i belonging to community c
G	G_{ij} is the probability of edges between community i and j
Q	Q_{ic} is the probability of node v_i belonging to text cluster c
V	V_{jc} is the propensity of a node in cluster c having attribute j
V	Set of nodes in a network
E	Set of edges in a network
C	Set of attributes in a network
k	Number of communities

explore the relationship between communities and this content. Tang et al. [36] proposed a weighted nonnegative factorization method for attributed graph clustering, which incorporates a weighting scheme to distinguish the importance of attributes. Jin et al. utilized both the community structure matrix and the node attribute matrix in NMF framework SCI [29]. Chen et al. [27] combined node attribute information and community structure information in the NMF framework to accurately find the relationships between networks. Some recent research work focus on building NMF model to learn low-dimensional representation of nodes for discovering communities in attributed networks [33, 37].

3. The GCDNMF Model

In this section, we present the proposed general community detection method GCDNMF, which incorporates topological information and node attribute in a collective NMF-based model. It consists of three main parts: network structure modeling, node attribute modeling, and joint modeling. Next, we describe each part in detail and give the optimization method.

3.1. Problem Formulation. We denote a network as $G = (\mathcal{V}, \mathcal{E}, \mathcal{C})$, where $\mathcal{V}$ is a set of n nodes, $\mathcal{E}$ is the set of edges between nodes, and $\mathcal{C}$ is a set of attribute vectors. The n nodes and their connections are interpreted by an adjacency matrix A, if nodes i and j are connected, the corresponding entry $A_{ij} = 1$, otherwise $A_{ij} = 0$. The attributes of nodes in the network are represented by an attribute matrix $C \in R^{n\times m}$, where m is the dimension of node attributes. Our proposed method aims to partition the network G into k communities by jointly decomposing the adjacency matrix A and the attribute matrix C. In this paper, we summarize the notations and their definitions in Table 1.

3.2. Modeling Network Structures. Community detection refers to find those nodes with relatively close relationship from a network and divide them into different communities. The idea that the parts constitute the whole in the nonnegative matrix factorization provides an effective solution to this problem. To model the topological structure of nodes, we improve the traditional NMF method and propose a three-factor factorization method to decompose the adjacency matrix A. The objective function can be expressed as:

$$\min \left\|A - HGH^T\right\|_F^2 s.t. H \geq 0, G \geq 0, \tag{1}$$

where $H \in R^{n\times k}$ is the community membership matrix, in which H_{ij} indicates the propensity of node i belonging to community j; $G \in R^{k\times k}$ is the community relation matrix and G_{ij} is the probability of edges existing between community i and community j. Intuitively, G is used as a measure of the strength of relationships between communities. Compared with the traditional NMF method, this method adopts a trifactor decomposition instead of a two-factor decomposition. On the one hand, the trifactor NNF model is suitable for both directed and undirected networks. More importantly, it has a clear physical meaning for H and G. The relation matrix G is further combined with node attribute modeling to exploit the generalized community structure.

3.3. Modeling Node Attributes. In attribute networks, nodes and their correlated attributes can be regarded as the relationship between documents and keywords. Using the bag-of-words approach, the attribute matrix is denoted as $C \in R^{n\times m}$, where n represents the number of documents and m is the number of keyword features. Assuming that the documents consists of k clusters, based on NMF text clustering, C can be decomposed into two nonnegative matrices $Q \in R^{n\times k}$ and $V \in R^{m\times k}$. We then have the following objective function related to the node attributes:

$$\min \left\|C - QV^T\right\|_F^2 s.t. Q \geq 0, V \geq 0, \tag{2}$$

where Q is the probability distribution matrix between nodes and communities, Q_{ic} represents the membership degree of node i belonging to community c. V is the probability distribution matrix between node attributes and communities, and V_jc indicates the propensity that community c can be describe by keyword j. In this way, we can divide the

Input: Adjacency matrix A;
Attribute matrix C;
Number of communities k;
Number of iterations *iter*;
Output: Community label l_i for each node i
1: initialize G, Q and V randomly
2: initialize H by K-rank-D
3: for $t = 1 : iter$ do
4: $G = G \otimes ((H^{\mathrm{T}}AH + \beta Q^{\mathrm{T}}Q)/(H^{\mathrm{T}}HGH^{\mathrm{T}}H + \beta G))$
5: $Q = Q \otimes ((\alpha CV + 2\beta QG)/(\alpha QV^{\mathrm{T}}V + 2\beta QQ^{\mathrm{T}}Q))$
6: $V = V \otimes (C^{\mathrm{T}}Q/(VQ^{\mathrm{T}}Q))$
7: $H = H \otimes (AHG^{\mathrm{T}}/(HGH^{\mathrm{T}}HG^{\mathrm{T}}))$
8: end for
9: return $l_i = \mathrm{argmax}_{r \le k} H_{ir}$

Algorithm 1: GCDNMF Algorithm.

Table 2: Details of the real-world network datasets utilized in the experiments, where N, E, M, and K are the number of nodes, edges, attributes, and communities, respectively, where 'AVD', 'CC', and 'ASC' represent the average degree, clustering coefficient, and assortative coefficient.

Networks	N	E	M	K	'AVD'	'CC'	'ASC'
Cora	2708	5429	1433	7	3.89	0.241	-0.065
Citeseer	3312	4732	3703	6	2.73	0.142	0.047
Cornell	195	304	1703	5	2.90	0.157	-0.241
Texas	187	328	1703	5	3.09	0.303	-0.250
Wisconsin	265	530	1703	5	3.18	0.278	-0.179
Washington	230	446	1703	5	3.54	0.197	-0.223

communities by node attributes and obtain attribute community matrix Q, which indicates that nodes in the same attribute community have a large attribute similarity.

3.4. Joint Community Detection Model. In the above two subsections, the community detection results are obtained from two perspectives of structural information and node attributes, respectively. To ensure that the final result is consistent, GCDNMF introduces a consistency module to jointly formulate the above two aspects. Different from traditional methods that focus on embedding structures and attributes into the common node-community space, we also concentrate on the relationship matrix between communities. Intuitively, based on attribute community matrix Q, we can further obtain the matrix describing the relationship between communities in the attribute information by $Q^{\mathrm{T}}Q$. Specifically, each entry in $Q^{\mathrm{T}}Q$ portrays two communities in which nodes have attribute similarity and indicates the propensity of edges existing between two communities based on attribute similarity.

Considering that the topology is consistent with the clustering structure of the node attributes, the structure community relation matrix G and the attribute community relation matrix $Q^{\mathrm{T}}Q$ obtained are inclined to be approximated. Therefore, we derive the following objection function:

$$\min \left\| G - Q^{\mathrm{T}}Q \right\|_F^2 s.t. G \ge 0, Q \ge 0. \tag{3}$$

Then, we propose a weighted joint NMF-based framework to integrate above objectives. Our goal is to minimize the following optimization problem over H, G, Q, and V:

$$\begin{aligned} L = &\left\| A - HGH^{\mathrm{T}} \right\|_F^2 + \alpha \left\| C - QV^{\mathrm{T}} \right\|_F^2 + \beta \left\| G - Q^{\mathrm{T}}Q \right\|_F^2, \\ &s.t. H \ge 0, G \ge 0, Q \ge 0, V \ge 0, \end{aligned} \tag{4}$$

where α and β are positive weights to balance the structure/attribute fusion and the strength of consistency constraint on the community relationship.

3.5. Model Optimization. Since minimizing Eq. (4) with respect to H, G, Q, and V is not convex in all variables together, we utilize an alternating iterative updating scheme to optimize the objective for convergence to a local minimum. First, according to matrix properties $Tr(AB) = Tr(BA)$ and $Tr(A^{\mathrm{T}}) = Tr(A)$, the objective function can be derived to the following form:

$$\begin{aligned} \min_{H,G,Q,V \ge 0} L = &Tr\left(AA^{\mathrm{T}} + HGH^{\mathrm{T}}HG^{\mathrm{T}}H^{\mathrm{T}} - 2AHG^{\mathrm{T}}H^{\mathrm{T}}\right) \\ &+ \alpha Tr\left(CQC^{\mathrm{T}} + QV^{\mathrm{T}}VQ^{\mathrm{T}} - 2CVQ^{\mathrm{T}}\right) \\ &+ \beta Tr\left(GG^{\mathrm{T}} + Q^{\mathrm{T}}QQ^{\mathrm{T}}Q - 2GQ^{\mathrm{T}}Q\right). \end{aligned} \tag{5}$$

To optimize Eq. (5), w.r.t. H, G, Q, and V, four Lagrangian multipliers are introduced, $\lambda_H \in R^{n \times k}$, $\lambda_G \in R^{n \times n}$, $\lambda_Q \in R^{n \times k}$, and $\lambda_V \in R^{m \times k}$. According to Karush-Kuhn-Tucker(KKT) condition, which characterizes the necessary and sufficient condition that the optimal solutions need to satisfy:

$$\begin{aligned} &\lambda_H \otimes H = 0, \quad \lambda_G \otimes G = 0, \\ &\lambda_Q \otimes Q = 0, \quad \lambda_V \otimes V = 0. \end{aligned} \tag{6}$$

where $\otimes$ is the Hadamard product operator(like the operator '.*' in matlab), for example, $[A \otimes B]_{ij} = A_{ij}.*B_{ij}$. Thus, we derive the Lagrange function:

$$\begin{aligned} \min_{H,G,Q,V \ge 0} L = &Tr\left(AA^{\mathrm{T}} + HGH^{\mathrm{T}}HG^{\mathrm{T}}H^{\mathrm{T}} - 2AHG^{\mathrm{T}}H^{\mathrm{T}}\right) \\ &+ \alpha Tr\left(CQC^{\mathrm{T}} + QV^{\mathrm{T}}VQ^{\mathrm{T}} - 2CVQ^{\mathrm{T}}\right) \\ &+ \beta Tr\left(GG^{\mathrm{T}} + Q^{\mathrm{T}}QQ^{\mathrm{T}}Q - 2GQ^{\mathrm{T}}Q\right) \\ &+ Tr\left(\lambda_H H^{\mathrm{T}}\right) + Tr\left(\lambda_G G^{\mathrm{T}}\right) \\ &+ Tr\left(\lambda_Q Q^{\mathrm{T}}\right) + Tr\left(\lambda_V V^{\mathrm{T}}\right). \end{aligned} \tag{7}$$

Setting partial derivatives of G, Q, V, and H to zero, we

have:

$$\begin{aligned}
\frac{\partial L}{\partial G} &= 2H^{\mathrm{T}}HGH^{\mathrm{T}}H - 2H^{\mathrm{T}}AH + 2\beta G - 2\beta Q^{\mathrm{T}}Q + \lambda_G = 0,\\
\frac{\partial L}{\partial Q} &= 2\alpha QV^{\mathrm{T}}V - 2\alpha CV + 4\beta QQ^{\mathrm{T}}Q - 4\beta QG + \lambda_Q = 0,\\
\frac{\partial L}{\partial V} &= 2\alpha VQ^{\mathrm{T}}Q - 2\alpha C^{\mathrm{T}}Q + \lambda_V = 0,\\
\frac{\partial L}{\partial H} &= 4HGH^{\mathrm{T}}HG^{\mathrm{T}} - 4AHG^{\mathrm{T}} + \lambda_H = 0.
\end{aligned} \tag{8}$$

To eliminate Lagrangian multipliers by Eq. (7), we obtain:

$$\begin{aligned}
\left(H^{\mathrm{T}}HGH^{\mathrm{T}}H - H^{\mathrm{T}}AH + \beta G - \beta Q^{\mathrm{T}}Q\right)_{ij} G_{ij} &= 0,\\
\left(\alpha QV^{\mathrm{T}}V - \alpha CV + 2\beta QQ^{\mathrm{T}}Q - 2\beta QG\right)_{ij} Q_{ij} &= 0,\\
\left(VQ^{\mathrm{T}}Q - C^{\mathrm{T}}Q\right)_{ij} V_{ij} &= 0,\\
\left(HGH^{\mathrm{T}}HG^{\mathrm{T}} - AHG^{\mathrm{T}}\right)_{ij} H_{ij} &= 0.
\end{aligned} \tag{9}$$

From Eq. (9), we obtain the following updating formulas:

$$\begin{aligned}
G &= G \otimes \frac{H^{\mathrm{T}}AH + \beta Q^{\mathrm{T}}Q}{H^{\mathrm{T}}HGH^{\mathrm{T}}H + \beta G},\\
Q &= Q \otimes \frac{\alpha CV + 2\beta QG}{\alpha QV^{\mathrm{T}}V + 2\beta QQ^{\mathrm{T}}Q},\\
V &= V \otimes \frac{C^{\mathrm{T}}Q}{VQ^{\mathrm{T}}Q},\\
H &= H \otimes \frac{AHG^{\mathrm{T}}}{HGH^{\mathrm{T}}HG^{\mathrm{T}}}.
\end{aligned} \tag{10}$$

3.6. Initialization of GCDNMF. Since the NMF-based approach is sensitive to random initial values of the variables, to overcome this problem, the node membership matrix H of our model is initialized based on our previously presented work K-rank-D [38], which utilizes the centrality and dispersion of the network to determine the cluster centers.

To formulate the centrality of nodes in the network, according to the algorithm in [38], by modifying the transition probability matrix P, where $P_{ij} = A_{ij}/\sum_j A_{ij}$, we get the centrality of nodes by:

$$\bar{P}_i = P_i + \sum_j \exp\left(-\frac{d_{ij}^2}{P_i}\right), \tag{11}$$

where d_{ij} is the Euclidean distance between node i and node j. The nodes with higher centrality are more likely to be selected as the center points. What is more, as in real networks, the distance between community centers is usually far from each other, to measure the degree of dispersion among centers, a dispersion measurement by computing the distance between node i and other nodes with higher centrality is defined as:

$$\delta_i = \min_{j:\bar{P}_j > \bar{P}_i} d_{ij}. \tag{12}$$

According to the centrality and dispersion of nodes formulated as Eq. (11) and Eq. (12). The CV (comprehensive value) of any node i in the network can be defined as:

$$CV(i) = \frac{\bar{P}_i \delta_i}{\max_{1\le i\le n}\left(\bar{P}_i\right) \max_{1\le i\le n}\left(\delta_i\right)}. \tag{13}$$

We sort the CV values of all nodes in descending order and select the k node with the highest CV value as the center of the network. Then we chose the k columns(corresponding to the selected k centers) in the similarity matrix $S = (A + I)^{\tau}$ to obtain the initialized H matrix, where I is an n-dimensional identity matrix and τ is the step length of signal propagation, and we take this value as $\tau = 3$ in this paper.

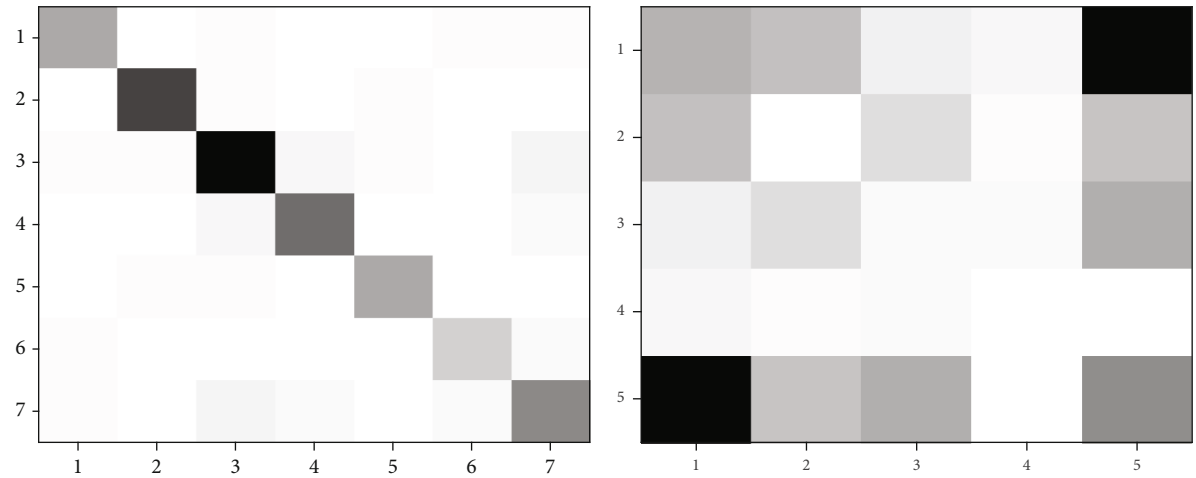
3.7. The GCDNMF Algorithm. Algorithm 1 outlines our proposed GCDNMF. Taking the adjacency matrix A, the attribute matrix C and the number of communities k as input, after initializing the initial matrix, the membership matrix of nodes H is obtained by iteratively update. Finally, the community partition result of nodes is obtained by the maximum assignment. The GCDNMF algorithm is decreasing with step 4 to step 7 and converges to a local optimum. Since $k \ll n$, β, and α are constant, the complexity of updating G, Q, V, and H is $O(T(n^2k) + mnk))$ for T iterations. With well initialization based on centrality and dispersion of nodes, GCDNMF can converge quickly and reduce required iterations largely in partitioning nodes of a network. Generally, 100 iterations will give a promising performance.

4. Experimental Results and Analysis

In this paper, to verify the effectiveness of our proposed method, we compared against the sate-of-the-art community detection method based on NMF model. The average results with 10 trials were recorded. All the algorithms were ran on a PC with RAM:8.0GB, CPU: Intel i7-4600 U, and Platform: MATLAB 2014b.

4.1. Data Description. We used both synthetic and real-world networks to test the effectiveness of our proposed algorithm. The details of these datasets are given below, and the detailed parameters of the data are given in Table 2.

(1) Cora the Cora dataset is a subset of the large Cora citation dataset. It contains 2708 research papers from seven subfields of machine learning. In this network, each node is characterized by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary, which consists of 1703 unique words.



(a) Cora (b) Washington

FIGURE 1: The gray-scale images of community-relation matrices, namely, the block matrices of the networks. Each block represents the link probabilities between the corresponding community pair, and darker colors of the blocks correspond to larger link probability.

(2) Citeseer the Citeseer dataset is a citation network of computer science publications. It contains 3312 publications, each of which is labeled as one of 6 categories. Similar to Cora dataset, each publication is described as one binary vector indicating the presence or absence of the corresponding word from a dictionary of 3703 unique words.

(3) WebKB the WebKB dataset consists of 877 scientific publications and 1608 links, which includes Web page networks of four universities: Cornell, Texas, Washington, and Wisconsin.

According to the block matrices of the networks, the first two datasets are assortative mixing(traditional community structure) as shown in Figure 1(a) taking the Cora dataset as example, while the WebKB datasets is mixture structure, which is neither assortative mixing nor disassortative mixing(e.g.,bipartite structure or multipartitie structure) show in Figure 1(b) taking the Washington dataset as example.

4.2. Evaluation Measurements. In this study, three commonly used metrics are used to measure the performance of an algorithm, accuracy (ACC) [39], normalized mutual information (NMI) [39], and Pairwise F-measure (PWF) [40]. These metrics are defined as follows.

(1) Accuracy (ACC). Given node i, l_{pi} is the assigned label by an algorithm, and l_{ti} is the true label. The accuracy is defined as the fraction of all nodes whose predicted labels are the same with the true labels. The ACC of a particular division of a network is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta\left(l_{ti}, p_{\text{map}}\left(l_{pi}\right)\right)}{n}, \tag{14}$$

where $\delta(x, y)$ is a Kronecker function that the value is 1 if $x = y$, otherwise, 0. $p_{\text{map}}(l_{pi})$ is a permutation mapping function that maps the label l_{pi} of node i to the corresponding label in the ground-truth. n is the overall number of nodes in a network.

(2) Normalized mutual information (NMI). The NMI is defined by:

$$NMI\left(C, C'\right) = \frac{-2\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij} \log\left(n_{ij} n / n_i^C n_j^{C'}\right)}{\left(\sum_{i=1}^{K} n_i^C \log\left(n_i^C / n\right)\right) + \left(\sum_{j=1}^{K'} n_j^{C'} \log\left(n_j^{C'} / n\right)\right)}, \tag{15}$$

where C is the ground-truth cluster label, C' is the computed cluster label, K is the number of communities, n_i^C is the number of nodes in the ground-truth community i, $n_j^{C'}$ is the number of nodes in the computed community j, n_{ij} is the number of nodes in the ground-truth community i that are assigned to the computed community j. In general, the higher NMI, the better result an algorithm get.

(3) Pairwise F-measure (PWF). Let L_T denote the set of nodes having the same community label in the ground-truth, and L_S be the set of nodes in the same community divided by a given algorithm. $|X|$ is the cardinality of X. The balanced PWF is the harmonic mean of precision and recall. It is defined as follows:

$$PWF = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} \times \text{recall}}, \tag{16}$$

where precision $= |L_S \cap L_T| / |L_S|$ and recall $= |L_S \cap L_T| / |L_T|$. The higher the PWF, the closer the division is to the ground-truth.

Table 3: Average comparison results of different methods, the best results are highlighted bold, and the second best are highlighted underline.

Metric	Algorithm	Cora	Citeseer	Cornell	Texas	Wisconsin	Washington
ACC	stdNMF	0.335	0.250	0.338	0.455	0.386	0.433
	ANMF	0.381	0.263	0.324	0.423	0.322	0.374
	CDCN	0.277	0.217	0.422	0.395	0.375	0.500
	Mtrinmf	0.229	0.216	0.270	0.286	0.283	0.272
	SCI	0.344	0.263	0.456	0.565	**0.544**	0.525
	SA-cluster	0.233	0.264	0.415	0.401	0.404	0.491
	GSB	0.335	0.284	0.456	**0.597**	0.536	0.366
	PCL-DC	**0.564**	**0.412**	0.329	0.348	0.336	0.380
	$GCDNMF_{rand}$	0.434	0.233	0.378	0.494	0.388	0.457
	GCDNMF	0.478	0.343	**0.492**	**0.630**	0.476	**0.585**
NMI	stdNMF	0.186	0.050	0.087	0.124	0.075	0.113
	ANMF	0.245	0.066	0.105	0.165	0.069	0.103
	CDCN	0.047	0.058	0.115	0.079	0.084	0.105
	Mtrinmf	0.043	0.012	0.046	0.042	0.037	0.029
	SCI	0.145	0.116	0.079	0.080	**0.169**	0.139
	SA-cluster	0.047	0.117	0.064	0.082	0.101	0.077
	GSB	0.207	0.085	0.122	**0.265**	0.157	0.068
	PCL-DC	**0.416**	**0.170**	0.073	0.061	0.060	0.092
	$GCDNMF_{rand}$	0.287	0.037	0.108	0.182	0.081	0.130
	GCDNMF	0.369	0.151	**0.147**	0.220	0.065	**0.162**
PWF	stdNMF	0.257	0.211	0.309	0.435	0.325	0.397
	ANMF	0.359	0.251	0.283	0.391	0.288	0.327
	CDCN	0.224	0.271	0.376	0.381	0.369	0.490
	Mtrinmf	0.180	0.178	0.240	0.270	0.256	0.250
	SCI	0.300	0.292	0.437	0.545	**0.507**	0.503
	SA-cluster	0.233	0.264	0.415	0.401	0.404	0.491
	GSB	0.258	0.232	0.402	0.570	**0.507**	0.296
	PCL-DC	**0.441**	0.299	0.281	0.316	0.274	0.326
	$GCDNMF_{rand}$	0.372	0.222	0.331	0.478	0.345	0.415
	GCDNMF	0.385	**0.300**	**0.452**	**0.611**	0.407	**0.521**

4.3. Experimental Results and Analysis. In this section, to validate the effectiveness of GCDNMF for community detection, in addition to comparing with structure based nonnegtive matrix factorization methods, such as stdNMF, ANMF, and Mtrinmf, we also compare with the state-of-the-art NMF-based methods that combines structure and attribute, like CDCN and SCI. In addition, we compare GCDNMF with promising attributed clustering method SA-cluster and probabilistic model PCL-DC (focusing on traditional community detection) and GSB (aiming to detect general community detection). What is more, we compare the performance of GCDNMF with different initialization settings, where $GCDNMF_{rand}$ is with random initialization and GCDNMF is with the centrality-based initialization.

The average results of ten random trials are shown in Table 3. From the table, we notice that GCDNMF perform well on most of the datasets and achieves the best and second-best on all metrics, ACC, NMI, and PWF. Furthermore, we have the following observations.

Different from the method only considering topology, such as Mtrinmf, stdNMF, and ANMF, the method of integrating topological connections and node attributes can significantly improve the performance of community detection. The best and second-best results are obtained by methods that integrate two types of information. For instance, the highest ACC score among the methods that focus on network topology information is 0.455(stdNMF), and the best ACC score among the methods that combine both the topology and attribute is 0.565(GSB) on Texas dataset. Compared with those methods, our method GCDNMF achieves 0.630.

Among all the comparison methods of structure and attribute fusion, PCL-DC achieves the best results and GCDNMF is second on Cora and Citeseer datasets. This is in accordance with the original intention of PCL-DC, which is mainly used for mining traditional community structures.

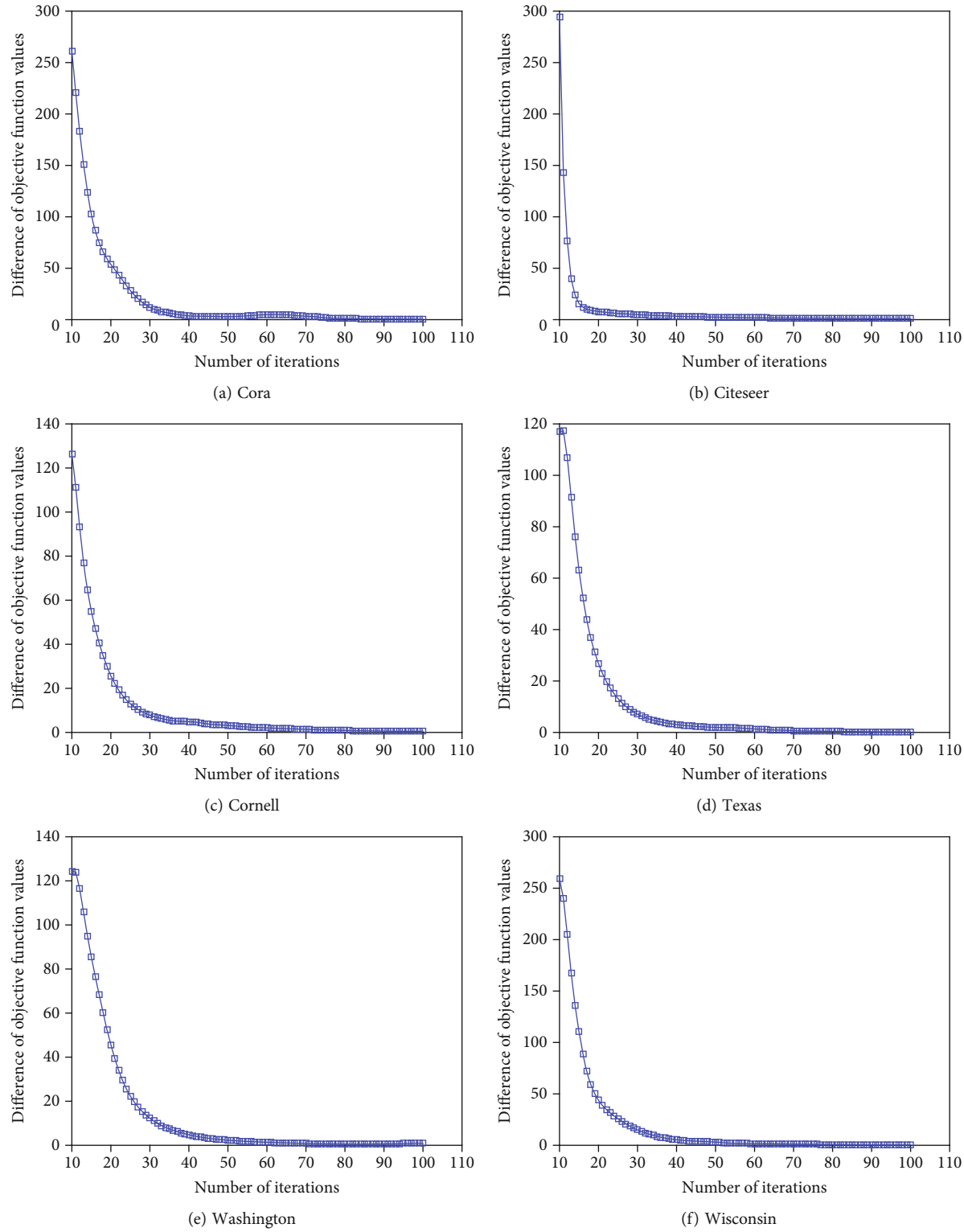
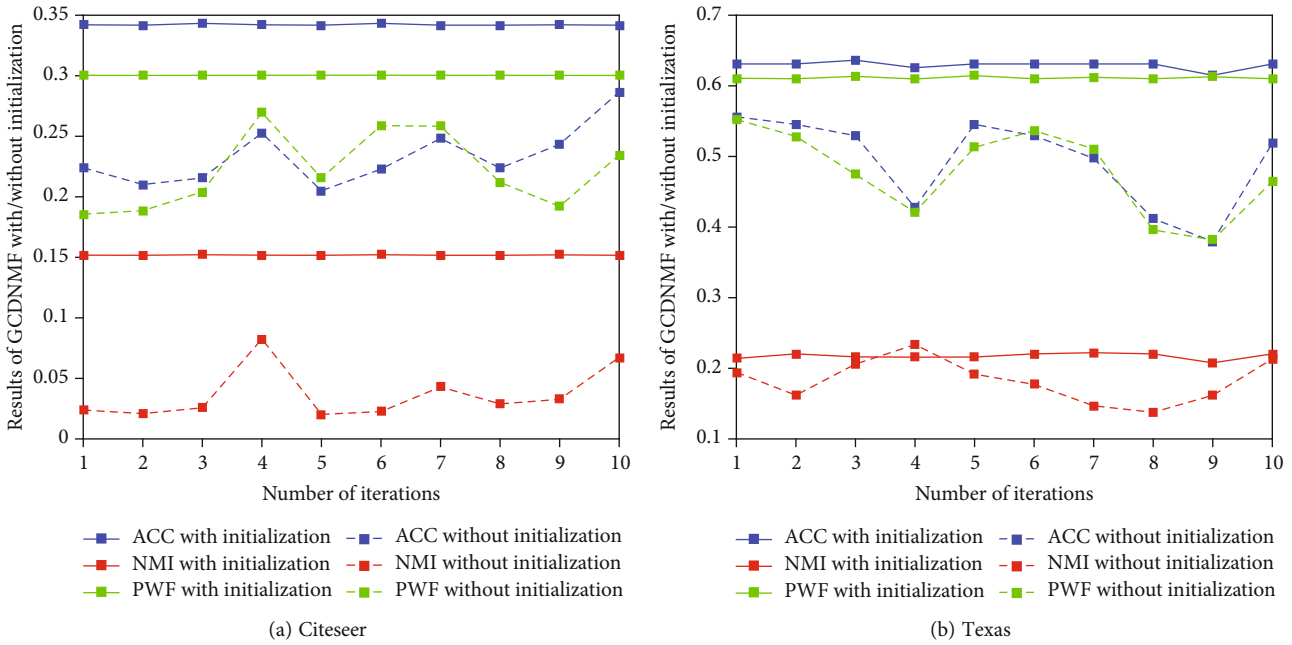


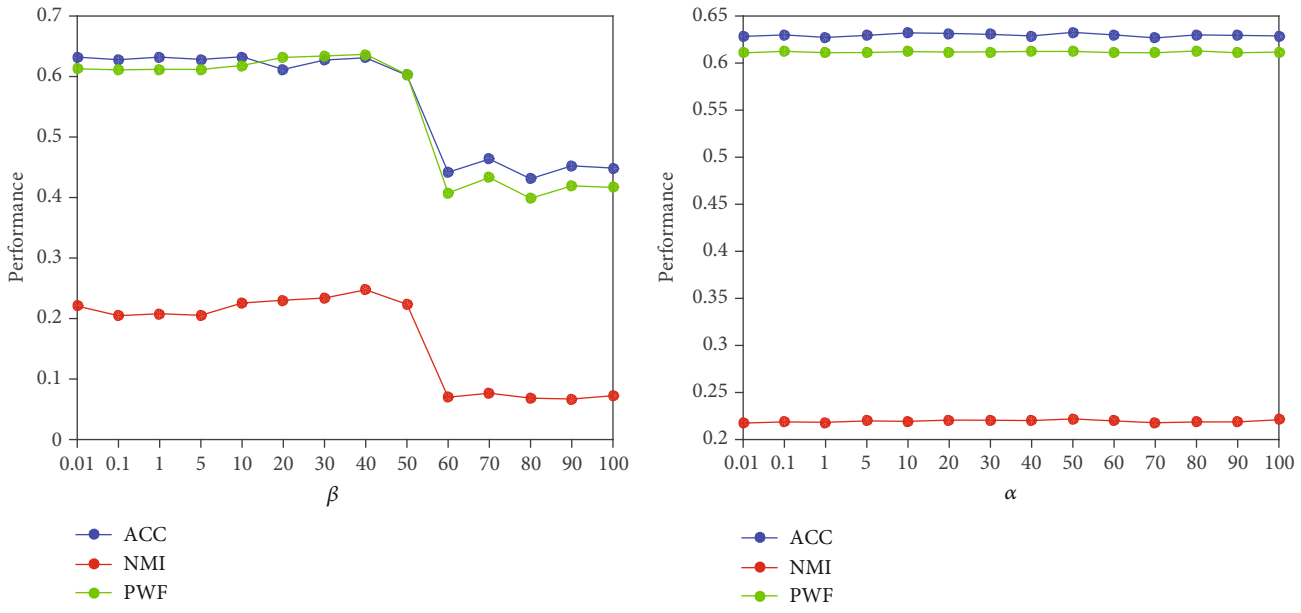
Figure 2: Convergence curve of our proposed GCDNMF. The horizontal axis is the number of iterations, and the vertical axis is the difference of the objective function values. The value of the vertical axis gradually decreases to close to 0, indicating that the algorithm converges.

But on the other four datasets with mixture community structures, methods based on joint matrix decomposition (such as GCDNMF and SCI) and statistical inference model GSB have better performance. Specifically, compared with PCL-DC, GCDNMF improve ACC and PWF by almost 20 % on Washington dataset. For the general community structure discovery, i.e., on the WebKB datasets (including Cornell, Texas, Wisconsin and Washington), overall,



(a) Citeseer (b) Texas

Figure 3: Results of GCDNMF with and without initialization on Citeseer and Texas datasets. The blue, red, and green lines represent the ACC, NMI, and PWF, respectively, where the solid line is the result with our introduced initialization, and the dashed line is the result with random initialization.



(a) β (b) α

Figure 4: The effect of parameter β and α.

GCDNMF and GSB have comparable effects, but GCDNMF achieves the best on three of the four datasets in terms of ACC and PWF. In addition, GCDNMF is able to extract traditional community structures better than GSB, such as on the Cora and Citeseer datasets. This observation implies that GCDNMF can capture more complex community structures, which is shown in that it can not only discover traditional community structures, but also detect mixture structures in networks. This verifies the effectiveness of our proposed GCDNMF by introducing consistency constraints to explore the community-interaction between linkages and attributes.

4.4. Convergence and Stability Study. To solve the proposed joint formulation, we adopt an iterative update technique. In this subsection, we experimentally study the convergence of our proposed GCDNMF. The convergence rate on six

datasets are shown in Figure 2. From these figures, we can see that our proposed GCDNMF converge within 50 iterations on all datasets.

Further, we experimentally validate the effectiveness of the initialization strategy in our work, and we compare the performance of GCDNMF with random initialization. Due to limited space, we take the Citeseer and Texas networks as examples to test the results of GCDNMF with and without initialization. As shown in Figure 3, the blue, red, and green lines, respectively, reflect the results of ACC, NMI, and PWF in ten iterations. The solid line is the results of using initialization, and the dotted line is the result of without initialization. From the figures, it can be noted that the initialization mechanism significantly improves the stability of the results. More specifically, the standard deviation of ten-round results is 0.0006, 0.0002, and 1.07E-4, respectively, in these three metrics with initialization strategy, while the standard deviation without initialization strategy is 0.025, 0.021, and 0.314. Additionally, we find that the accuracy of network community detection is significantly improved through the initialization strategy. Similarly, the same conclusion can be obtained on other networks. In conclusion, the experimental results show that GCDNMF can converge quickly while maintaining high community detection quality compared to random initialization.

4.5. Parameters Analysis. The GCDNMF model has two hyperparameters: α indicates the contribution of the attribute information to the community detection results, and β controls the strength of consistency between community-relation matrices derived from structure and attribute. Because the results of different networks have similar tends, here we demonstrate and analyze the performance effect of hyper-parameters on Texas in experiments. We vary α and β in range of $\{0.01,0.1,1, 10, \cdots, 100\}$ and observe the results while holding the other parameter fixed. Figures 4(a) and 4(b) demonstrate the performance of GCDNMF when $\alpha = 1$ and $\beta = 1$ and ranging the other parameter from 0.01 to 100, respectively. From the figures, we observe that GCDNMF is sensitive to β and is relatively stable with different settings of α when β fixed. Specifically, from Figure 4(a), when β becomes larger, the performance of GCDNMF first keeps rising slightly and then drops sharply at a certain value. This indicates that too large β will introduce noise by excessive consistency constraint from attribute clusters. Therefore we suggest α to be 1 (with equal attention to structure and attribute information) and properly tune β in $\{0.01,0.1, \cdots, 50\}$ so as to achieve a high performance.

5. Summary

In this paper, GCDNMF is proposed for general attribute network community detection by exploring the consistency relationship between node-community structures based on structural connectivity and node attributes. By comparing with several state-of-the-art methods, it is demonstrated that the GCDNMF method has better performance in revealing the general community structure for all benchmarks. In addition, we demonstrate that GCDNMF has stable performance after adopting initialization. However, there is still space for improvements in future work. Interesting issues include the proposed approach to overlapping community detection and semisupervised general community detection.

Data Availability

The datasets used in the manuscript are from the hyperlink: https://linqs-data.soe.ucsc.edu/public/lbc/

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Beijing Science and Technology Planning Project under grant KM202010005015 and the National Natural Science Foundation of China under grant 62006009.

References

[1] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[2] E. J. Mark, "Newman and Elizabeth A. Leicht. Mixture models and exploratory analysis in networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9564–9569, 2007.

[3] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Uncovering fuzzy community structure in complex networks," *Physical Review E*, vol. 76, no. 4, article 046103, 2007.

[4] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 927–935, Paris, France, 2009.

[5] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, "Community discovery using nonnegative matrix factorization," *Data Mining and Knowledge Discovery*, vol. 22, no. 3, pp. 493–521, 2011.

[6] H. Shen, X. Cheng, and J. Guo, "Exploring the structural regularities in networks," *Physical Review E Statistical Nonlinear and Soft Matter Physics*, vol. 84, no. 5, article 056111, 2011.

[7] Y. He, L. Hongtao, L. Huang, and X. Shi, "Non-negative matrix factorization with pairwise constraints and graph laplacian," *Neural Processing Letters*, vol. 42, no. 1, pp. 167–185, 2015.

[8] Y. Pei, N. Chakraborty, and K. P. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *Twenty-fourth international joint conference on artificial intelligence*, pp. 2083–2089, Buenos Aires, Argentina, 2015.

[9] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, pp. 265–271, 2016.

[10] Z. Huang, Y. Ye, X. Li, L. Feng, and H. Chen, "Joint weighted nonnegative matrix factorization for mining attributed graphs," in *Advances in Knowledge Discovery and Data*

Mining, J. Kim, K. Shim, L. Cao, J. G. Lee, X. Lin, and Y. S. Moon, Eds., vol. 10234 of Lecture Notes in Computer Science, pp. 368–380, Springer, Cham, 2017.

[11] L. Yang, L. Zhang, Z. Pan, G. Hu, and Y. Zhang, "Community detection based on co-regularized nonnegative matrix tri-factorization in multi-view social networks," in *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing*, Shanghai, China, 2018.

[12] H. Zhang, X. Niu, I. King, and M. R. Lyu, "Overlapping community detection with preference and locality information: a non-negative matrix factorization approach," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–14, 2018.

[13] C. Yan and Z. Chang, "Modularized tri-factor nonnegative matrix factorization for community detection enhancement," *Physica A: Statistical Mechanics and its Applications*, vol. 533, p. 122050, 2019.

[14] L. Hong, Q. Zhao, X. Sang, and L. Jianfeng, "Community detection in complex networks using nonnegative matrix factorization and density-based clustering algorithm," *Neural Processing Letters*, vol. 51, no. 2, pp. 1731–1748, 2020.

[15] L. Hong, X. Sang, Q. Zhao, and L. Jianfeng, "Community detection algorithm based on nonnegative matrix factorization and pairwise constraints," *Physica A: Statistical Mechanics and its Applications*, vol. 545, article 123491, 2020.

[16] X. Luo, Z. Liu, M. Shang, and M. Zhou, "Highly-accurate community detection via pointwise mutual information-incorporated symmetric non-negative matrix factorization," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 463–476, 2021.

[17] M. Zhang and Z. Zhou, "Structural deep nonnegative matrix factorization for community detection," *Applied Soft Computing*, vol. 97, article 106846, 2020.

[18] Z. Ye, H. Zhang, L. Feng, and Z. Shan, "CDCN: a new nmf-based community detection method with community structures and node attributes," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5517204, 12 pages, 2021.

[19] A. Hintze and C. Adami, "Modularity and anti-modularity in networks with arbitrary degree distribution," *Biology Direct*, vol. 5, no. 32, 2010.

[20] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, article 016107, 2011.

[21] R. K. Behera, D. Naik, B. Sahoo, and S. K. Rath, "Centrality approach for community detection in large scale network," in *Proceedings of the 9th Annual ACM India Conference*, pp. 115–124, Gandhinagar, India, 2016.

[22] R. K. Behera and S. K. Rath, "An efficient modularity based algorithm for community detection in social network," in *2016 International Conference on Internet of Things and Applications (IOTA)*, pp. 162–167, Pune, India, 2016.

[23] R. K. Behera, D. Naik, S. K. Rath, and R. Dharavath, "Genetic algorithm-based community detection in large-scale social networks," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9649–9665, 2020.

[24] K. Berahmand, S. Haghani, M. Rostami, and Y. Li, "A new attributed graph clustering by using label propagation in complex networks," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 5, 2020.

[25] K. Berahmand, E. Nasiri, and Y. Li, "Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding," *Computers in Biology and Medicine*, vol. 138, article 104933, 2021.

[26] K. Berahmand, M. Mohammadi, A. Faroughi, and R. P. Mohammadiani, "A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix," *Cluster Computing*, vol. 25, no. 2, pp. 869–888, 2022.

[27] Y. Chen, L. Wang, D. Qi, T. Ma, and W. Zhang, "Community detection based on deepwalk model in large-scale networks," *Security and Communication Networks*, vol. 2020, Article ID 8845942, 13 pages, 2020.

[28] J. Jia, P. Liu, D. Xiaojin, and Y. Zhang, "Multilayer social network overlapping community detection algorithm based on trust relationship," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 9268039, 14 pages, 2021.

[29] D. Jin, Y. Zhizhi, P. Jiao et al., "A survey of community detection approaches: from statistical modeling to deep learning," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2021.

[30] A. Kumari, R. K. Behera, A. S. Shukla, S. P. Sahoo, S. Misra, and S. K. Rath, "Quantifying influential communities in granular social networks using fuzzy theory," in *Computational Science and Its Applications – ICCSA 2020. ICCSA 2020*, vol. 12252 of Lecture Notes in Computer Science, Springer, Cham, 2020.

[31] A. Kumari, R. K. Behera, K. S. Sahoo, A. Nayyar, A. K. Luhach, and S. P. Sahoo, "Supervised link prediction using structured-based feature extraction in social network," *Concurrency and Computation: practice and Experience*, vol. 34, no. 13, article e5839, 2022.

[32] Z. Kuncheva and G. Montana, "Community detection in multiplex networks using locally adaptive random walks," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1308–1315, Paris, France, 2015.

[33] Y. Li, C. Jia, and Y. Jian, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 321–334, 2015.

[34] Y. Li, C. Jia, X. Kong, L. Yang, and Y. Jian, "Locally weighted fusion of structural and attribute information in graph clustering," *IEEE Transactions on Cybernetics*, vol. 49, no. 1, pp. 247–260, 2019.

[35] M. Li and J. Liu, "A link clustering based memetic algorithm for overlapping community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 410–423, 2018.

[36] F. Tang, C. Wang, S. Jinxia, and Y. Wang, "Spectral clustering-based community detection using graph distance and node attributes," *Computational Statistics*, vol. 35, no. 1, pp. 69–94, 2020.

[37] Y. Li, C. Sha, X. Huang, and Y. Zhang, "Community detection in attributed graphs: an embedding approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 338–345, 2018.

[38] J. Zhang, J. Fei, X. Song, and J. Feng, "An improved louvain algorithm for community detection," *Mathematical Problems in Engineering*, vol. 2021, Article ID 1485592, 14 pages, 2021.

[39] X. Zhou, K. Yang, Y. Xie, C. Yang, and T. Huang, "A novel modularity-based discrete state transition algorithm for community detection in networks," *Neurocomputing*, vol. 334, pp. 89–99, 2019.

[40] J. Zhu, B. Chen, and Y. Zeng, "Community detection based on modularity and k-plexes," *Information Sciences*, vol. 513, pp. 127–142, 2020.