Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5517204, 12 pages
https://doi.org/10.1155/2021/5517204

Research Article

CDCN: A New NMF-Based Community Detection Method with Community Structures and Node Attributes

Zhiwen Ye,[1] Hui Zhang,[1,2] Libo Feng,[3,4] and Zhangming Shan[1]

[1]*State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China*
[2]*Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China*
[3]*Engineering Research Center of Cyberspace, Yunnan University, Kunming 650500, China*
[4]*School of Software, Yunnan University, Kunming 650500, China*

Correspondence should be addressed to Libo Feng; fenglibo@buaa.edu.cn

Received 27 January 2021; Revised 15 April 2021; Accepted 8 May 2021; Published 26 May 2021

Academic Editor: Yuanlong Cao

Community discovery can discover the community structure in a network, and it provides consumers with personalized services and information pushing. It plays an important role in promoting the intelligence of the network society. Most community networks have a community structure whose vertices are gathered into groups which is significant for network data mining and identification. Existing community detection methods explore the original network topology, but they do not make the full use of the inherent semantic information on nodes, e.g., node attributes. To solve the problem, we explore networks by considering both the original network topology and inherent community structures. In this paper, we propose a novel nonnegative matrix factorization (NMF) model that is divided into two parts, the community structure matrix and the node attribute matrix, and we present a matrix updating method to deal with the nonnegative matrix factorization optimization problem. NMF can achieve large-scale multidimensional data reduction processing to discover the internal relationships between networks and find the degree of network association. The community structure matrix that we proposed provides more information about the network structure by considering the relationships between nodes that connect directly or share similar neighboring nodes. The use of node attributes provides a semantic interpretation for the community structure. We conduct experiments on attributed graph datasets with overlapping and nonoverlapping communities. The results of the experiments show that the performances of the F1-Score and Jaccard-Similarity in the overlapping community and the performances of normalized mutual information (NMI) and accuracy (AC) in the nonoverlapping community are significantly improved. Our proposed model achieves significant improvements in terms of its accuracy and relevance compared with the state-of-the-art approaches.

1. Introduction

The science of networks is a modern discipline spanning the natural, social, and computer sciences, as well as engineering. There are different kinds of networks in the real world, such as citation networks, social networks, and collaboration networks [1]. Community detection algorithms are important methods for analyzing the networks' structure and understanding the node semantics, which play important roles in the era of network intelligence [2–5]. First, analyzing the community structure of the network is helpful for people to study the composition and evolution of the whole network, and it can better explain the intrinsic characteristics and causes of the network. Furthermore, community detection algorithms are the key to understanding complex network systems and have important applications in different networks in various fields. For example, they are very useful for social networks to recommend friends and groups to users by analyzing the inherent structure and characteristics of their social network and clustering the user nodes. Community detection is very useful in the application of disease spread analysis, e.g., it can be used in reality epidemic spreading [6].

Most existing community detection algorithms analyze the network by using the original network topology information [7–15]. Girvan and Newman [7] analyzed the community structures in social and biological networks. Newman and Leicht [8] analyzed the mixture models of networks. Rosvall and Bergstrom [9] revealed the community structure of complex networks utilizing the maps of random walks. Xie et al. [1] proposed a method for uncovering overlapping communities in social networks named SLPA via a speaker-listener interaction dynamic process. Coscia et al. [12] proposed a local first discovery method for overlapping communities. He [13] utilized the Markov random field approach for community detection in a specific network. Clauset et al. [14] and Li et al. [15] analyzed the community structure in large scale networks. Cui and Wang [16] used the key bicommunity and intimate degree uncover the overlapping community structures in bipartite networks.

However, for some networks, there are not only network topology information but also node attribute information that is a semantic interpretation of the community structure. For example, papers in citation networks contain titles, abstracts, and keywords that may be represented using binary-valued vectors. We binarize the categorical input so that they can be thought of as vectors in Euclidean space (we call this embedding the vector in Euclidean space).

Such networks with node attributes are named attributed graphs [17, 18]. It is a great challenge to discover the community structure with node attributes in an effective way. To characterize a community, the existing community detection methods mainly rely on the original network topology. The missing and meaningless information in the network topology often leads to poor results. The node attributes of a network may carry essential community information that is complementary to the network topology information. Therefore, even though two nodes are not directly connected, they may belong to the same community according to the node attributes. Several algorithms that consider both structural and attribute information have been proposed in [17, 19–22]. Yang and Leskovec [17] used the nonnegative matrix factorization approach to find the overlapping communities in large scale networks. Atzmueller et al. [19] proposed an exhaustive subgroup discovery method for description-oriented community detection. Wang et al. [20] proposed a semantic community identification method to find the community structures in large attribute networks. Huang et al. [21] analyzed the attributes of community networks. Yang et al. [22] proposed a discriminative approach that combined links and content for community detection. However, all of those methods assign each edge of the attribute graphs the same value. This will lose information about the network. For example, edges that form densely connected subgraphs are much more likely to be in the same community than edges that connect separate subgraphs. Thus, utilizing the original network topology directly causes indiscriminate penalizing of node pairs, whether they are in densely connected structures or not. It means that we should assign the different characteristic nodes various values.

Nonnegative matrix factorization (NMF) is an effective method in community detection. Some scholars have studied it. Luo et al. [23] proposed a symmetric NMF method via pointwise mutual information-incorporated that has highly accurate. Lu et al. [24] used the NMF method to improve density peak clustering in community detection. Zhang and Zhou [25] studied the structure of deep NMF in community detection. Wang et al. [26] used the constraint NMF to detect the community in dynamic networks. These studies laid the foundation for our experiments.

The state-of-the-art methods (e.g., CDE [27]) use both a community structure matrix and a node attribute matrix in the NMF framework, and CDE also considers the densely connected subgraphs. However, CDE only considers the relationship between two nodes directly connected while they analyze the community structure matrix. This behavior will lose information about the community structure.

Scholars have proposed some important methods for large-scale community detection such as neighborhood, maximal subgraph, intimate degree, and core-vertices [28]; these studies provide important ideas for our paper. More importantly, the main contributions of this paper are as follows.

(1) We propose a novel method that generates the community structure matrix, which retains the relationship between two nodes that are directly connected or share the same neighbors

(2) We combine node attribute information and community structure information in an effective way. Then, we propose our method, named Community Detection with Community Structure and Node Attributes (CDCN), to identify the network communities with semantic annotation and community structures using nonnegative matrix factorization framework [29, 30]

(3) Extensive experiments were conducted on public datasets to demonstrate the effectiveness of CDCN, and its accuracy and performance were better than those of the state-of-the-art methods

The remainder of the paper is organized as follows. Section 2 briefly summarizes the three different types of community detection models. Section 3 describes the community detection model, deduces its theory and formula, and describes the solution algorithm for our model. Section 4 conducts extensive comparative experiments to evaluate the effectiveness of our proposed CDCN model on real graph datasets with the ground-truth communities delineated. Section 5 presents the conclusions of this paper and discusses future research directions.

2. Related Works

This section will briefly summarize the three different types of community detection models that use different information to determine the network information. We briefly summarize the three types of community detection models in Table 1.

The first type of community detection method focuses on the original network topology. GN [7] was built around the idea of using centrality indices to find community boundaries. NMM [8] was used to find the mechanism of

Table 1: A brief summarization of community detection.

Types	Representative works
Original network topology	GN [7], NMM [8], InfoMap [9], CPM [10], SLPA [11], DEMON [12], SNMF [31], NetMRF [13], SBM [32], and SPAEM [33]
Node attributes	CAN [34], SMR [35], and NC [36]
Both original network topology and node attributes	COMODO [19], PCL-DC [22], SCI [20], BIGCLAM [17], and CDE [27]

probabilistic mixture models and the expectation maximization algorithm to understand the structure of networks. CPM [10] is a clique percolation method, and it consists of two steps. The first step is to construct the vertices of the k-clique graph, and the second is to find the connected components and set-union vertices within each connected component to get a new community. SLPA [11] was presented as general framework for detecting and analyzing both individual overlapping nodes and entire communities. In it, the nodes exchange labels according to dynamic interaction rules. The stochastic block model (SBM [32]) is the simplest node-based community detection model. The nodes of the network randomly fall into K communities, which are denoted as $z_i \in \{1, 2, \cdots, k\}$, and the edges are independently generated at a probability $w_{z_i z_j}$. McDaid et al. [37] improved Bayesian inference for the stochastic block model for large networks.

However, the above community detection methods directly utilized the original network topology and ignore the inherent community structures (e.g., node attributes). The missing and meaningless information in the network topology often leads to poor results. Therefore, the second type of method focuses on node attributes, and it includes some classical or state-of-the-art clustering methods. Strictly speaking, those methods are not community methods, but they could use node attributes information to discover communities. Thus, in this paper, we also regard them as related work. CAN [34] was proposed as a clustering model to learn the data similarity matrix by assigning the adaptive and optimal neighbors for each data point based on the local distances. SMR [35] uses for the kernelized random walks on the global KNN graph and the Smooth Representation Clustering to improve the clustering result. NC [5, 36] uses the eigenvectors of the matrix representations of the network to solve the community detection problem.

The third type of community detection method considers both the original network topology and node attribute information. Several algorithms that consider both structural and attribute information have been proposed in [17, 19–22]. However, all of those methods assign each edge of the attributed graph the same value. This will lose information about the network. The state-of-the-art method SCI [20] uses both the community structure matrix and the node attribute matrix in NMF framework, and CDE [27] encodes the inherent community structures for community detection via the underlying community memberships. However, they only considered the relationship between two directly connected nodes when they analyze the community structure matrix.

Nonnegative matrix factorization (NMF) [38] is an effective means of data dimensionality reduction. It can discover the hidden information and the relationship between multidimensional data and lay the foundation for data mining and knowledge discovery. At present, NMF has been widely used in data mining [39], image retrieval [40], community discovery [41], hotspot prediction [42], social network privacy protection [43], signal processing [44], and other fields. In terms of community discovery [45, 46], NMF can find the associations between networks based on network node attributes, which is an important community detection method [47]. Many scholars have studied the application of NMF in community detection and provided some ideas for the research of our paper.

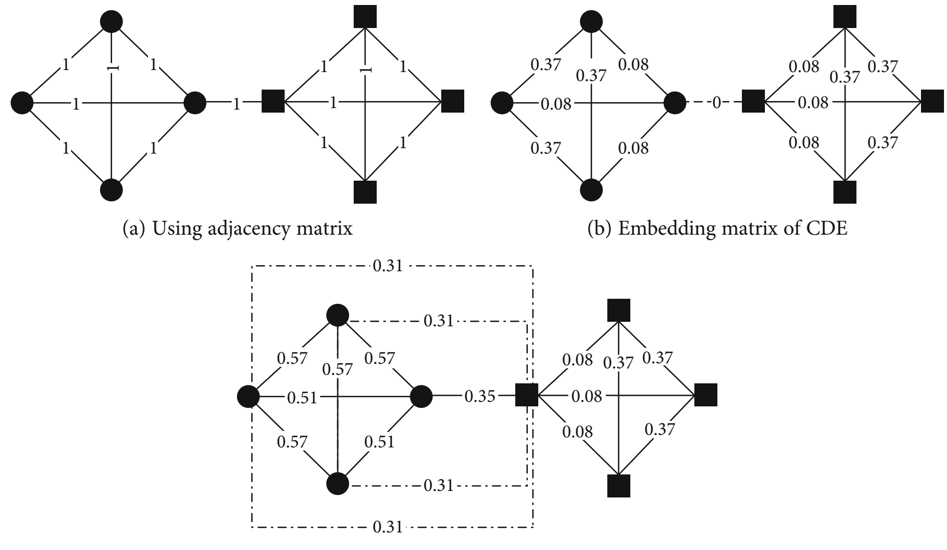
Different from all those methods, we combine node attribute information and community structure information by generating the community structure matrix, which retains the relationship between two directly connected nodes or nodes that share the same neighbors. This method can more accurately find the relationships between networks. The topology diagram of the three methods can be seen in Figure 1.

Figure 1 shows an unweighted graph with two communities, where the different shapes stand for the nodes of different communities. Figure 1(a) is the result of using the adjacency matrix directly, and Figure 1(b) is the community structure embedding matrix of the CDE model. They both focus on the relationship between two directly connected nodes. Figure 1(c) shows the community structure matrix of our models, where the dotted lines are the correlations of the nodes that share the same neighbors but are not directly connected. In other words, compared with existing models, we can not only express the node relationship using a continuous numerical value, but we can also describe the relationship between nonadjacent nodes. Otherwise, the isolated nodes (cold nodes) have no neighbor nodes. It increases the scale of the system and can be ignored when constructing the adjacent matrix. In this paper, the influence of the isolated node on the system is not considered.

3. CDCN: The Community Detection Model

In this section, we propose a novel algorithm for community detection that combines the community structure and node attribute information. We will introduce the community structure, node attributes, the overall model, and the algorithm in detail.

3.1. Community Structure Part. The community structure part models the network structure. Given an undirected network $G = (P, E)$ with n nodes P and e edges E, we could get a binary-valued adjacency matrix A from G. If node i and node



(a) Using adjacency matrix

(b) Embedding matrix of CDE

(c) Community structure matrix of our models

Figure 1: The community structure matrix of different methods.

j have a direct connection, the value of A_{ij} is 1, and otherwise, it is 0, where $i = (1, 2, \cdots, n)$ and $j = (1, 2, \cdots, n)$.

As one of the state-of-the-art methods, SCI directly regards the binary-valued adjacency matrix A as community structure matrix. It will degrade the effectiveness of the community detection model without embedding the adjacency matrix due to the sparsity of A. The CDE model is proposed as a novel community structure embedding method to quantify the structural closeness of nodes to offer a good depiction of the inherent community structures in graphs. Though CDE solved the problem of the sparsity of the adjacency matrix, it still has limitations in that CDE only considers the relationship between two directly connected nodes when it analyze the network structure. Both SCI and CDE will lose the relationship information between nodes that are not directly connected.

We start with a concise and reasonable observation regarding whether two nodes belong to the same community. They may be surrounded by the similar environment that means the two nodes may share the similar neighbor nodes.

Therefore, we measure the similarity in a community memberships as follows:

$$\text{Similarity}(i, j) = \log\left(\frac{P(i, j)}{P(i)P(j)}\right), \tag{1}$$

where $P(i, j)$ can be expressed by $\text{Jaccard}(i, j)/D$, and $\text{Jaccard}(i, j)$ is the Jaccard index value if node i and node j. $P(i) = d_i/D$, where $d(i) = \sum_{j=1}^{J} A_{i,j}$ is the degree of node i and $D = \sum_{i=1}^{I} d_i$ is the total degree for network G. The Jaccard index value can be expressed as follows:

$$\text{Jaccard}(i, j) = \frac{|N(i)| \cap |N(j)\,|}{|N(i)| \cup |N(j)\,|}, \tag{2}$$

where $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, n$; and $N(i)$ is the neighboring nodes of node i.

According to the similarity of the node community member ships, we could get a community structure matrix $S \in R^{n\times n}$, where $S_{i,j} = \text{Similarity}(i, j)$.

In our daily lives, if two people like the same movie, can we think that they share similar hobbies? That may be right but insufficient if the movie is popular and everyone would like it. Thus, if the movie is unfashionable, we could be sure that the two people share similar hobbies. Obviously, this is also suitable for community detection. If two nodes have a same cold neighboring node (few nodes are connected to it), it will make a great contribution to the similarity between the two nodes. In other words, we should add a penalty to the hot neighboring nodes (many nodes are connected to it). By replacing $\text{Jaccard}(i, j)$ with $L(i, j)$, $P(i, j)$ can be expressed as $L(i, j)/D$, $L(i, j) = (\sum_{p=1}^{P} 1/(1 + N(t_p)))/|N(i)\,| \cup |N(j)|$, where P is the count of $|N(i)\,| \cap |N(j)|$, $t_p \in |N(i)\,| \cap |N(j)|$, and $p = (1, 2, \cdots, P)$; $N(t_p)$ is the count of t_p. If $N(t_p) = 0$, Jaccard (i, j) and $L(i, j)$ will be the same.

Figure 1(a) illustrates an unweighted graph with two communities. It uses the SCI method. The edge value is 1 if two nodes are directly connected, and it cannot describe which edges are more important when detecting two communities' structure. Figure 1(b) is the result of the embedding method of CDE. It can assign more weights to the edges that form densely connected subgraphs while assigning less weight to the connection between two communities. Figure 1(c) is the display of our proposed community structure matrix. We show more information about the network structure by considering the relationships between nodes that are directly connected or share similar neighboring nodes.

We define $U \in R^{N\times K}$ as the probability distribution matrix between nodes and communities. U_{ij} stands for the propensity of node i belonging to community j, where $i = (1, 2, \cdots, N)$ and $j = (1, 2, \cdots, K)$. The community structure matrix $S \in R\, n \times n$ can be approximately decomposed into the multiplication between the probability distribution matrix U and its transpose. For a formal model, the following holds:

$$\min_{U \geq 0} \left\| S - UU^T \right\|_F^2, \tag{3}$$

where $S_{i,j} = \sum_{k=1}^{K} U_{i,k} U_{k,j}^T$. The process implies that if node i and node j have similar community memberships, they have a high similarity.

3.2. Community Node Structure Part. We define $T \in R^{N \times F}$ as the node attribute matrix, where N is the count of the nodes in network and F is the feature dimension of a node. The attributes of a node are in the form of an N -dimensional binary-valued vector, and T_{i*} represents the vector of node i, where $i = (1, 2, \cdots, N)$. The node attribute function is as follows:

$$\min_{U \geq 0, M \geq 0} \| T - UM \|_F^2, \tag{4}$$

where $M \in R^{K \times F}$ and$T_{i,j} = \sum_{k=1}^{K} U_{i,k} M_{k,j}$.

The node attribute matrix T is decomposed into two matrixes, U and M. As mentioned above, U is the probability distribution matrix between nodes and communities, $U_{i,k}$ stands for the propensity of node i belonging to community k, M is the probability distribution matrix between nodes feature and communities, and $M_{k,j}$ is the weight of the j-th node attribute feature for community k.

In this way, we can use the node attribute information to divide the communities.

3.3. The CDCN Model. In this subsection, we will elaborate the overall model of our CDCN method. There are two parts of our method, which include the community structure part and the node attribute part.

We combine the community structure part in equation (3) and the node attribute part in equation (4) together. Therefore, our proposed model is written as follows:

$$\min_{U \geq 0, M \geq 0} \left\| S - UU^T \right\|_F^2 + \alpha \| T - UM \|_F^2, \tag{5}$$

where α is a positive parameter that adjusts the weight for the two items. If $\alpha > 1$, it means that we are more inclined to the node attribute information; in contrast, we are more inclined to the network topology information.

To make the model more generalized, we do not strictly restrict the symmetric decomposition of S. In regard to the original model in equation (3), it can be equivalently converted into the following:

$$\min_{U \geq 0, V \geq 0} \left\| S - UV^T \right\|_F^2, s.t. U = V, \tag{6}$$

where V is a surrogate variable for U. For further relaxation, equation (5) can be changed to the following:

$$\begin{aligned} L(U, V, M) = \min_{U \geq 0, V \geq 0, M \geq 0} & \left\| S - UV^T \right\|_F^2 \\ & + \alpha \| T - UM \|_F^2 + \beta \| U - V \|_F^2, \end{aligned} \tag{7}$$

where β is a positive parameter to adjust the closeness between U and V. The higher it is, the closer the two variables are.

In practice, it is common to set β to a moderate value for real applications. In this way, we can enhance the generalization ability of the model. This means that we do not need to limit the community structure matrix S to be decomposed into the same matrix. In other words, we relax the matrix decomposition condition.

In summary, we use the same variable U to combine the two parts, the community structure matrix and the node attribute matrix, and to make the model more generalized, we do not strictly restricted the symmetric decomposition of S. In this way, we get the optimization function of the model.

As for the detection of overlapping communities, we identify that node v_i belongs to the j-th community when U_{ij} is higher than a predefined threshold ε. Following CDE, we set $\varepsilon = 0.1$ in our paper.

3.4. The Algorithm for CDCN. In this subsection, we will share the solution algorithm for our proposed model. The learning process algorithm for CDCN can be seen in Algorithm 1.

The matrix U_{mk},V_{mk}, and M_{kf} in Algorithm 1 can be derived from the following derivation. According to equation (7), take the derivatives of $L(U, V, M)$ with respect to U, V, and M, we can get the formulas (8), (9), and (10), respectively:

$$\begin{aligned} \frac{\partial L(U, V, M)}{\partial U} = & \, 2UV^T V - 2SV + \alpha \left(2U^T MM^T - 2TM^T \right) \\ & + \beta (2U - 2V), \end{aligned} \tag{8}$$

$$\frac{\partial L(U, V, M)}{\partial V} = 2VU^T U - 2SU + \beta (2V - 2U), \tag{9}$$

$$\frac{\partial L(U, V, M)}{\partial M} = \alpha \left(2U^T UM - 2U^T T \right). \tag{10}$$

Based on this, the updating rules for the variables are given as follows:

$$U_{mk} \longleftarrow U_{mk} - \rho_{mk} \frac{\partial L(U, V, M)}{\partial U_{mk}}, \tag{11}$$

$$V_{mk} \longleftarrow V_{mk} - \theta_{mk} \frac{\partial L(U, V, M)}{\partial V_{mk}}, \tag{12}$$

$$M_{kf} \longleftarrow M_{kf} - \phi_{kf} \frac{\partial L(U, V, M)}{\partial M_{kf}}, \tag{13}$$

where ρ_{mk}, θ_{mk}, and ϕ_{kf} denote the step sizes for the (mk)th element of matrix U, the (mk)th element of matrix V and the

```
**Input:**network graph $G$, node attributes matrix $T$, hyper-parameters $\alpha$ and $\beta$, number of communities K and maximum number of iterations *maxIter*.
**Output:**the probability distribution matrix U.
Begin
According to network graph $G$ and Eq.(1), generate the community structure matrix and randomly initialize the probability distribution matrix
 $U^{(0)} \sim (0,1)^{N\times K}$, $V^{(0)} \sim (0,1)^{N\times K}$
        $M^{(0)} \sim (0,1)^{K\times F}$ ; $i = 0$
**While**$i \le maxIter$**do**
$U_{mk} \longleftarrow U_{mk}([SV + \alpha TM^T + \beta V]/[UV^TV + \alpha U^TMM^T + \beta U])$;
$V_{mk} \longleftarrow V_{mk}([SU + \beta U]/[VU^TU + \beta V]))$;
$M_{kf} \longleftarrow M_{kf}([U^TT]/[U^TUM])$;
End While.
End
```

Algorithm 1: The learning process: CDCN.

(kf)th element of matrix M, respectively, in the gradient descent methods. If we set $\rho_{mk} = U_{mk}/2(UV^TV + \alpha U^TMM^T + \beta U)$, $\theta_{mk} = V_{mk}/2(VU^TU + \beta V)$, and $\phi_{kf} = M_{kf}/2\alpha U^TU M$, then the following holds:

$$U_{mk} \longleftarrow U_{mk}\frac{\left[SV + \alpha TM^T + \beta V\right]}{\left[UV^TV + \alpha U^TMM^T + \beta U\right]}, \tag{14}$$

$$V_{mk} \longleftarrow V_{mk}\frac{[SU + \beta U]}{\left[VU^TU + \beta V\right]}, \tag{15}$$

$$M_{kf} \longleftarrow M_{kf}\frac{\left[U^TT\right]}{\left[U^TUM\right]}. \tag{16}$$

The algorithm for the optimization (7) is summarized in Algorithm 1. Note that the updating with variable step sizes will naturally maintain the nonnegative constraints. Because the number of nodes N is far greater than the number of node features F, the time complexity is $O(\mathrm{KN}^2)$.

4. Experiments and Analysis

In this section, we have conducted extensive comparative experiments to evaluate the effectiveness of our proposed CDCN model on real graph datasets with ground-truth communities.

4.1. Datasets. We consider 7 widely accepted network ground-truth community datasets, i.e., Karate, Polbooks, Football (http://www-personal.umich.edu/ mejn/netdata/), Citeseer, WebKB (http://linqs.cs.umd.edu/projects/projects/lbc/), Facebook (http://snap.stanford.edu/data/ego-Facebook.html), and HEP-TH (http://snap.stanford.edu/data/cit-HepTh.html). The network statistics are reported in Table 2. The Citeseer network consists of 3312 scientific publications with 4732 edges. The number of node attribute features in Citeseer is 3703. The WebKB network includes 4 subnetworks (i.e., Cornell, Texas, Washington, and Wisconsin), and each subnetwork consists of 5 communities. There are 877 web pages with 1608 edges, and each webpage is annotated by 1703-dimensional binary-valued word attributes. Karate, Polbooks, and Football are nonoverlapping communities without node attributes. The HEP-TH (high energy physics theory) citation graph is from the ePrint arXiv and covers all the citations within a dataset of 27770 papers with 352807 edges, and we believe that the papers published in the same journal belong to the same community. There are some communities that contain very few nodes; therefore, we exclude these communities (less than 10 nodes) and then get the dataset of 20048 papers with 236230 edges.

The node attributes of Cornell, Texas, Washington, Wisconsin, Citeseer, and Facebook are binary vectors where the elements are either 0 or 1. The node attributes of HEP-TH are dense vector with a dimension of 300. We extract the paper titles and abstracts and then the train word vector model [41] to get the vector.

We compare different methods in these networks to prove the effectiveness of our community structure matrix. The detailed information of the datasets can be seen in Table 2.

4.2. Evaluation Methods. The compared methods may include nonoverlapping and overlapping communities, and so we choose different evaluation metrics.

(i) For nonoverlapping communities:

In terms of the measures to evaluate the quality of nonoverlapping communities, we use two evaluation metrics. We adopt the same evaluation procedure used in [17] that every detected community is matched with its most similar ground-truth community.

The first metric is the accuracy (AC [48]). Given a network containing $|V|$ nodes, for each node, $\mathrm{predict}_i$ is the community label we obtain by applying different algorithms, and real_i is the ground-truth label provided by the datasets. The accuracy is defined as follows:

$$C = \frac{\sum_{i=1}^{n}\delta(\mathrm{real}_i, \mathrm{map}(\mathrm{predict}_i))}{|V|}, \tag{17}$$

Table 2: Dataset statistics. $|V|$: number of nodes; $|E|$: number of edges; f: number of node attribute features; K: number of communities.

Datasets	$\|V\|$	$\|E\|$	f	K
Karate	34	156	—	2
Polbooks	105	882	—	3
Football	115	1226	—	12
Cornell	195	283	1703	5
Texas	187	280	1703	5
Washington	230	366	1703	5
Wisconsin	265	459	1703	5
Citeseer	3312	4536	3703	6
Facebook	4039	88243	10	193
HEP-TH	20048	236230	300	537

where $\delta(x, y)$ is the function δ that equals 1 if $x = y$ and it equals 0 otherwise, and $\text{map}(\text{predict}_i)$ is the mapping function that maps each community label real_i to the equivalent label from the datasets. The best mapping can be found by using the Kuhn-Munkres algorithm [49].

The second metric is the normalized mutual information (NMI [48]). In clustering applications, mutual information is used to measure the similarity of two sets of clusters. Given the discovered communities C of the results of community detection methods and a set of ground-truth communities C^*, their mutual information metric $\text{NMI}(C, C^*)$ is defined as follows:

$$\text{NMI}(C, C^*) = \sum_{c_i \in C, c_j^* \in C^*,} p\left(c_i, c_j^*\right) .\log \frac{p\left(c_i, c_j^*\right)}{p(c_i).p\left(c_j^*\right)}, \quad (18)$$

where $p(c_i)$ and $p(c_j^*)$ denote the probabilities that a node arbitrarily selected from the network belongs to the community c_i and c_j^*, respectively, and $p(c_i, c_j^*)$ denotes the joint probability that this arbitrarily selected node belongs to clusters c_i and c_j^* at the same time.

(ii) For overlapping communities:

We compare a set of detected communities M with the ground-truth communities $M*$ as in [15], and the evaluation function is as follows:

$$\frac{1}{2\,|C*|} \sum_{c_i^* \in C^*} \max_{c_j \in C} \delta\left(c_i^*, c_j\right) + \frac{1}{2\,|C|} \sum_{c_i \in C} \max_{c_j \in C} \delta\left(c_i^* \in C^*\right), \quad (19)$$

where $\delta(c_i^*, c_j)$ is the similarity measure between communities c_i^* and c_j. We consider a standard metric $\delta(.)$ to quantify the similarity between a pair of communities, and the similarity score will be between 0 and 1.1 indicates the perfect recovery of the ground-truth communities.

4.3. Parameter Sensitivity Analysis. In this section, we perform the parameter sensitivity analysis of CDCN on the Wisconsin and Washington dataset. The number of nodes in these two is appropriate, which makes it easier to see the effect. Our algorithm has two hyperparameters: α is a nonnegative constant that controls the balance between the original network topology information and node attribute information, and β is a nonnegative constant for the relaxation of matrix U. For each hyperparameter value, we repeated the experiments ten times and took the average of the ten results. The results of other datasets are similar. These two parameters are also applicable to other datasets because they have similar topological structures and network characteristics. The results of the two parameters can be seen in following figures.

Figures 2 and 3 illustrate that CDCN achieves better performances in the range of $\alpha = 0.15$ through $\alpha = 0.3$, and specifically, CDCN achieves the highest AC and NMI scores when $\alpha = 0.2$. This indicates the different importance of the original network topology information and the node attribute information. In terms of the parameter β, we set $\alpha = 0.2$ and vary β from 0 to 0.05.

Figures 4 and 5 illustrate that CDCN achieves the highest AC and NMI scores when $\beta = 0.02$. These results indicate that if we use both node attribute information and community structure information, we will get better results for real networks.

4.4. Experimental Setups. We compared our algorithm against six topology based methods, i.e., SNMF, SLPA, DEMON, CPM, Louvain, and InfoMap; three node attributes based methods, i.e., CAN, SMR, and NC; and three methods that consider both network topologies and node attributes, i.e., PCL-DC, SCI, and CDE.

As with the experiments in [27], it is hard to compare the quality of community results when the numbers of communities are different for baseline methods. Therefore, we set the number of detected communities K as the number of ground-truth communities. We applied our proposed method and other baseline methods on the public datasets, repeat the tests ten times, and take the average of the ten results.

For all baseline methods, we set their parameters by default to achieve the best results for those methods. For example, for CDE, we set $\alpha = 1$, $\beta = 2$, and $\kappa = 5$, and for SCI, we set $\alpha = 50$ and $\beta = 1$. For more information, please refer to their papers. Regarding the parameters of our CDCN approach, *maxIter* is set to100 to achieve convergence, and the hyperparameters α and β are set as 0.2 and 0.02, respectively. Our algorithms are implemented in python, and all experiments are performed on a PC with Windows 7, Intel(R) Core(TM) i7-4790 CPU @ 3.60 GHz and 24 GB of main memory.

4.5. Evaluation on Nonoverlapping Communities. In this subsection, we evaluate the results on nonoverlapping communities. We report the ACs and NMIs of all methods in Table 3. The results indicate that CDCN outperforms all

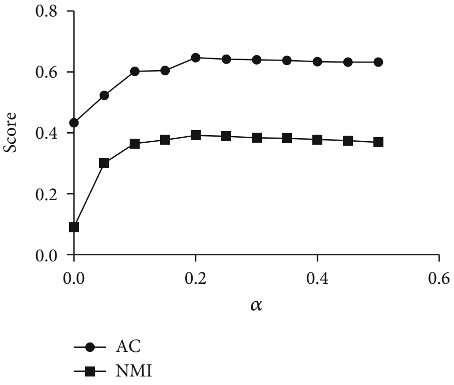


FIGURE 2: Fixing $\beta = 0.01$ then varying α from 0 to 0.5 on the Wisconsin dataset.

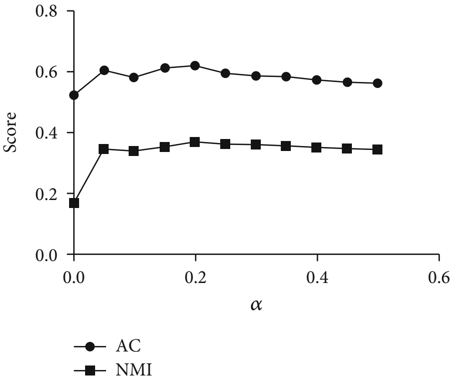


FIGURE 3: Fixing $\beta = 0.01$ then varying α from 0 to 0.5 on the Washington dataset.

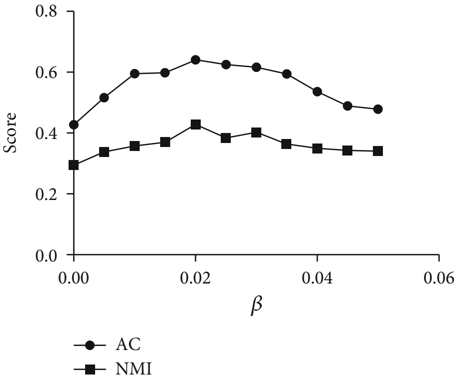


FIGURE 4: Fixing $\alpha = 0.2$ then varying β from 0 to 0.05 on the Wisconsin dataset.

comparison algorithms for the nonoverlapping community detection task.

The baseline comparison methods include InfoMap, CPM, SLPA, Louvain, DEMON, SNMF, CAN, SMR, NC, PCL-DC, SCI, and CDE. The real datasets include Cornell, Texas, Washington, Wisconsin, and Citeseer. All the datasets are independent of each other, and there is no connection between them; therefore, they are nonoverlapping communities. We apply our method to the above datasets by using the different baseline methods.

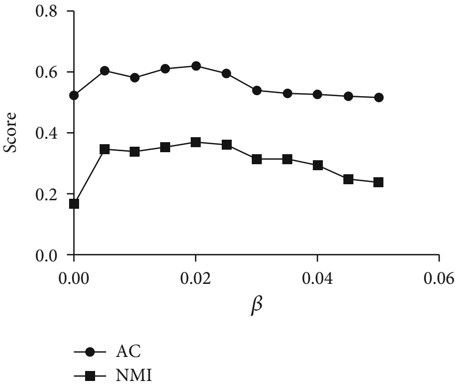


FIGURE 5: Fixing $\alpha = 0.2$ then varying β from 0 to 0.05 on the Washington dataset.

Compared with the algorithms that focus on the original network topology or node attributes, the results show that combining both the original network topology and the inherent community structure information together will result in making great improvements. For example, among the algorithms that focus on the original network topology or node attributes, the highest ACs on Cornell, Texas, Washington, Wisconsin, and Citeseer are, respectively, 0.446, 0.545, 0.508, 0.471, 0.314, and 0.168, and ACs on our methods are 0.569, 0.641, 0.695, 0.694, 0.544, and 0.215, respectively. The values increased by 0.123, 0.096, 0.187, 0.223, 0.230, and 0.047, respectively. The same as the AC, our method also greatly improved the NMI.

The experiments results can be seen in Table 3. From Table 3, we can see that compared with the PCL-DC, SCI, and CDE models that combine both original network topology information and node attribute information, obviously, PCL-DC, SCI, and CDE will lose the information on the relationships between nodes that are not directly connected, and our model considers the relationships between nodes that share the same neighbors. The results in Table 3 prove this. In Table 3, CDE gets the best results compared with PCL-DC and SCI. The highest ACs of CDE on Cornell, Texas, Washington, Wisconsin, Citeseer, and HEP-TH are, respectively, 0.499, 0.498, 0.568, 0.645, 0.474, and 0.201. Our model improves the AC values by 0.07, 0.143, 0.127, 0.049, 0.07, and 0.014, respectively. Regarding the AC, compared with the methods that combine both the original network topology information and node attribute information, there are also huge improvements in the results.

4.6. Evaluation on Overlapping Communities. Some scholars have proposed effective methods in the detection of overlapping communities, such as subspace decomposition, maximal cliques, maximal subgraph, and the clustering coefficient. Li et al. [50] proposed a method to measure the performance of the overlapping communities. They are the state-of-the-art methods in recent years. With reference to

Table 3: The performances of different community detection algorithms on nonoverlapping communities measured by the AC and NMI.

Metric	Datasets	InfoMap	CPM	SLPA	Louvain	DEMON	SNMF	CAN	SMR	NC	PLC-DC	SCI	CDE	*CDCN*
AC	Cornell	0.2	0.446	0.239	0.266	0.377	0.371	0.446	0.415	0.317	0.348	0.354	0.449	0.569
	Texas	0.214	0.471	0.523	0.269	0.475	0.496	0.545	0.470	0.540	0.369	0.488	0.498	0.641
	Washington	0.143	0.469	0.434	0.204	0.381	0.410	0.491	0.508	0.456	0.408	0.401	0.568	0.695
	Wisconsin	0.152	0.471	0.262	0.223	0.430	0.386	0.471	0.471	0.422	0.354	0.396	0.645	0.694
	Citeseer	0.053	0.178	0.090	0.238	0.208	0.309	0.212	0.211	0.314	0.452	0.327	0.474	0.544
	HEP-TH	0.041	0.123	0.091	0.135	0.113	0.168	0.135	0.162	0.143	0.193	0.184	0.201	0.215
NMI	Cornell	0.147	0.05	0.138	0.109	0.051	0.061	0.045	0.061	0.084	0.081	0.073	0.311	0.358
	Texas	0.102	0.077	0.040	0.059	0.043	0.097	0.021	0.09	0.115	0.068	0.097	0.252	0.328
	Washington	0.117	0.006	0.158	0.087	0.079	0.035	0.046	0.117	0.038	0.103	0.086	0.341	0.406
	Wisconsin	0.111	0.027	0.106	0.078	0.047	0.070	0.037	0.070	0.077	0.071	0.069	0.406	0.427
	Citeseer	0.214	0.199	0.214	0.228	0.166	0.096	0.003	0.007	0.003	0.221	0.083	0.208	0.263
	HEP-TH	0.035	0.092	0.084	0.112	0.093	0.137	0.114	0.123	0.114	0.152	0.145	0.179	0.191

these methods, we proposed evaluation metrics under overlapping communities.

For the overlapping communities, we use the F1-Score and Jaccard-Similarity to evaluate the partitioned results of all the methods, except the clustering methods that could not discover the overlapping communities. The tested network is the complete Facebook data, and it contains 10 different ego-networks with manually identified circles. We select 4 representative ego-networks from them. The experiment results can be seen in Table 4.

The baseline comparison methods include InfoMap, CPM, SLPA, Louvain, DEMON, SNMF, PCL-DC, SCI, and CDE. The real datasets include FaceBook Ego-network 107, FaceBook Ego-network 698, FaceBook Ego-network 1912, FaceBook Ego-network 3908, and FaceBook. There are some intersections between datasets, and some overlapping areas appear; therefore, there are overlapping communities. We apply our method to the above datasets by using the different baseline methods.

Ego-network 107 has the most nodes, Ego-network 698 has the fewest nodes, Ego-network 1912 has the highest intensive degree, and Ego-network 3908 has lowest intensive degree.

As shown in Table 4, CDCN achieves the best performances on all the tested networks. In addition, it also shows that our model greatly improves community detection by combining both the original network topology and the inherent community structure information together. For instance, the highest F1-Score among the methods that focus on network topology information is 0.517, and the highest F1-Score among the methods that combine both the original network topology and the inherent community structures information together is 0.474 on FaceBook ego-network 107. Compared with those methods, our model achieves an F1-Score of 0.539. The other ego-network results are similar with those of FaceBook ego-network 107. On the complete Facebook data, our model gets an F1-Score of 0.372, which is greater than the best results of the other methods. For the Jaccard-Similarity, the results in Table 4 also indicate that CDCN outperforms all comparison algorithms.

4.7. Evaluation on Nonnode Attributes Communities. There are some communities without node attributes, and it is hard to divide these communities using the majority methods that use node attributes. However, it is easy to deal with the problem using CDCN since we could use only the community structure part of our method. Therefore, our method will be simplified as follows:

$$L(U, V) = \min_{U\geq 0, V\geq 0} \left\| S - UV^T \right\|_F^2 + \beta \| U - V \|_F^2. \tag{20}$$

The update formulas are changed into the following:

$$U_{mk} \longleftarrow U_{mk} \frac{[SV + \beta V]}{\left[UV^T V + \beta U \right]}, \tag{21}$$

$$V_{mk} \longleftarrow V_{mk} \frac{[SU + \beta U]}{\left[VU^T U + \beta V \right]}. \tag{22}$$

To prove the usefulness of our community structure matrix, we add two more baselines, which are called Adj-Mat and Emb-Mat. Adj-Mat just replaces the community structure matrix S in equation (20) with the adjacency matrix of network. Similarly, Emb-Mat just replaces the community structure matrix in equation (20) with the embedding matrix of CDE. Then, the optimization algorithm is changed, as shown in Algorithm 2.

The baseline comparison methods include InfoMap, CPM, SLPA, Louvain, DEMON, SNMF, Adj-Mat, and Emb-Mat. We assessed the NMI and AC values on the karate, football, and polbooks datasets.

In addition, in this part, we compare our method with the methods that focus on the original network topology and that do not node attribute information on four datasets. The four datasets are nonoverlapping communities, and so, we use AC and NMI to evaluate the partitioning result of all the methods. The results can be seen in Table 5.

As shown in Table 5, the results indicate that CDCN outperforms all comparison algorithms for the nonoverlapping community detection task. The NMI reached 1.0, 0.916,

Table 4: The performances of different community detection algorithms on overlapping communities measured by the F1-Score and Jaccard-Similarity.

Metric	Datasets	InfoMap	SLPA	Louvain	DEMON	SNMF	PCL-DC	SCI	CDE	CDCN
F1-Score	FaceBook ego-network 107	0.448	0.510	0.264	0.517	0.378	0.384	0.405	0.474	0.539
	FaceBook ego-network 698	0.636	0.628	0.588	0.576	0.612	0.345	0.239	0.574	0.640
	FaceBook ego-network 1912	0.372	0.323	0.366	0.328	0.378	0.312	0.316	0.322	0.379
	FaceBook ego-network 3908	0.579	0.528	0.567	0.387	0.410	0.421	0.388	0.471	0.580
	FaceBook	0.330	0.351	0.321	0.214	0.134	0.224	0.213	0.324	0.372
Jaccard-Similarity	FaceBook ego-network 107	0.372	0.410	0.205	0.421	0.267	0.304	0.294	0.369	0.432
	FaceBook ego-network 698	0.556	0.529	0.482	0.464	0.487	0.256	0.141	0.441	0.571
	FaceBook ego-network 1912	0.286	0.272	0.251	0.229	0.249	0.200	0.202	0.215	0.290
	FaceBook ego-network 3908	0.432	0.441	0.421	0.313	0.287	0.310	0.268	0.352	0.441
	FaceBook	0.206	0.213	0.242	0.178	0.169	0.182	0.168	0.224	0.262

```
Input:network graph $G$, hyper-parameters $\beta$, number of communities K and maximum number of iterations $maxIter$.
Output:the probability distribution matrix U.
Begin
According to the network graph $G$ and Eq.(1), generate the community structure matrix and randomly initialize the probability distribution matrix
$U^{(0)} \in (0,1)^{N\times K}$, $V^{(0)} \in (0,1)^{N\times K}$ ; $i = 0$
While$i \le maxIter$do
Update $U^{(i+1)}$ according to Eq.(21);
Update $V^{(i+1)}$ according to Eq.(22);
End While
End
```

Algorithm 2: The optimization process of CDCN.

Table 5: The performances of the different community detection algorithms on nonoverlapping as community measured by the AC and NMI.

Metric	Datasets	InfoMap	CPM	SLPA	Louvain	DEMON	SNMF	Adj-Mat	Emb-Mat	CDCN
NMI	Karate	0.581	0.652	0.658	0.569	0.429	0.836	0.825	0.912	1.0
	Football	0.901	0.855	0.582	0.713	0.463	0.894	0.875	0.903	0.909
	Polbooks	0.412	0.538	0.498	0.547	0.383	0.508	0.487	0.524	0.570
AC	Karate	0.894	0.888	0.820	0.741	0.688	0.970	0.962	0.970	1.0
	Football	0.918	0.897	0.613	0.716	0.546	0.891	0.867	0.911	0.923
	Polbooks	0.695	0.819	0.785	0.823	0.740	0.743	0.724	0.798	0.838

and 0.570 on the karate, football, and polbooks datasets, respectively, which are better than the above methods. The AC value reached 1.0, 0.909, and 0.838 on karate, football, and polbooks datasets, respectively, which are also better than the above methods. Furthermore, the results compared with the Adj-Mat and Emb-Mat also show the great usefulness of our community structure matrix.

5. Conclusions

Community detection has been widely used in recommendation systems, social networks, and network security. Efficient and fast community detection algorithms contribute to the development of intelligent networks. Based on the analysis of the network characteristics, in this paper, in order to solve the problem of community detection in attributed graphs, we propose a novel method to generate the community structure matrix, which retains the relationship between two directly nodes connected or nodes that share the same neighbors, and named it CDCN. We combine node attribute information and community structure information in an effective way in the nonnegative matrix factorization framework. We used two indicators named AC and NMI on nonoverlapping communities and two indicators named F1-Score and Jaccard-Similarity on overlapping communities to evaluate our method. On nonoverlapping communities, the AC and NMI values of CDCN are better than those of other methods. On overlapping communities, the F1-score and Jaccard-Similarity value of CDCN are better than those of other methods. The extensive experimental results demonstrated that our algorithm can effectively discover the communities in real networks.

Data Availability

The original dataset used in this work is available from the corresponding author on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported partly by the National Key Research and Development Program of China under Grant No. 2017YFB1400200 and the Science and Technology Plan in Key Fields of Yunnan under Grant No. 202001BB050076.

References

[1] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks," *Acm computing surveys (csur)*, vol. 45, no. 4, pp. 1–35, 2013.

[2] C. Pizzuti, "Evolutionary computation for community detection in networks: a review," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 3, pp. 464–483, 2018.

[3] S. Fortunato and D. Hric, "Community detection in networks: a user guide," 2016, https://arxiv.org/abs/1608.00163/.

[4] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," 2016, https://arxiv.org/abs/1608.00763/.

[5] M. E. J. Newman, "Spectral methods for network community detection and graph partitioning," 2013, https://arxiv.org/abs/1307.7729/.

[6] G. Ren and X. Wang, "Epidemic spreading in time-varying community networks," *Chaos*, vol. 24, no. 2, article 023116, 2014.

[7] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

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

[9] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, article P10008, no. 10, 2008.

[11] J. Xie, B. K. Szymanski, and X. Liu, "Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 344–349, Vancouver, BC, Canada, December 2011.

[12] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 615–623, Beijing, China, August 2012.

[13] D. He, X. You, Z. Feng, X. Y. Di Jin, and W. Zhang, "A network-specific markov random field approach to community detection," in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, February 2018.

[14] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, article 066111, 2004.

[15] Y. Li, K. He, D. Bindel, and J. E. Hopcroft, "Uncovering the small community structure in large networks: a local spectral approach," in *Proceedings of the 24th international conference on world wide web*, pp. 658–668, Florence, Italy, May 2015.

[16] Y. Cui and X. Wang, "Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks," *Physica A: Statistical Mechanics & Its Applications*, vol. 407, pp. 7–14, 2014.

[17] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 587–596, Rome, Italy, February 2013.

[18] X. Huang, H. Cheng, and J. X. Yu, "Dense community detection in multi-valued attributed networks," *Information Sciences*, vol. 314, pp. 77–99, 2015.

[19] M. Atzmueller, S. Doerfel, and F. Mitzlaff, "Description-oriented community detection using exhaustive subgroup discovery," *Information Sciences*, vol. 329, pp. 965–984, 2016.

[20] X. Wang, D. Jin, X. Cao, Y. Liang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 265–271, Phoenix, Arizona, USA, 2016.

[21] X. Huang, H. Cheng, and J. X. Yu, "Attributed community analysis: global and ego-centric views," *IEEE Database Engineering Bulletin*, vol. 39, no. 3, pp. 29–40, 2016.

[22] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 927–936, Paris, France, 2009.

[23] X. Luo, Z. Liu, M. Shang, J. Lou, and M. C. 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.

[24] H. Lu, X. Sang, Q. Zhao, and J. Lu, "Community detection algorithm based on nonnegative matrix factorization and improved density peak clustering," *IEEE Access*, vol. 8, pp. 5749–5759, 2020.

[25] M. Zhang and Z. Zhou, "Structural Deep Nonnegative Matrix Factorization for community detection," *Applied soft computing*, vol. 97, no. Part B, article 106846, 2020.

[26] S. Wang, G. Li, G. Hu, H. Wei, Y. Pan, and Z. Pan, "Community detection in dynamic networks using constraint nonnegative matrix factorization," *Intelligent Data Analysis*, vol. 24, no. 1, pp. 119–139, 2020.

[27] Y. Li, C. Sha, X. Huang, and Y. Zhang, "Community detection in attributed graphs: an embedding approach," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on*

Educational Advances in Artificial Intelligence (EAAI-18), pp. 338–345, New Orleans, Louisiana, USA, 2018.

[28] X. Wang and J. Li, "Detecting communities by the core-vertex and intimate degree in complex networks," *Physica A: Statistical Mechanics & Its Applications*, vol. 392, no. 10, pp. 2555–2563, 2013.

[29] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS)*, pp. 556–562, Denver, CO, USA, 2000.

[30] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

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

[32] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," 2010, https://arxiv.org/abs/1008.3926/.

[33] W. Ren, G. Yan, X. Liao, and L. Xiao, "Simple probabilistic algorithm for detecting community structure," *Physical review. E, Statistical, nonlinear and soft matter physics*, vol. 79, no. 3, article 036111, 2009.

[34] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14*, pp. 977–986, New York, NY, USA, 2014.

[35] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 3834–3841, Columbus, OH, USA, 2014.

[36] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[37] A. F. McDaid, T. B. Murphy, N. Friel, and N. J. Hurley, "Improved bayesian inference for the stochastic block model with application to large networks," *Computational Statistics & Data Analysis*, vol. 60, pp. 12–31, 2013.

[38] Y. Chen, H. Zhang, X. Zhang, and R. Liu, "Regularized semi-non-negative matrix factorization for hashing," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1823–1836, 2018.

[39] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between LDA and NMF based schemes," *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.

[40] Y. Chen, H. Zhang, Y. Tong, and M. Lu, "Diversity regularized latent semantic match for hashing," *Neurocomputing*, vol. 230, pp. 77–87, 2017.

[41] C. He, X. Fei, H. Li, Y. Tang, H. Liu, and S. Liu, "Improving NMF-based community discovery using distributed robust nonnegative matrix factorization with SimRank similarity measure," *The Journal of Supercomputing*, vol. 74, no. 10, article 2500, pp. 5601–5624, 2018.

[42] J. Wang, Y. Fan, L. Feng, Z. Ye, and H. Zhang, "Research hotspot prediction and regular evolutionary pattern identification based on NSFC grants using NMF and semantic retrieval," *IEEE Access*, vol. 7, pp. 123776–123787, 2019.

[43] D. Yu, N. Chen, F. Jiang, B. Fu, and A. Qin, "Constrained NMF-based semi-supervised learning for social media spammer detection," *Knowledge-Based Systems*, vol. 125, pp. 64–73, 2017.

[44] J. Gao, Y. Lu, J. Qi, and L. Shen, "A radar signal recognition system based on non-negative matrix factorization network and improved artificial bee colony algorithm," *IEEE Access*, vol. 7, pp. 117612–117626, 2019.

[45] N. Y. Chen, Y. Liu, and H.-C. Chao, "Overlapping community detection using non-negative matrix factorization with orthogonal and sparseness constraints," *IEEE Access*, vol. 6, pp. 21266–21274, 2018.

[46] X. Li, Z. Hu, and H. Wang, "Combining non-negative matrix factorization and sparse coding for functional brain overlapping community detection," *Cognitive Computation*, vol. 10, no. 6, article 9585, pp. 991–1005, 2018.

[47] 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, 2018.

[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, Lake Tahoe, NV, USA, 2013.

[49] X. Cai and F. Sun, "Supervised and constrained nonnegative matrix factorization with sparseness for image representation," *Wireless Personal Communications*, vol. 102, no. 4, article 5325, pp. 3055–3066, 2018.

[50] J. Li, X. Wang, and J. Eustace, "Detecting overlapping communities by seed community in weighted complex networks," *Physica A: Statistical Mechanics & Its Applications*, vol. 392, no. 23, pp. 6125–6134, 2013.