Accepted Manuscript

Graph regularized nonnegative matrix tri-factorization for overlapping community detection

Hong Jin, Wei Yu, ShiJun Li

PII: S0378-4371(18)31225-1
DOI: https://doi.org/10.1016/j.physa.2018.09.093
Reference: PHYSA 20154

To appear in: *Physica A*

Received date : 23 June 2017
Revised date : 10 August 2018

Please cite this article as: H. Jin, et al., Graph regularized nonnegative matrix tri-factorization for overlapping community detection, *Physica A* (2018), https://doi.org/10.1016/j.physa.2018.09.093

First, through the consideration of the intrinsic geometric information of the network graph represented by the low-dimensional manifold, we found that it could help achieve better community detection results.

Second, with the help of spectrum properties analysis of relative non-backtracking matrix for certain complex network, it is a good guide to the number of communities.

Third, when we added an orthogonal constraint condition on the membership indicator matrix, the NMF based model we proposed can combine the advantages of both fuzzy and non-fuzzy overlapping community detection methods.

Graph Regularized Nonnegative Matrix tri-factorization for Overlapping Community Detection

Hong Jin[a] , Wei Yu[a] and ShiJun Li [a]
[a]School of Computer Science, Wuhan University, Wuhan 430072, China
E-mail: anya_1024@163.com

Abstract:

Non-negative Matrix Factorization technique has attracted many interests in overlapping community detection due to its performance and interpretability. However, when adapted to discover community structure the intrinsic geometric information of the network graph is seldom considered. In view of this, we proposed a novel NMF based algorithm called Graph regularized nonnegative matrix tri-factorization (GNMTF) model, which incorporates the intrinsic geometrical properties of the network graph by manifold regularization. Moreover, by using three factor matrices we can not only explicitly obtain the community membership of each node but also learn the interaction among different communities. The experimental results on two well-known real world networks and a benchmark network demonstrate the effectiveness of the algorithm over the representative non-negative matrix factorization based method.

Key words: graph regularized; geometrical structure; nonnegative matrix tri-factorization; community detection

1. Introduction

Networks (or graphs) have become ubiquitous as data from diverse areas can naturally be mapped to graph structures [1]. Social networks, Information networks, Technological networks and so on have been shown a structure of modules [2]. Modules are often referred to as communities within which nodes are much more connected to each other than to the rest of the network [3]. Modules or communities can reveal meaningful information of these networks [4]. The problem of extracting communities in an efficient and effective way has become crucial. Towards this end, graph mining and analysis methods constitute prominent tools [5].

Graph mining algorithms usually identify network patterns in the form of subgraphs that satisfy a given criterion. Which in most cases is related to the structure of the graph. There are some recent studies about graph mining that can be adapted to community detection. The Bag Constrained Structure Pattern Mining for Multi-Graph Classification formulate a multi-graph learning task to build a learning model from a number of labeled training bags to predict previously unseen test bags with maximum Accuracy [6]. Boosting for Multi-Graph Classification formulate a novel graph-based learning problem, which aims to learn a classifier from a set of labeled bags each containing a number of graphs inside the bag [7]. Multiple Structure-View Learning for Graph Classification advance graph classification to handle multigraph learning for complicated objects from multiple structure views, where each object is represented as a bag containing several graphs and the label is only available for each graph bag but not individual graphs inside the bag [8].Positive and Unlabeled Multi-Graph Learning handles multi-graph learning for complicated objects, where each object is represented as a bag of graphs and the label is only available to each bag but not individual graphs[9].

Though they are about multi-graph classification, their ideas are relatively new and can provide a better range of methods.

Recently, overlapping community detection is becoming a new hotspot of community detection research. To this issue, most existing partitioning algorithms and hierarchical clustering methods are not workable solutions. The first method for overlapping community detection called the clique percolation method (CPM) [10]. Subsequently, several extensions based on CPM have been proposed. Meanwhile, by considering with the different interpretations of the factor matrices, the NMF could be used for either feature extraction and clustering. Then many NMF based clustering algorithms have been proposed to solve overlapping community detection problem and exhibited competitive performance [11]. For example, Zhang et al proposed a bounded nonnegative three factor matrix factorization model called BNMTF to discover the community membership of each node explicitly and also the interaction among different communities [12]. It was argued that outperforms other NMF-based methods.

Moreover, the NMF model itself has developed rapidly in recent years. These studies provided new feasible way for the NMF based overlapping community detection algorithm design. Robust Dual Clustering with Adaptive Manifold Regularization simultaneously performs dual matrix factorization tasks with the target of an identical cluster indicator in both of the original and projected feature spaces, respectively [13]. Ensemble manifold regularized sparse low-rank approximation for multi-view feature embedding encodes multi-view feature into a unified and discriminative embedding for a given task [14]. Truncated Cauchy Nonnegative Matrix Factorization for Robust Subspace robustly learns the subspace on noisy datasets contaminated by outliers [15].

Even though the NMF-based methods show good performance in overlapping community detection, the concerned research should be strengthened in both breadth and depth. At present, most of the NMF based methods when adapted to overlapping community detection problem design the strategies from two directions. On the one hand, they focus on how to construct the data matrix to be decomposed in the NMF based model, like adjacency matrix based, physical process based, shortest path based and common neighbor based methods [16]. On the other hand, they put particular emphasis on how to decrease computational complexity of the NMF model itself [17]. In this paper, we intended to solve this problem from a different viewpoint. The idea comes from the recent research results in matrix factorization [18]. By manifold learning, NMF based approaches are able to preserve the intrinsic geometric structure of the original data space which is essentially useful for classification and clustering [19]. It was argued that learning performance can be significantly enhanced if the geometrical structure is exploited [20].

In view of the above mentioned, here we introduce a novel method called graph regularized nonnegative matrix tri-factorization (GNMTF). It incorporated the graph structure as a regularization term into the objective function of the standard NMF model. Specifically, we use the symmetric three factor matrices factorization denoted as HSH^T, in which H represents the membership indicators while S represents

the relationship among all different communities[21]. Each entry in H indicates the membership strength that a node belongs to certain community [22]. Furthermore, with the orthogonal constraint H can explicitly assign the node membership to a certain community [23]. Therefore, we are able to combine the advantages of both fuzzy and non-fuzzy overlapping community detection methods. The former estimates the strength of memberships while not being able to provide clear node membership to each community. The latter gives crisp partitions allowing each node to have multiple community labels but without any information about the strength of the nodes' membership to each community.

Obviously, factor matrices H and S have clear physical meanings. Moreover, by incorporating the intrinsic geometric structure as an additional regularization term we can acquire a new object function. Hence, our algorithm is particularly applicable to datasets which have apparent geometrical structure. As a novelty method, it is able to maintain competitive efficiency when discovering community structure. And shows great potential in preserving the geometrical structure, but further research and development is still required.

Particularly, for NMF based algorithms another key issue the parameterization should be pointed out [24]. Most of the relative parameters have reasonable default values. But the number of communities referred to as the underlying factor of the NMF model is an exception. Its value is uncertain and several methods have been developed to infer the number of communities [25]. In this paper, it chooses a simple way to determine the most appropriate number of communities by the eigenvalue distribution of the relative non-backtracking matrix of the network [26]. Recent studies have already demonstrated its effectiveness [27].

The rest of the paper is organized as follows: Section II presents how the graph regularized 3-factor NMF to deal with community detection problem. Section III gives a brief view of the parameterization for the underlying factor by non-backtracking matrix. Section V shows the experimental results. Finally, Section VI concludes.

2. Community Detection via GNMTF

In this section, we described the GNMTF model in detail including two parts such as the model formulation and the graph regularized item analysis. While the context of parameter learning will be presented in the next section.

2.1 Model Formulation

Given a complex network $G=(V,E)$, in which V represents a set of nodes and E represents a set of edges with each of them connecting a pair of nodes in V. Let A denote the adjacency matrix for G, if there is an edge between the ith and jth nodes then the $(i,j)th$ entry of A is 1, otherwise the $(i,j)th$ entry of A is 0.

With the assumption that all the networks considered in this paper are undirected, A is symmetric accordingly. Due to the pairwise similarities of the input matrix, it is a special case of 3-factor NMF called symmetric 3-factor NMF [28]. Which are the

basic model we actually used. Another key parameter for standard NMF model the underlying factor k will be discussed in the next section. Here we assumed that its value is given in advance. Specific to community detection, the factor matrix $H \in R_+^{N\times K}$ represents the community membership of the N nodes in V while $S \in R_+^{K\times K}$ represents the relationship among the K different communities. Moreover, the physical meaning of the $(i,j)th$ entry in H is the membership strength that the ith node belongs to the jth community[12]. The higher the value of H_{ij}, the greater the possibility of the ith node to the jth community.

Usually in real networks each node just participates in relatively small number of communities and H is sparse accordingly. Here we use l_1 norm to enhance the sparsity of H [28]. Under the constraint of l_1 norm the effective number of free parameters in H are reduced to a large extent so the complexity of the basic 3-factor NMF model is greatly decreased[29].

Next, let us get down to the adaption of the 3-factor NMF model when deal with community detection. The product form HSH^T indicates the interaction between any two nodes in view of the community structure. It will be used to approximate the adjacency matrix A. Equivalently, in the form of nonnegative matrix tri-factorization we approximate the adjacency matrix A with the above mentioned product form as formula (1) shows [8]:

$$A \approx \hat{A} = HSH^T \ s.t.\ H \geq 0,\ S \geq 0 \quad (1)$$

The approximation error can be measured by loss function such as squared loss and generalized KL-divergence [25]. They are respectively defined as formula (2) and formula (3) show:

$$L_{sq}(A,H,S) = \left\| A - HSH^T \right\|_F^2 \quad (2)$$

$$L_{KL}(A,H,S) = \sum \left(A_{ij} \log \frac{A_{ij}}{\left(HSH^T\right)_{ij}} - A_{ij} + \left(HSH^T\right)_{ij} \right), (3)$$

Where $\left(HSH^T\right)_{ij}$ is the $(i,j)th$ entry of $\hat{A}$.

As discussed above, H is usually sparse. Before combing the graph regularized item the above optimization problem underlying 3-factor NMF can be formulated as equation (4) shows to minimize the new objective function denoted by O [18]

$$O = L(A,H,S) + \alpha \sum_{ij} H_{i,j}$$

$$= \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{\left(HSH^T\right)_{ij}} - A_{ij} + \left(HSH^T\right)_{ij} \right) + \alpha \sum_{ij} H_{ij} \quad (4)$$

$$\text{s.t. } H,S \geq 0, \alpha \geq 0$$

Theoretically the loss function $L(A,H,S)$ in the objective function can be chosen as any from in Eqs.(2) or (3). Here we chose the KL-divergence. Moreover, the parameter $\alpha \geq 0$ would help balance the tradeoff between the approximation error and the complexity of H.

2.2 NMTF with Manifold Regularization

From the perspective of data representation, the relative adjacency matrix $A = [\alpha_1, ..., \alpha_n] \in R^{N \times N}$ can be regarded as a high dimensional dataset and each column of A is a sample point. Due to its symmetry, NMTF aims to actually find two nonnegative matrices $H = [h_{ik}] \in R^{N \times K}$, $S = [s_{pq}] \in R^{K \times K}$ and $F = H^T = [h_{kj}] \in R^{K \times N}$ whose product can well approximate the original matrix A as formula (5) shows.

$$A \approx HSF = HSH^T \quad (5)$$

Moreover, by mapping $Z \leftarrow HS$ the above 3-factor NMF can be reduced to the unconstrained 2-factor NMF [21]. That is to say the degree of freedom of HSH is the same as HH^T. Then the ikth row and kth column element of matrix Z can be described as formula (6) shows

$$z_{ik} = \sum_{j=1}^{K} \sum_{p=1}^{K} h_{ij} s_{pk} \quad (6)$$

In fact, in most cases $K << N$. Thus NMF is able to find a compressed approximation of the original data matrix [18]. This approximation process can be explained column by column as formula (7) shows

$$\alpha_j \approx \sum_{k=1}^{K} \mathbf{z}_k h_{jk} \quad (7)$$

In which $\mathbf{z}_k$ represents the kth column vector of Z. Thus, all the columns of Z together can be regarded as a new basis. Each data vector α_j is approximated by a linear combination of the new basis and weighted by the components of H. If y_j^T

represents the jth row of H, which denoted as $y_j = \left[h_{j1},...,h_{jK}\right]^T$. Then y_j can be viewed as the new representation of the jth data point in the new coordinate axes defined by the new basis Z. Noted that only when the structure discovered by the basis vectors is latent in the data, a good approximation can be achieved.

For the above given complex network $G=(V,E)$, each column of the adjacency matrix A represents a node of the network. It is in a N-dimensional space, where N is often very large. Inspired by the idea of hidden metric space, in a lower dimensional manifold the key properties of the original network can be preserved.

By incorporating a geometrically based regularized constraint, NMF can maintain its performance on overlapping community detection meanwhile the accuracy potentially improved. Because the intrinsic geometrical and discriminating properties of the network is essential to the real-world applications. While recent studies in manifold learning have demonstrated that the geometric structure can be effectively encoded through constructing a nearest neighbor graph on a scatter of data points. A detailed description of the procedure is as follows [30].

Step 1. For the given complex network each node corresponds to a data point represented by $\alpha_i (i=1,...,n)$. The defined weight matrix W is as formula (8) shows:

$$W_{ij} = \begin{cases} 1, & if\ \alpha_i \in N_p(\alpha_j) \text{ or } \alpha_j \in N_p(\alpha_i) \\ 0, & otherwise, \end{cases} \quad (8)$$

For each data point α_i all its p nearest neighbors were found and denoted as $N_p(\alpha_i)$. Additionally, then W_{ij} can be used to measure the closeness of two data points α_i and α_j.

Step 2. With respect to the new basis Z the original data point α_j can be represented in the low-dimensional space as $y_j = \left[h_{j1},...,h_{jK}\right]^T$. We formalize the mapping process as function $f_k(\alpha_j) = h_{jk}$ shows. Then the Euclidean distance was used to measure the "dissimilarity" between the low-dimensional representations of two data points.

Combining with the above definition of weight matrix W, the following two terms can be used to measure the smoothness of the low-dimensional representation in the new basis. It is formalized as formula (9) shows

$$
\begin{aligned}
R &= \frac{1}{2}\sum_{i,j=1}^{N}\left\| f_k(\alpha_i) - f_k(\alpha_j)\right\|^2 W_{ij} \\
&= \frac{1}{2}\sum_{i,j=1}^{N}\left\| y_i - y_j\right\|^2 W_{ij} \\
&= \sum_{i=1}^{N} y_i^T y_i D_{ii} - \sum_{i,j=1}^{N} y_i^T y_j W_{ij} \qquad (9) \\
&= Tr\left(H^T DH\right) - Tr\left(H^T WH\right) \\
&= Tr\left(H^T LH\right)
\end{aligned}
$$

In which $Tr(\cdot)$ denotes the trace of a matrix, D is a diagonal matrix whose unit entries are column or row sums of W since W is symmetric. Then $D_{ii} = \sum_j W_{ij}$. Furthermore by defining $L = D - W$, the simplify expression can be obtained.

Step 3. By minimizing R, the mapping function f_k is sufficiently smooth on the data manifold. An intuitive explanation is that if two data points α_i and α_j are close for example W_{ij} is big, then $f_k(\alpha_i)$ and $f_k(\alpha_j)$ are also close to each other.

As a new constraint the geometrically-based regularizer was incorporated into the original NMF objective function. Consequently, the relative new objective function for our GNMF method can be shown in formula (10)

$$O_1 = \left\| A - ZH^T\right\|_{ii}^2 + \lambda Tr\left(H^T LH\right) \qquad (10)$$

In which $\lambda \geq 0$ it was a regularization parameter and controls the smoothness of the new representation.

2.3 Time complexity analysis

In this method, the major computational cost consists of standard NMF and the manifold regularization. For standard NMF the major operations are the matrix multiplication, the time complexity of multiplication for two matrices such as a $n \times k$ matrix and a $k \times n$ matrix is $O\left(n^2 k\right)$. Supposing the multiplicative updates stops after t iterations, the overall cost for standard NMF is $O\left(tn^2 k\right)$. For the manifold regularization, it needs $O\left(n^3\right)$ to construct the p nearest neighbor graph. Therefore, the time complexity for GNMTF is $O\left(tn^2 k + n^3\right)$.

3. Parameter Learning

As above mentioned, the key parameter for nonnegative matrix factorization is the underlying factor k. Specific to community detection, this parameter is amount to the

number of communities. Here we chose a simple and fast method to estimate the number of communities. Which previous studies have demonstrated its effectiveness, it was based on the spectral properties of certain graph operators such as non-backtracking matrix [31].

3.1 Estimating k from the non-backtracking matrix

Recall A is the adjacency matrix of the given complex network G. Let $d_i = \sum_{j=1}^{n} A_{ij}$ be the degree of node i. Next, we give the definition of the non-backtracking matrix which will be used to estimate the number of communities.

Here we denote m as the number of edges in the given undirected complex network. While the corresponding non-backtracking matrix denoted as B. When building matrix B, we present the edge between node i and node j by two directed edges, one from i to j and the other from j to i [32]. Then the $2m \times 2m$ matrix B can be defined by formula (11)

$$B_{i\to p,q\to l} = \begin{cases} 1 & if\ p = q\ and\ i \neq l \\ 0 & otherwise \end{cases} \quad (11)$$

The spectrum of B is proved to be made up of ± 1 and the eigenvalues of an $2n \times 2n$ matrix denoted by formula (12) [33]

$$\tilde{B} = \begin{pmatrix} 0 & U - I_n \\ -I_n & A \end{pmatrix} \quad (12)$$

In which 0_n represents a $n \times n$ matrix with all of its elements be zeros. While I_n is the $n \times n$ identity matrix $U = diag(d_i)$ is $n \times n$ diagonal matrix with d_i on the diagonal. It was argued that if a network has K communities then we can find that the first K largest eigenvalues are real valued in magnitude of $\tilde{B}$ [34]. Especially, they were separated from the bulk, which is contained in a circle of radius $\left\|\tilde{B}\right\|^{1/2}$. These K eigenvalues were regarded as the informative eigenvalues of $\tilde{B}$ [35]. Moreover, it was also proved that the spectrum norm of the non-backtracking matrix can be approximated by formula (13)

$$\tilde{d} = \left(\sum_{i=1}^{n} d_i\right)^{-1} \left(\sum_{i=1}^{n} d_i^2\right) - 1 \quad (13)$$

Since the informative eigenvalues of the non-backtracking matrix are real-valued and separated from the bulk of radius $\left\|\tilde{B}\right\|^{1/2}$, we can estimate K by counting the number of real eigenvalues of $\tilde{B}$ that are at least $\left\|\tilde{B}\right\|^{1/2}$. Noted that the parameter learning process performs well especially when the communities of the known

complex network have similar sizes and edge densities.

4. Experiments and Results

In this section, we chose a representative overlapping community detection method called BNMTF as the comparison algorithm [12]. It was argued that outperforms other NMF-based methods. To test the performance of GNMTF and the BNMTF algorithms, we carried out both of the algorithms on several typical network datasets. The applied datasets include two real world networks such as Zachary's Karate Club Network and American College Football Network and one benchmark network. The results show that for our algorithm the communities discovered are more close to the actual situation. Moreover, evaluation index normalized mutual information indicates that the manifold regularization improves the accuracy of the proposed community detection method.

4.1 Real World and Benchmark Networks Testing

4.1.1 Zachary's Karate Club Study

In this subsection, we took the well-known karate club friendship network studied by Zachary as an example. During the course of the Karate club study, a disagreement happened between the administrator and the club's instructor. Which resulted in the instructor's leaving and constructing a new club, taking about a half members of the original club with him. Here we applied our algorithm to this network as an attempt to identify the factions involved in the split of club.

First, we estimated the number of communities by the spectral properties of the relative non-backtracking matrix. According to the background knowledge this network is actually divided into two different communities. The spectrum of the non-backtracking matrix shown in figure 1 confirmed this. There were a compact circular spectral band and two outlying eigenvalues. Then the underlying factor of the NMF based model was determined and set $K = 2$. Next, the communities discovered by the proposed algorithm GNMTF was shown in figure 2. Meanwhile, the communities discovered by the comparison algorithm BNMTF was shown in figure 3.

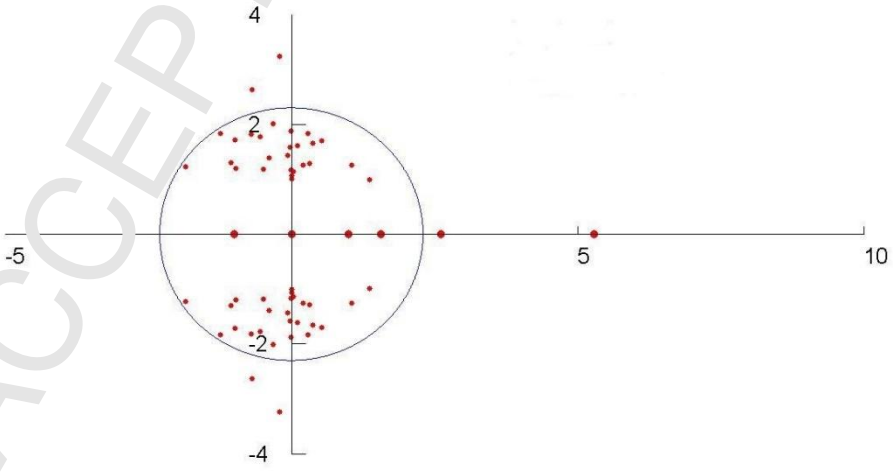

Fig.1. the spectrum norm of the non-backtracking matrix for Zachary's Karate Club Network

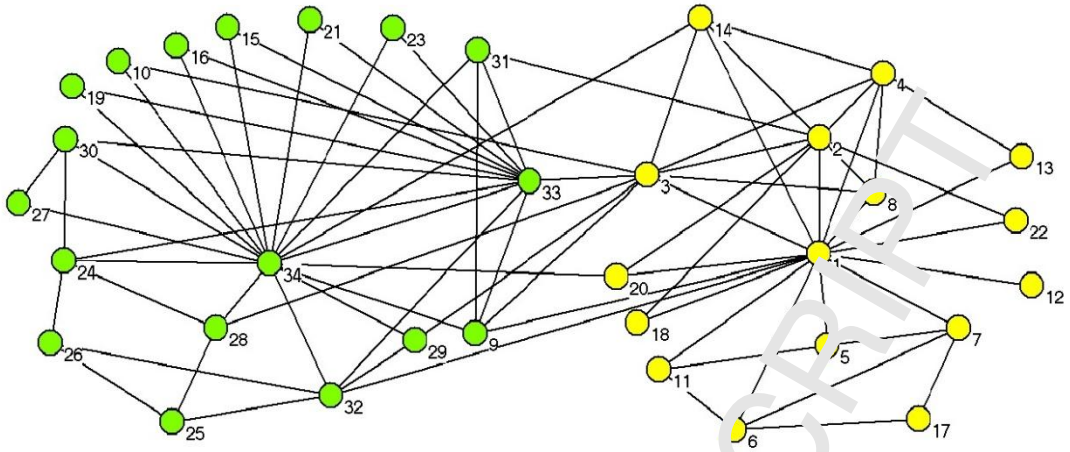

Fig.2. the communities discovered by our proposed algorithm GNMTF

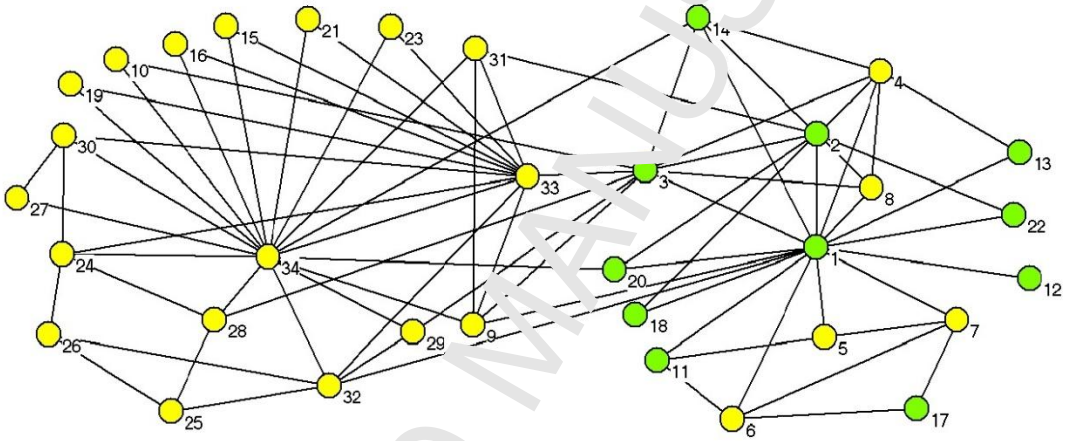

Fig.3. the communities discovered by the comparison algorithm BNMTF

It can be seen that for our proposed algorithm GNMTF the communities discovered were in accordance with the actual situation. It was able to reflect the factions involved in the split of club. While for the comparison algorithm BNMTF, the community detection result was not very well. As figure 3 showed, the Zachary's Karate Club network was also divided into two communities. However, the structural composition for one of the communities was a little different from the actual situation. For example, the nodes 4,5,6,7,8 painted in yellow were in fact belonged to the community represented by green. Noted that for the comparison algorithm the community detection result of the network greatly depended on one of the relevant parameters.

4.1.2. American College Football Network

We apply our algorithm in American College Football network in this subsection. The network represents the game schedule of the 2000 season of Division I of the U.S. college football league. In which nodes represent teams and edges represent regular season games between the two teams they connect. Previous study indicated that it can be divided into 12 communities each containing around 8-12 teams. Note that there exist four independent teams which do not belong to any community.

The same as before, we should first estimate the number of communities by the

spectral properties of the corresponding non-backtracking matrix. The spectrum of the relative non-backtracking matrix showed in figure 4. It indicated that there was about 10 different communities. Accordingly, it can be seen that there were 10 eigenvalues which lay outside of the compact circular spectral band. Consequently, the value of the key parameter the underlying factor of the NMF model was $K = 10$. Then we can use the proposed algorithm GNMTF to discover the community structures. The corresponding result was shown in figure 5. While the community discovering result of the comparison algorithm was shown in figure 6.

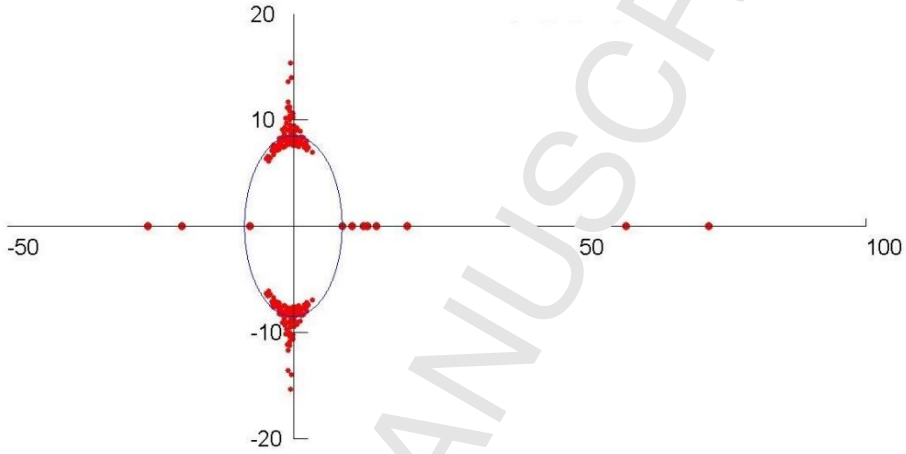

Fig.4. the spectrum norm of the non-backtracking matrix for American College Football Network

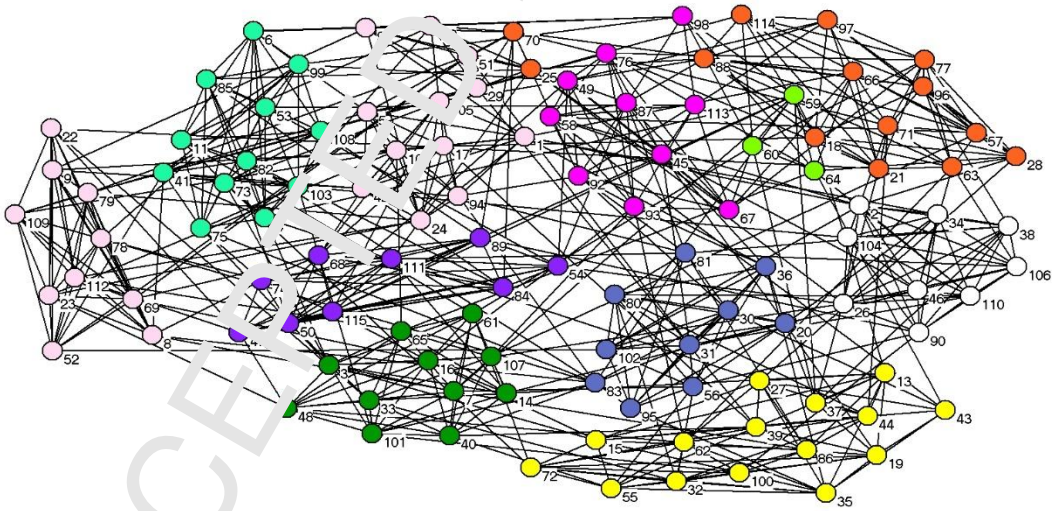

Fig.5. the communities discovered by our proposed algorithm GNMTF

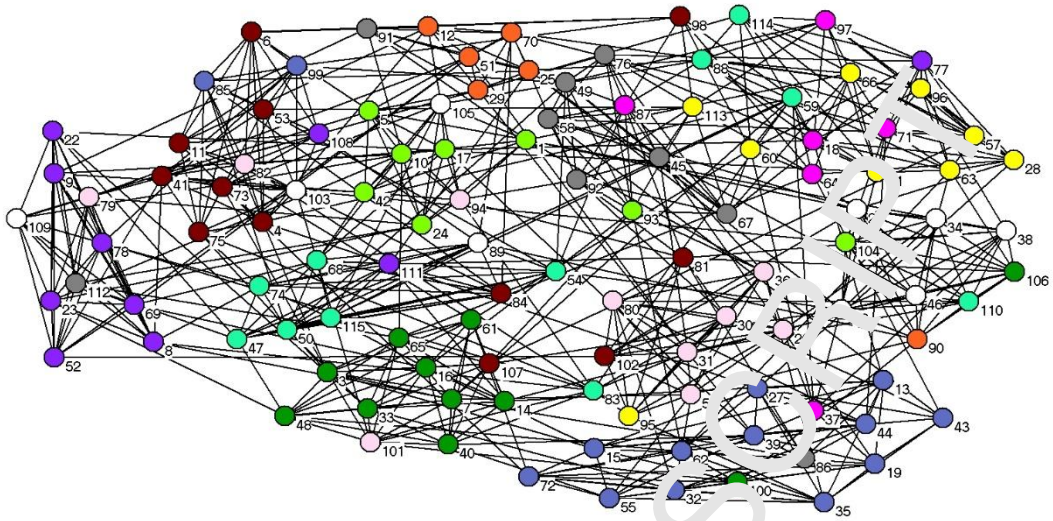

Fig.6. the communities discovered by the comparison algorithm BNMTF

As shown in figure 5 it suggested that in the American College Football network there might be 10 groups rather than 12. Which was a little difference from the group assignment given by the ground truth. Compared with the actual situation, for our method it discovered most of the communities accurately. Noted that the community represented by pink actually contained three different communities in reality. The first constructed by nodes 8,9,22,23,52,69,78,79,109 and 112. The second consisted of nodes 1,5,10,17,24,42,94 and 105. While the third consisted of nodes 12,29,51 and 91. Additionally, the community represented by green including nodes 59,60,64 was in fact belonged to the community represented by orange.

Especially, for the comparative algorithm we set the community number $K=12$ in accordance with the known situation. The community discovering result was shown in figure 6. It can be seen that most of the communities discovered were different from the background truth. For example, the community represented by blue contained nodes 8,9,22,23,52,69,77,78,108 and 111 was differ from the standard result which included nodes 8,9,22,23,52,69,78,79,109 and 112. Similarly, the community represented by white contained nodes 2,16,34,38,46,89,103,105 and 109 was also different from the standard result in the background truth which included nodes 2,16,34,38,46,89,104,106 and 110. With respect to the other similar results, we would not describe here. They were quite distinct in the figure.

4.1.3. Benchmark network

At last, we use LFR-benchmark generator to produce an overlapping network within implanted communities. LFR-benchmark model considered the important features of real world networks such as the fat-tailed distributions of node degree and community size [36]. Then we tested the proposed algorithm on the generated benchmark network. The relative parameters were set as follows $N=150, \langle k \rangle = 8,$ $k_{\max}=25, \mu=0.3, c_{\min}=13, c_{\max}=35, on=5$ and $om=3$.

In which N represents the number of nodes, $\langle k \rangle$ represents the average degree,

k_{max} represents the maximum degree, μ represents the mixing parameter , c_{min} represents the minimum community size while c_{max} represents the maximum community size, on represents the number of overlapping nodes and om represents the number of memberships of the overlapping nodes

Under this circumstances, the partition results of the proposed algorithm and the comparative algorithm both compared with the predefined communities of the above LFR-benchmark. Similarly, we first estimated the number of communities by the spectrum of the relative non-backtracking matrix. It was shown in figure 7 which indicated that the network can be divided into 6 communities. Therefore, the value of the underlying factor of the NMF based model was $k=6$. Noted that in this experiment, we divided the node to the community with the largest membership.

For contrast, figure 8 showed the planted community structure of the LFR-benchmark network. Then the community structures discovered by our proposed algorithm GNMTF was shown in figure 9. While the community discovering result of the comparison algorithm was shown in figure 10.

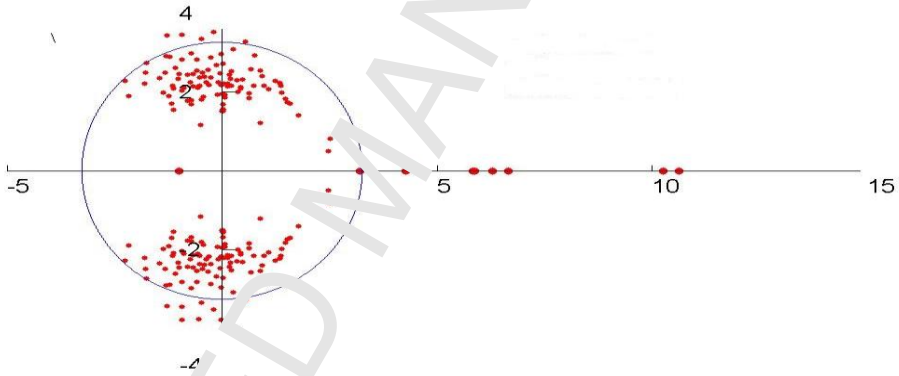

Fig.7. the spectrum norm of the non-backtracking matrix for LFR-benchmark network

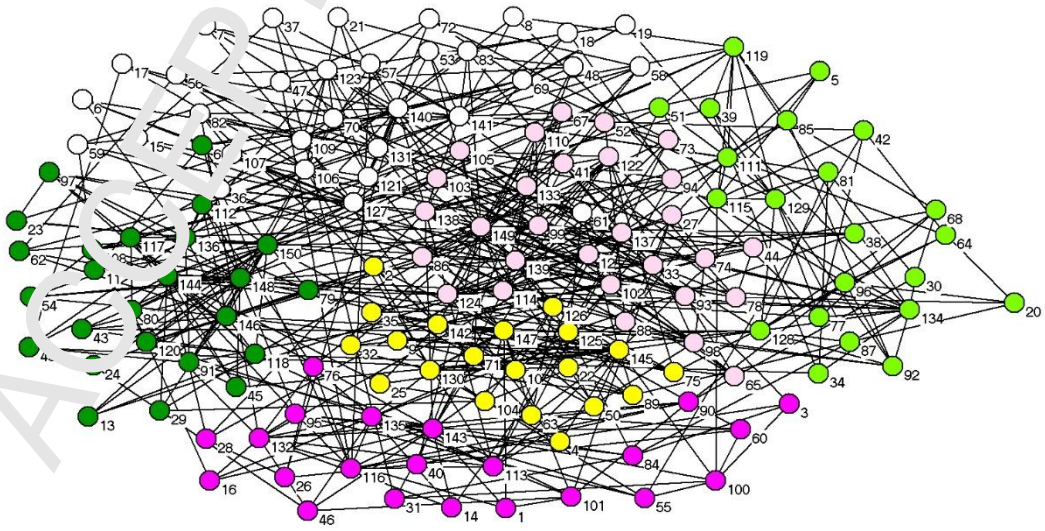

Fig.8. the planted communities of the benchmark network

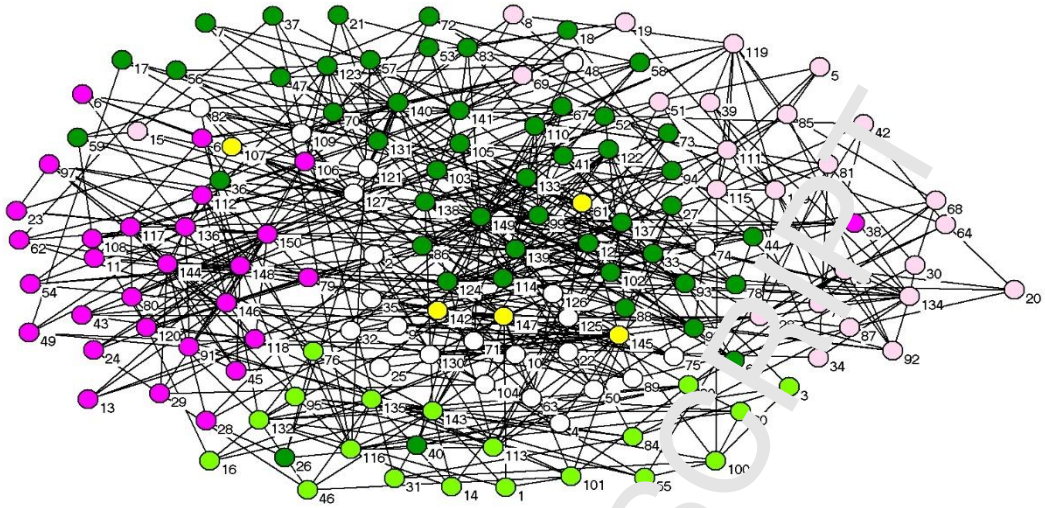

Fig.9. the communities discovered by our proposed algorithm GNMTF

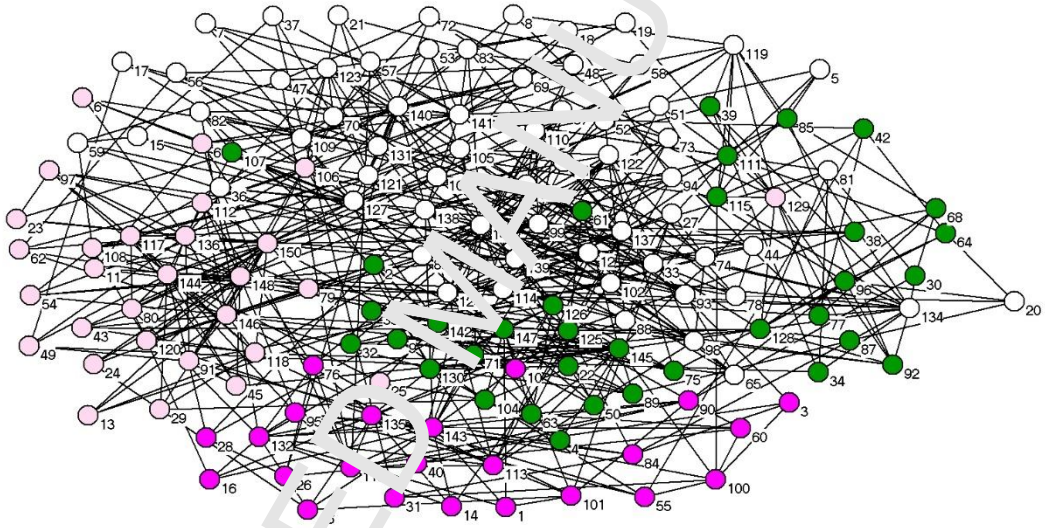

Fig.10. the communities discovered by the comparison algorithm BNMTF

As shown in figure 9 it suggested that for our method the communities discovered basically met the predefined community structures. The most obvious difference was the largest community which was painted in green. It contained nodes 7,12,17,18,21,26,27,33,36,37,40,41,44,47,52,53,56,57,58,59,65,67,70,72,73,78,83,86, 88,93,94,98,99,102,103,105,110,114,122,123,124,131,133,137,138,139,140,141 and 149. Actually, it matched with most of the nodes painted in white and all the nodes painted in pink of the planted community structures. While the rest community structures were almost in accordance with the predefined community structures.

Noted that for the comparison algorithm according to the LFR model we set the community number $K = 6$, but the actual community discovering result had four communities due to the memberships of certain communities were zeros. From figure 10 it can be seen that the community discovering results were similar as ours in structure. The community painted in white contained the two communities of the predefined community structures represented by white and pink. Besides, for our method the community represented by dark green which were constructed by nodes

2,4,9,22,30,32,34,35,38,39,42,50,61,63,64,68,71,75,77,85,87,89,92,96,104,107,111,115,126,128,130,142,145 and 147 included nearly most of the nodes in the communities painted in yellow and green of the LFR benchmark network within planted community structures. Except for nodes 81, 129 painted in green and nodes 10,25 painted in yellow.

4.2. Criterion for Accuracy Evaluation

We compared the quality of the proposed algorithm with the chosen representative comparison algorithm by calculating the normalized mutual information. It is used extensively in measuring the performance of clustering algorithms. It is argued that the larger the NMI value is, the better the community discovering result is. The formal definition of NMI is as formula (14) shown [37]

$$NMI(X,Y)=\frac{-2\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}C_{ij}\log\frac{C_{ij}C}{C_{i.}C_{.j}}}{\sum_i C_{i.}\log\frac{C_{i.}}{C}+\sum_j C_{.j}\log\frac{C_{.j}}{C}} \tag{14}$$

In which X represents the community structures under the actual situation. While Y represents the community structures discovered by certain algorithm. N_X and N_Y are the community numbers of X and Y respectively. Especially, C represents the confusion matrix. The physical meaning of C_{ij} is that the number of vertices which are divided in the community j discovered by certain algorithm, but actually should be assigned to the ground truth community i. $C_{i.}$ is the sum over row i of C and $C_{.j}$ is the sum over column j of C. Note that for overlapping community discovering, we use the generalized normalized mutual information.

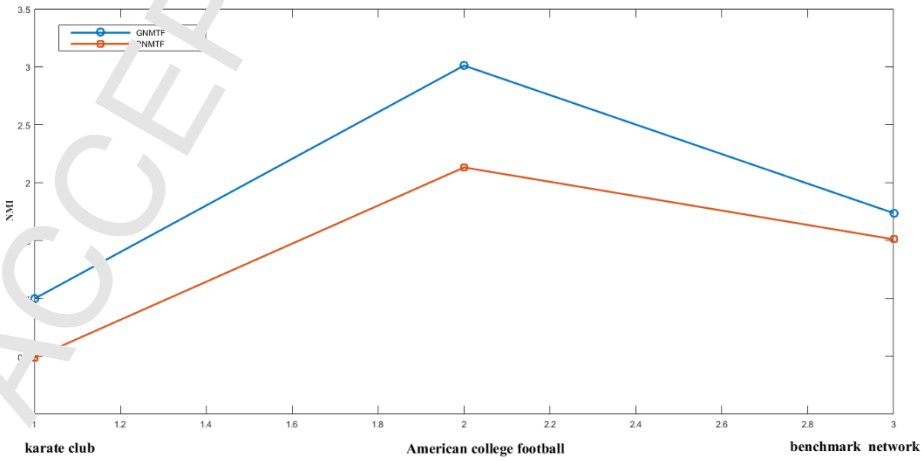

Fig.11. the NMI value of different networks for our method and the comparative algorithm

As for the evaluation of community detection methods, NMI indicated the similar degree between the discovered community structure and the actual situation. The larger the NMI value is, the closer the detected community structure to the true partition. For the above two real world networks and one benchmark network, the NMI values were computed for both our method and the comparative algorithm as figure 11 shown. It can be seen that the proposed method can achieve a higher NMI values on the above three representative networks. Moreover, for the comparison algorithm its community detection results on some networks rely too much on the preset value of relative parameters.

5. Conclusions and Future Works

While recent advances have made nonnegative matrix factorization for community detection far more scalable than in the past, NMF based algorithms are highly competitive because of its interpretability and applicability when used to discover overlapping community structure. However, most of the NMF based methods investigated community detection problem from one of the two perspectives. One is how to construct the similarity matrix to be decomposed. The other is how to make the NMF based model more efficiency. They indeed can improve the performance of the algorithms. Inspired by the recent progress in matrix factorization and manifold learning, we made a new attempt. In this paper, through the consideration of the intrinsic geometric information of the network graph represented by the low-dimensional manifold, we found that it could help achieve better community detection results. Moreover, with the help of spectrum properties analysis of relative non-backtracking matrix for certain complex network, it is a good guide to the number of communities. While in the future work, both the theories and the empirical evidences of how to make the graph regularized item work well with NMF based model deserves much work.

Acknowledge

This work was supported by the National Natural Science Foundation of China(No.61272109, 61502350).

References

[1]M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences 99(12) (2002) 7821–7826.

[2]M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E (2004) 69026113.

[3]M.E.J. Newman, Networks: An Introduction, New York: Oxford University Press (2010).

[4]S. Fortunato, Community detection in graphs, Phys. Rep.486 (2010) 75-174.

[5]A.L. Barabasi, Network Science, www.barabasilab.neu.edu/networksciencebook (2012).

[6]J. Wu, X. Zhu, C. Zhang, et al, Bag Constrained Structure Pattern Mining for Multi-Graph Classification, IEEE Transactions on Knowledge & Data Engineering 26(10)(2014) 2382-2396.

[7]J. Wu, S. Pan, X. Zhu, et al, Boosting for Multi-Graph Classification, IEEE Transactions on Cybernetics 45(3)(2015) 416-429.

[8] J. Wu, S. Pan, X. Zhu, et al, Multiple Structure-View Learning for Graph Classification, IEEE Transactions on Neural Networks & Learning Systems 29(7)(2018) 3236-3251.

[9] J. Wu, S. Pan, X. Zhu, et al, Positive and Unlabeled Multi-Graph Learning, IEEE Transactions

on Cybernetics 47(4)(2016) 818-829.
[10]S. Gregory, Finding overlapping communities in networks by label propagation, New J Phys 12(10) (2010) 103018.
[11]D.X. He, D. Jin, C. Baquero, et al., Community detection using generative model and nonnegative matrix factorization, PloS One9(1) (2014) e86899.
[12]Y. Zhang, D.Y. Yeung, Overlapping community detection via bounded nonnegative matrix tri-factorization, Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (2012) 606-614.
[13]N. Zhao, L. Zhang, B. Du, et al, Robust Dual Clustering with Adaptive Manifold Regularization, IEEE Transactions on Knowledge & Data Engineering pp(99)(2017) 2498-2509.
[14]L. Zhang, Q. Zhang, L.P. Zhang, et al, Ensemble Manifold Regularized Sparse Low-Rank Approximation for Multiview Feature Embedding, Pattern Recognition 48(10)(2015) 3102-3112.
[15]N. Guan, T. Liu, Y. Zhang, et al, Truncated Cauchy Non-negative Matrix Factorization for Robust Subspace Learning, IEEE Transactions on Pattern Analysis & Machine Intelligence PP(99)(2018) 1-14.
[16]I. Psorakis, S. Roberts, M. Ebden, B. Sheldon, Overlapping community detection using Bayesian non-negative matrix factorization, Physical Review E (2011) 83.
[17]X.C. Cao, X. Wang, D. Jin, et al., Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization, Scientific Reports 3 (2013) 2993.
[18]D. Cai, X. F. He, J. W. Han, T. S. Huang, Graph regularized non-negative matrix factorization for data representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8) (2011) 1548–1560.
[19]M. Belkin, P. Niyogi, V. Sindhwani, Manifold Regularization: A Geometric Framework for Learning from Examples, J. Machine Learning Research7 (2006) 2399-2434.
[20]D. Cai, X. He, X. Wu, J. Han, Non-negative Matrix Factorization on Manifold, Proc. Int'l Conf. Data Mining (2008).
[21]C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix tri-factorizations for clustering, ACM SIGKDD 12th (2006) 126-135.
[22]D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems 13 (2000) 556–562.
[23]F. Wang, T. Li, X. Wang, S. Zhu, C. H. Q. Ding, Community discovery using nonnegative matrix Factorization, Data Mining and Knowledge Discovery 22(3) (2011) 493–521.
[24]Z.Y. Zhang, Y. Ahn, Community detection in bipartite networks using weighted symmetric binary matrix factorization, Internat. J. Modern Phys. C 26 (09) (2015) 1550096.
[25]Z.Y. Zhang, Y. Wang, Y.Y. Ahn, Overlapping community detection in complex networks using symmetric binary matrix factorization, Physical Review E 87(6-1) (2013) 062803.
[26]F. Krzakala,C. Moore, E. Mossel, J. Neeman, et al., Spectral redemption in clustering sparse networks, Proceedings of the National Academy of Sciences110(52) (2013) 20935–20940.
[27]C.M. Le, E. Levina, Estimating the number of communities in networks by spectral methods, Computer Science (2015) 769-774.
[28]P.O. Hoyer, Non-Negative Matrix Factorization with Sparseness Constraints, J. Machine Learning Research 5 (2004) 1457-1469.
[29]S. Gao, I. Tsang, L. Chia, Laplacian Sparse Coding, Hypergraph Laplacian Sparse Coding, and Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence35(1) (2013)

92-104.
[30]R.S. Wang, S.H. Zhang, Y. Wang, et al., Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures, Neurocomputing72(1/3) (2008) 134-141.
[31]K. Rohe, S. Chatterjee, B.Yu, Spectral clustering and the high-dimensional stochastic block model, Annals of Statistics 39(4) (2011) 1878–1915.
[32]A. Saade, F. Krzakala, L. Zdeborov´a, Spectral density of the non-backtracking operator on random graphs, EPL 107(5) (2014) 50005.
[33]R. Fitzner, R. vander Hofstad, Non-backtracking random walk. J. Stat. Phys. 150 (2013) 264–284.
[34]A. Saade, F. Krzakala, L. Zdeborová, Spectral Clustering of Graphs with the Bethe Hessian, Advances in Neural Information Processing Systems 1 (2014) 406-414.
[35]C. Bordenave, M. Lelarge, L. Massoulie, Non-backtracking Spectrum of Random Graphs: Community Detection and Non-regular Ramanujan Graphs, IEEE 56th Annual Symposium on Foundations of Computer Science (2015) 1347-1357.
[36]A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E80 (2009) 016118.
[37]A.F. Mcdaid, D. Greene, N. Hurley, Normalized Mutual Information to evaluate overlapping community finding algorithms, Computer Science (2011).