Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5750645, 15 pages
http://dx.doi.org/10.1155/2016/5750645

Research Article

Efficient Community Detection in Heterogeneous Social Networks

Zhen Li, Zhisong Pan, Yanyan Zhang, Guopeng Li, and Guyu Hu

College of Command Information Systems, PLA University of Science & Technology, Nanjing, Jiangsu, China

Correspondence should be addressed to Guyu Hu; huguyu@189.cn

Received 13 June 2016; Revised 21 September 2016; Accepted 23 October 2016

Academic Editor: Michael Small

Community detection is of great importance which enables us to understand the network structure and promotes many real-world applications such as recommendation systems. The heterogeneous social networks, which contain multiple social relations and various user generated content, make the community detection problem more complicated. Particularly, social relations and user generated content are regarded as link information and content information, respectively. Since the two types of information indicate a common community structure from different perspectives, it is better to mine them jointly to improve the detection accuracy. Some detection algorithms utilizing both link and content information have been developed. However, most works take the private community structure of a single data source as the common one, and some methods take extra time transforming the content data into link data compared with mining directly. In this paper, we propose a framework based on regularized joint nonnegative matrix factorization (RJNMF) to utilize link and content information jointly to enhance the community detection accuracy. In the framework, we develop joint NMF to analyze link and content information simultaneously and introduce regularization to obtain the common community structure directly. Experimental results on real-world datasets show the effectiveness of our method.

1. Introduction

The past few years witnessed the emergence and popularity of online social media. Millions of users participate in online social media such as Twitter and Facebook, making up many social networks. Then social network analysis attracts enormous attention and community detection is one of the fundamental tasks in this field. A community can be defined as a group of users that (1) interact with each other more frequently than with those outside the group and (2) are more similar to each other than to those outside the group [1]. The research on community detection is beneficial for a variety of real-world applications such as online marketing and recommendation systems.

Many community detection methods utilizing only one type of social relation are developed [2–4]; however, one type of relation can only provide limited and incomplete information for detecting communities [1, 5]. In real-world social media, the networks are often heterogeneous containing multiple types of social relations and user generated content [5–7], combining the social relations and content information is a better strategy for community detection [1, 8–13]. For example, Facebook has different interactions among the same set of users, users can be friends with each other, and a user can also follow someone. These interactions among users are regarded as link information and modeled as graph [6, 14, 15]. There is also user generated content in Facebook: users can post photos and share articles and other things they are interested in, and then users are related to certain items. The relations between user and content are regarded as content information [9, 14] and represented by user-content matrix. Figure 1 is an example of heterogeneous social network containing various link information and content information. Figure 1(a) shows three types of link information among the same set of users. Each type of relation provides part of topological information of the whole social network. Figure 1(b) shows the content information; users are related to various items; these relations reflect users' intrinsic features. Based on the definition of community, we know that both users interacting closely and users with high similarity tend

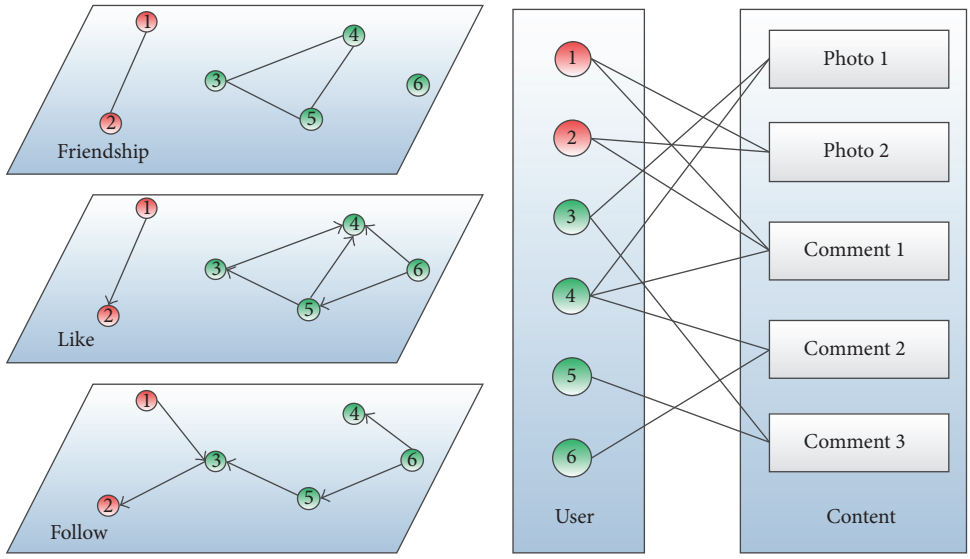

Figure 1: Example of heterogeneous social network which contains different types of link information and content information. Different colors of nodes represent different communities. (a) The link information: three types of social relations, that is, friendship, like relation, and follow relation. (b) The content information: user-content relations; there are two types of content, that is, photos and comments.

to be in the same community. In heterogeneous social networks, different types of link information reveal the network topology from different perspectives, and various content information implicates user similarity. Furthermore, network topology and user similarity indicate a common community structure, which provide complementary information from different views for mining the common community structure [9, 11, 13, 16]. Thus, researchers turn to combine the two kinds of information to make the community detection results more reliable. Note that the research is based on an assumption that different information indicates a common community structure [1, 8–12].

Community detection in heterogeneous networks has several challenges. First, there exists some noise in link data; how to handle the noisy data to extract the accurate community structure is challenging. Moreover, the link information (represented by adjacency matrix) and the content information (represented by user-content matrix) are different types of data, and how to combine them efficiently and effectively is a complicated problem. Some methods are proposed to detect the communities in heterogeneous networks. Gu and Zhou [8] propose a dual regularized coclustering method based on seminonnegative matrix factorization (semi-NMF) for community detection, where a feature graph is constructed based on user similarity. The clustering method can explore the geometric structure of both link manifold and content manifold by coclustering. Pei et al. [1] propose nonnegative matrix trifactorization (NMTF) based clustering framework, in which user similarity, message similarity, and user interaction are captured explicitly from user generated content to improve the accuracy of community detection. Hidru and Goldenberg [10] develop a graph regularized multiview NMF-based method for data integration, in which the parameters are automatically learnt from data. Ni et al. [13] discuss the case in which multiple clustering structures across different networks are allowed and construct a network of networks based on the domain similarity to regularize the clustering structure in different networks. Hu et al. [17] develop a scalable method which can handle any large number of networks; it identifies recurrent patterns across multiple networks to discover biological modules. In the method, different networks are compressed into two metagraphs and clustering is performed in these two graphs. Lukk et al. [18] construct a global gene expression map by integrating microarray data from different cell and tissue types, allowing researchers to do further researches. Tang et al. [6] review four existing strategies for integrating multiple relational information to find out the shared community structure: network integration, utility integration, feature integration, and partition integration. Ruan et al. [14] first calculate the content similarity and combine link strength with content similarity to construct the final backbone graph and then use community detection algorithm for single view on the backbone graph to derive the common structure.

Although many algorithms have developed, community detection in heterogeneous networks remains an issue to be addressed. The existing methods have some shortcomings: (1) the methods in paper [6, 13, 17, 18] are developed to integrate the multiple link information for community detection and cannot deal with content information. (2) The strategies in paper [8, 14] first turn the content information into link information and then process the constructed network. However, compared with processing the user-content matrix directly, preprocessing adds extra time consumption. Besides, constructing network according to content information is usually based on k-nearest neighbor sets; the best value of

parameter k has to be learnt. (3) The methods in paper [1, 8] concentrate on dealing with one type of content and one type of link information; it is hard to extend them to deal with multiple social relations; besides, the constraints introduced in paper [1, 10] are so strict that the methods have poor performances on some datasets. (4) Moreover, in most methods, the individual community structures of different views are detected simultaneously; then one of the detection results is chosen as the common community structure [1, 9, 11, 13, 17]. Although community structures of different views are restricted to be similar, each individual view contains some private components [19] which make the detected individual result not consistent with the common structure. Furthermore, to choose which result as the common structure is also an unsolved problem.

In this paper, we try to develop a community detection framework in heterogeneous social networks, to derive the common community structure by utilizing multiple link information and multiple content information. Particularly, we derive content information from matrix factorization rather than constructing similarity network and derive the common structure directly from combining multiple link information and content information rather than taking one of individual results as the common structure. Specifically, we propose a regularized joint NMF (RJNMF) based community detection framework. Our contributions are summarized as follows:

(1) We investigate the community detection problem in heterogeneous networks and develop a framework to deal with multiple link information and content information simultaneously; the content information can be processed without being turned into link information, and the framework is simple and effective.

(2) We introduce regularization into the optimization function to control the similarity of community structures explored from different data sources; thus, the effects of the noise can be reduced. Furthermore, the common community structure can be derived directly by obtaining the common community indicator from solving the optimization problem; although the individual result of each view can be also detected with the assistance of other views, we do not take the individual result as the common community structure.

(3) We carry out experiments on the real-world datasets to evaluate the effectiveness of the proposed method.

The rest of this paper is organized as follows: Section 2 includes a brief review of related work. In Section 3, we introduce the basic problem definition and related notions; then we describe our method in detail. Experimental results on real-world data are presented in Section 4. Finally, some conclusions are given in Section 5.

2. Related Work

Community detection using link and content information has been studied for years. The methods are mainly classified into two categories. In the first category, the methods deal with the link and content information in the same way. Gu and Zhou [8] proposed a regularized coclustering method based on semi-NMF to find community structures. In the method, the content information was turned into link information, and two types of regularization were introduced to explore geometric structures, requiring the cluster labels of data points to be smooth with respect to the link manifold, while the cluster labels of features are smooth with respect to the content manifold. However, the method can only deal with one type of link information. He et al. [9] extended NMF for multiview clustering by jointly factorizing the multiple matrices through coregularization and proposed a coregularized NMF framework for combining multiple content information without turning content information into link information. In the framework, pairwise regularization and cluster-wise regularization were developed to enforce similarity on different views. Pei et al. [1] proposed a clustering framework by integrating NMTF with three types of graph regularization. In the framework, user similarity, message similarity, and user interaction were captured to contribute to community detection. However, the constraints put on structures of different views were strict, resulting in bad performance on some datasets. Tang et al. [11] presented a joint NMF optimization framework to integrate multiple views. The components were separately constrained with L_1-norm regularization or Frobenius norm regularization to ensure sparsity and accuracy. However, the structures derived from multiple data sources were not restricted to be similar. Cheng et al. [12] discussed the multiview clustering problem in a complex scenario where users in different domain might not match and proposed a robust framework which allowed partial mapping and could handle graphs of different sizes. Cheng et al. [20] mined the common structure across multiple views by concatenating the low-rank matrices; particularly, they sought the sparsity-consistent low-rank affinities from the joint decomposition of multiple feature matrices into pairs of sparse and low-rank matrices, and the noisy data was removed by introducing $L_{2,1}$ norm. However, the method which was developed for image segmentation could not be used for social networks because the link data is neither sparse or low-rank. Guo et al. [21] enhanced codetection by extracting a shared low-rank representation of the object instances in multiple feature spaces. The representation was based on a linear reconstruction over the entire data set and the low-rank approach enables effective removal of noisy and outlier samples. The extracted low-rank representation could be used to detect the target objects by spectral clustering. However, the method only has good performance on content information; when it is used to deal with link information, the performance is poor. Xia et al. [22] proposed a shared transition probability matrix of the multiple views to conduct spectral clustering. They firstly constructed a transition probability matrix from each single view and then used these matrices to recover a shared low-rank transition probability matrix as a crucial input to the standard Markov chain method for clustering. Deng et al. [19] established a framework to capture both common components of all the views and private components of individual views. They decomposed an input data matrix concatenated

from multiple views as the sum of low-rank, sparse, and noisy parts. Then a unified optimization framework was established, where the low-rankness and group-structured sparsity constraints were imposed to simultaneously capture the shared and private components in both instance and view levels. However, because the link data is neither low-rank or sparse, the methods in paper [19, 22] perform poor on link data. Nguyen et al. [23] discussed community detection in multiplex social networks in which a user can have multiple accounts. They developed a unifying approach which aggregated multiple accounts of the same users and a coupling approach which used coupling techniques to find a consistent community structure. However, the method only considered the relations between users while neglecting the content information. Mahmood and Small [24] developed a community detection algorithm which was fundamentally different from most existing methods based on graph theoretics. The algorithm was based on the fact that each network community spanned a different subspace in the geodesic space, and each node was efficiently represented as a linear combination of nodes spanning the same subspace. Sparse linear coding with L_1 norm constraint was used to make the detection process more robust. The algorithm showed excellent performance on both benchmark and real-world networks. It is promising to extend this method to detect communities in heterogeneous networks. Guesmi et al. [25] proposed a method to deal with multiple types of objects and relationships derived from a bibliographic networks. The approach first constructed the Relation Context Family (RCF) to represent the different objects and relations using the relational concept analysis methods and then explored such RCF for community detection.

In the other category, the methods try to turn the content information into link information according to some rules. Ruan et al. [14] analyzed the community signal strength between nodes in the social network by fusing the link strength with content similarity. Content similarity was first estimated through cosine similarity or Jaccard coefficient; then the final network based on link information and similarity was constructed, and community detection algorithm for single view was used to analyze the network. Greene and Cunningham [15] proposed to produce a single unified graph based on the combination of k-nearest neighbor sets for users. With regard to content information, the neighbor was gained by estimating content similarity. However, the complexity is $O(n^2)$, where n is the number of nodes. Compared with processing the user-content matrices directly to derive community structure, turning content information into link information adds extra time consumption.

There is also ensemble method proposed. Zheng et al. [26] proposed a NMF-based ensemble clustering method which combined multiple clustering results into a single consolidated partition. In the method, community detection was implemented on each data source separately; then the detection results were integrated with different weights to capture the common community structure. Liu et al. [27] proposed spectral ensemble clustering which employed spectral clustering on the coassociation matrix to find the consensus partition. The method could deal with large scale datasets and had a good performance; moreover, it was robust to incomplete basic partitions with many missing values.

Table 1: Main notations.

Notations	Meaning	Dimension
n	Number of nodes	—
m	Number of content items	—
p/q	Number of adjacency/user-content matrices	—
k	Number of communities	—
$A^{(t)}$	The tth adjacency matrix	$n \times n$
$X^{(t)}$	The tth user-content matrix	$n \times m$
$U^{(t)}$	The individual community indicator matrix	$n \times k$
$W^{(t)}$	The individual community indicator matrix	$n \times k$
$H^{(t)}$	The basis matrix	$m \times k$
S	The common community indicator matrix	$n \times k$

In this paper, we develop our method based on NMF. NMF [28] is a popular factorization method where all the elements related are restricted to be nonnegative; it is widely used to model social networks and cluster users into communities [8, 23, 29], and it is extended to multiview clustering [1, 8–11]. Its core idea is to approximate a higher dimensional matrix with nonnegative lower dimensional matrices. The factorization can be described as $\|A - WH^T\|_F^2$, in social network analysis field; $W^{n\times k}$ is regarded as community indicator matrix which implicates the community membership [23, 29]. Here, we extend NMF with proper regularization to integrate link and content information for community detection and derive the common community structure by obtaining the common community indicator from solving the optimization problem.

3. Proposed Framework

In this section, we first introduce symbols used in the paper and then propose the framework for detecting common communities directly utilizing link and content information.

3.1. Problem Statement. The main notations used in this paper are listed in Table 1. Given a set of n nodes in the heterogeneous network, link information is represented by the adjacency matrix which describes the social relations between users and content information is represented by the user-content matrix which describes the relations between users and content. $A^{(t)}$ is tth adjacency matrix and $X^{(t)}$ is tth user-content matrix, p and q denote the number of adjacency and user-content matrices, respectively, and the nodes are grouped into k communities. $U^{(t)}$ and $W^{(t)}$ are individual community indicator matrices decomposed from $A^{(t)}$ and $X^{(t)}$, respectively. We use S to denote the common community indicator matrix for deriving the common community structure in the heterogeneous network. Particularly, indicator matrix represents the membership a node with respect to k communities [23, 29]. For example, S_{ij} ($S_{ij} \in S$) shows the degree of ith node belonging to the jth community.

Normally, the ith node belongs to the mth community if S_{im} is the max value of $S(i,:)$.

With the notations given in Table 1, the community detection problem in this paper is formally defined below.

Problem 1. Given adjacency matrices $\{A^{(1)}, \ldots, A^{(p)}\}$ and user-content matrices $\{X^{(1)}, \ldots, X^{(q)}\}$, find out k communities that (1) with maximal links within community and minimal links across communities in terms of link based views and (2) members are more similar to each other than to those outside the community in terms of content based views.

We first investigate how to combine the multiple matrices for community detection jointly. NMF has been shown useful in clustering problem [8, 29]. The 2-factor factorization NMF can be written as follows:

$$\begin{aligned} \min_{W \in R^{n\times k}, H \in R^{m\times k}} \quad & \left\| X - WH^T \right\|_F^2 \\ \text{s.t.} \quad & W_{ij} \geq 0, \\ & H_{ij} \geq 0, \\ & \forall i, j. \end{aligned} \tag{1}$$

In paper [1, 9, 13], NMF was extended for multiview clustering by establishing the joint optimization model. Given T data sources, the joint NMF can be presented as below:

$$\begin{aligned} \min_{W^{(t)} \in R^{n\times k}, H^{(t)} \in R^{m\times k}} \quad & \sum_{t=1}^{T} \lambda_t \left\| X^{(t)} - W^{(t)} \left(H^{(t)}\right)^T \right\|_F^2 \\ \text{s.t.} \quad & W_{ij}^{(t)} \geq 0, \\ & H_{ij}^{(t)} \geq 0, \\ & \forall i, j, \end{aligned} \tag{2}$$

where λ_t is the parameter to control the weight of each data source.

Since the multiple adjacency matrices and user-content matrices indicate a common underlying structure, similar indicator matrices are expected to be learnt. By introducing different regularization, various NMF-based integration methods have been developed. In paper [9], similarity constraints are imposed on each pair of indicator matrices; specifically, the regularization is defined as follows:

$$\min \quad \sum_{s=1}^{T} \sum_{t=1}^{T} \lambda_{st} \left\| W^{(s)} - W^{(t)} \right\|_F^2, \tag{3}$$

where λ_{st} is the parameter of each view pair. Based on the definition above, community detection in heterogeneous networks is a joint optimization problem, where the individual community indicators of multiple link and content data are learnt simultaneously. But the above methods just restrict the individual indicator matrices of different views to be similar and fail to find out the common indicator across multiple data. The popular detection method is to choose one of the individual detection results as the common community structure [1, 9, 11, 13, 17]; however, because each view contains some private components, the individual detection results are not completely consistent with the common structure; besides, the detection results of different views also have some differences, and how to choose a best result is an unsolved problem.

3.2. Proposed Method. In this subsection, the proposed community detection framework based on regularized joint NMF is presented. Given a set of n nodes in the network, the link and content data are represented with $\{A^{(1)}, \ldots, A^{(p)}\}$ and $\{X^{(1)}, \ldots, X^{(q)}\}$, respectively. The joint optimization problem is defined as follows:

$$\begin{aligned} \min_{U^{(t)}, W^{(t)} \in R^{n\times k}, H^{(t)} \in R^{m\times k}} \quad & \sum_{t=1}^{p} a_t \left\| A^{(t)} - U^{(t)} \left(U^{(t)}\right)^T \right\|_F^2 \\ & + \sum_{t=1}^{q} b_t \left\| X^{(t)} - W^{(t)} \left(H^{(t)}\right)^T \right\|_F^2 \\ \text{s.t.} \quad & U_{ij}^{(t)} \geq 0, \\ & W_{ij}^{(t)} \geq 0, \\ & H_{ij}^{(t)} \geq 0, \\ & \forall i, j, \end{aligned} \tag{4}$$

where a_t and b_t are parameters to control the contributions of different data sources. Particularly, the adjacency matrix can be weighted and symmetric. With regard to directed networks, the symmetrized graph can be gained by $A' = A + A^T$.

Based on the joint optimization framework developed above, we try to deal with the link and content information simultaneously. As the column vector of the indicator matrix W represents a cluster, when we adopt the vector-based L_2 norm, each entry of W^TW gives the cosine similarity between two clusters [9]; then W^TW can be interpreted as the cluster similarity matrix. In order to reduce the effects of noise and obtain the common community indicator, the individual indicators are constrained to be similar to the common indicator. Here, we introduce a consensus similarity regularization, which minimizes the disagreement between the similarity matrices for individual indicators and common indicator. It is defined as follows:

$$\min \quad \left\| \left(W^{(t)}\right)^T W^{(t)} - S^T S \right\|_F^2, \tag{5}$$

where S is the common community indicator. Incorporating the consensus similarity regularization into the joint NMF process, the proposed method can be presented as follows:

$$\begin{aligned} \min_{U^{(t)}, W^{(t)}, S \in R^{n\times k}, H^{(t)} \in R^{m\times k}} \quad & \sum_{t=1}^{p} a_t \left\| A^{(t)} - U^{(t)} \left(U^{(t)}\right)^T \right\|_F^2 \\ & + \sum_{t=1}^{q} b_t \left\| X^{(t)} - W^{(t)} \left(H^{(t)}\right)^T \right\|_F^2 \end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{p} c_t \left\| \left(U^{(t)}\right)^T U^{(t)} - S^T S \right\|_F^2 \\
&+\sum_{t=1}^{q} d_t \left\| \left(W^{(t)}\right)^T W^{(t)} - S^T S \right\|_F^2 \\
\text{s.t.}\quad & U_{ij}^{(t)} \geq 0, \\
& W_{ij}^{(t)} \geq 0, \\
& H_{ij}^{(t)} \geq 0, \\
& S_{ij} \geq 0, \\
& \forall i, j,
\end{aligned} \tag{6}
$$

where c_t and d_t are parameters to control the weights of regularization. In the framework, we can obtain the common community indicator by solving the optimization problem (6); then the common community structure in the heterogeneous network is derived directly.

3.3. Optimization. Since the objective function formula (6) is not convex, the optimal solution can be achieved using the iterative updating algorithm. To enforce the nonnegativity constraints, we need to incorporate Lagrange multipliers. Let $\alpha^{(t)}, \beta^{(t)}, \omega^{(t)}, \gamma$ be the Lagrange matrices for constraints $U^{(t)} \geq 0$, $W^{(t)} \geq 0$, $H^{(t)} \geq 0$, $S \geq 0$, respectively. Then, the Lagrangian function is as follows:

$$
\begin{aligned}
L = &\sum_{t=1}^{p} \left(a_t \left\| A^{(t)} - U^{(t)} \left(U^{(t)}\right)^T \right\|_F^2 \right. \\
&\left. + c_t \left\| \left(U^{(t)}\right)^T U^{(t)} - S^T S \right\|_F^2 \right) \\
&+ \sum_{t=1}^{q} \left(b_t \left\| X^{(t)} - W^{(t)} \left(H^{(t)}\right)^T \right\|_F^2 \right. \\
&\left. + d_t \left\| \left(W^{(t)}\right)^T W^{(t)} - S^T S \right\|_F^2 \right) - \alpha^{(t)} U^{(t)} - \beta^{(t)} W^{(t)} \\
&- \omega^{(t)} H^{(t)} - \gamma S = L_1 + L_2 - \alpha^{(t)} U^{(t)} - \beta^{(t)} W^{(t)} \\
&- \omega^{(t)} H^{(t)} - \gamma S,
\end{aligned} \tag{7}
$$

where

$$
\begin{aligned}
L_1 = &\sum_{t=1}^{p} \left(a_t \operatorname{tr} \left(A^{(t)} \left(A^{(t)}\right)^T - 2U^{(t)} \left(U^{(t)}\right)^T \left(A^{(t)}\right)^T + \left(U^{(t)} \left(U^{(t)}\right)^T\right)^2 \right) \right. \\
&\left. + c_t \operatorname{tr} \left(\left(\left(U^{(t)}\right)^T U^{(t)} \right)^2 + \left(S^T S\right)^2 + 2 \left(U^{(t)}\right)^T U^{(t)} S^T S \right) \right), \\
L_2 = &\sum_{t=1}^{q} \left(b_t \operatorname{tr} \left(X^{(t)} \left(X^{(t)}\right)^T + W^{(t)} \left(H^{(t)}\right)^T H^{(t)} \left(W^{(t)}\right)^T - 2W^{(t)} \left(H^{(t)}\right)^T \left(X^{(t)}\right)^T \right) \right. \\
&\left. + d_t \operatorname{tr} \left(\left(\left(W^{(t)}\right)^T W^{(t)} \right)^2 + \left(S^T S\right)^2 + 2 \left(W^{(t)}\right)^T W^{(t)} S^T S \right) \right).
\end{aligned} \tag{8}
$$

Then, the derivatives of L with respect to $U^{(t)}, W^{(t)}, H^{(t)}, S$ are

$$
\begin{aligned}
\frac{\partial L}{\partial U^{(t)}} = &-4a_t A^{(t)} U^{(t)} + 4a_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} \\
&+ 4c_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} - 4c_t U^{(t)} S^T S - \alpha^{(t)},
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial W^{(t)}} = &-2b_t X^{(t)} H^{(t)} + 2b_t W^{(t)} \left(H^{(t)}\right)^T H^{(t)} \\
&+ 4d_t W^{(t)} \left(W^{(t)}\right)^T W^{(t)} - 4d_t W^{(t)} S^T S \\
&- \beta^{(t)},
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial H^{(t)}} = &2b_t H^{(t)} \left(W^{(t)}\right)^T W^{(t)} - 2b_t \left(X^{(t)}\right)^T W^{(t)} \\
&- \omega^{(t)},
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial S} = &\sum_{t=1}^{p} 4c_t S S^T S - \sum_{t=1}^{p} 4c_t S \left(U^{(t)}\right)^T U^{(t)} \\
&+ \sum_{t=1}^{q} 4d_t S S^T S - \sum_{t=1}^{q} 4d_t S \left(W^{(t)}\right)^T W^{(t)} - \gamma.
\end{aligned} \tag{12}
$$

Using the KKT conditions that $\alpha_{ij}^{(t)} U_{ij}^{(t)} = 0$, $\beta_{ij}^{(t)} W_{ij}^{(t)} = 0$, $\omega_{ij}^{(t)} H_{ij}^{(t)} = 0$, and $\gamma S = 0$, we have $(\partial L/\partial U^{(t)})_{ij} U_{ij}^{(t)} = (\partial L/\partial U^{(t)})_{ij} (U_{ij}^{(t)})^2 = 0$, $(\partial L/\partial W^{(t)})_{ij} W_{ij}^{(t)} = (\partial L/\partial W^{(t)})_{ij} (W_{ij}^{(t)})^2 = 0$, $(\partial L/\partial H^{(t)})_{ij} H_{ij}^{(t)} = (\partial L/\partial H^{(t)})_{ij} (H_{ij}^{(t)})^2 = 0$, and $(\partial L/\partial S)_{ij} S_{ij} = (\partial L/\partial S)_{ij} (S_{ij})^2 = 0$. Solving the above equations, we derive the following update rules.

Input: nonnegative matrices $\{A^{(1)}, \ldots, A^{(p)}\}$ and $\{X^{(1)}, \ldots, X^{(q)}\}$, number of communities k, parameters $\{a_1, \ldots, a_p\}$, $\{b_1, \ldots, b_q\}$, $\{c_1, \ldots, c_p\}$, $\{d_1, \ldots, d_q\}$.
Output: common indicator matrix S
(1) Initialize $U_{ij}^{(t)} \geq 0$, $W_{ij}^{(t)} \geq 0$, $H_{ij}^{(t)} \geq 0$, $S_{ij} \geq 0$, $\forall i, j, t$
(2) **while** not convergent **do**
(3) **for** $t = 1$ **to** p **do**
(4) update $U^{(t)}$ according to Formula (13)
(5) **end for**
(6) **for** $t = 1$ **to** q **do**
(7) update $W^{(t)}$ and $H^{(t)}$ according to Formula (14) and (15)
(8) **end for**
(9) update S according to Formula (16)
(10) **end while**

Algorithm 1: RJNMF community detection.

Updating $U^{(t)}$. Update $U^{(t)}$ according to formula (13), while other variables are fixed:

$$U_{ij}^{(t)} \longleftarrow U_{ij}^{(t)} \sqrt{\frac{\left(a_t A^{(t)} U^{(t)} + c_t U^{(t)} S^T S\right)_{ij}}{\left(a_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} + c_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)}\right)_{ij}}}. \tag{13}$$

Updating $W^{(t)}$. Update $W^{(t)}$ according to formula (14), while other variables are fixed:

$$W_{ij}^{(t)} \longleftarrow W_{ij}^{(t)} \sqrt{\frac{\left(b_t X^{(t)} H^{(t)} + 2d_t W^{(t)} S^T S\right)_{ij}}{\left(b_t W^{(t)} \left(H^{(t)}\right)^T H^{(t)} + 2d_t W^{(t)} \left(W^{(t)}\right)^T W^{(t)}\right)_{ij}}}. \tag{14}$$

Updating $H^{(t)}$. Update $H^{(t)}$ according to formula (15), while other variables are fixed:

$$H_{ij}^{(t)} \longleftarrow H_{ij}^{(t)} \sqrt{\frac{\left(\left(X^{(t)}\right)^T W^{(t)}\right)_{ij}}{\left(H^{(t)} \left(W^{(t)}\right)^T W^{(t)}\right)_{ij}}}. \tag{15}$$

Updating S. Update S according to formula (16), while other variables are fixed:

$$S_{ij} \longleftarrow S_{ij} \sqrt{\frac{\left(\sum_{t=1}^{p} c_t S \left(U^{(t)}\right)^T U^{(t)} + \sum_{t=1}^{q} d_t S \left(W^{(t)}\right)^T W^{(t)}\right)_{ij}}{\left(\sum_{t=1}^{p} c_t S S^T S + \sum_{t=1}^{q} d_t S S^T S\right)_{ij}}}. \tag{16}$$

With the above updating rules, the optimization algorithm is presented in Algorithm 1. Note that the iterative process is stopped if these cluster matrices converge or the number of iterations reaches a given threshold. Besides, the community detection results are obtained from S by finding the max value in each row.

3.4. Theoretical Analysis

3.4.1. Correctness Analysis. In this subsection, we prove the correctness of the updating rules according to KKT condition. Taking formula (13) as an example, we show that the solution is a KKT fixed point.

From the derivatives of L with respect to $U^{(t)}$ in formula (9), and according to the KKT condition $\alpha_{ij}^{(t)} U_{ij}^{(t)} = 0$, then the fixed point must satisfy the following function at convergence:

$$\left(-a_t A^{(t)} U^{(t)} + a_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} + c_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} - c_t U^{(t)} S^T S\right)_{ij} U_{ij}^{(t)} = 0. \tag{17}$$

At convergence, $(U^{(t)})^{(\infty)} = (U^{(t)})^{(m+1)} = (U^{(t)})^{(m)} = U^{(t)}$, and according to solution of formula (13), we have

$$\left(-a_t A^{(t)} U^{(t)} + a_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} + c_t U^{(t)} \left(U^{(t)}\right)^T U^{(t)} - c_t U^{(t)} S^T S\right)_{ij} \left(U_{ij}^{(t)}\right)^2 = 0. \tag{18}$$

It is obvious that formula (18) is identical to formula (17); thus, the solution of formula (13) satisfies formula (17); that is, the solution satisfies the KKT condition. Similarly, the correctness of the other updating rules can be proved.

3.4.2. Convergence Analysis. We use the auxiliary function approach [30] to prove the convergence of formula (13). We first introduce the definition of auxiliary function according to paper [30] as follows.

Definition 2 (see [30]). $Z(h, \widetilde{h})$ is an auxiliary function for $J(h)$, if conditions $Z(h, \widetilde{h}) \geq J(h)$ and $Z(h, h) = J(h)$ are satisfied.

Lemma 3 (see [30]). *If Z is an auxiliary function for J, then J is nonincreasing under the update rules $h^{(t+1)} = \arg\min_h Z(h, h^{(t)})$.*

We give the auxiliary function for the objective function formula (6) with regard to $U^{(t)}$ according to the definition above. Let $J(U^{(t)})$ denote the sum of all terms in formula (6) which contain $U^{(t)}$; then the following function

$$\begin{aligned} Z\left(U^{(t)}, \widetilde{U}^{(t)}\right) &= -a_t \sum_{pqr} A^{(t)}_{rp} \widetilde{U}^{(t)}_{pq} \widetilde{U}^{(t)}_{rq} \left(1 + \log \frac{U^{(t)}_{pq} U^{(t)}_{rq}}{\widetilde{U}^{(t)}_{pq} \widetilde{U}^{(t)}_{rq}}\right) \\ &\quad + (a_t + c_t) \sum_{pq} \left(\widetilde{U}^{(t)} \left(\widetilde{U}^{(t)}\right)^T \widetilde{U}^{(t)}\right)_{pq} \frac{\left(U^{(t)}_{pq}\right)^2}{\widetilde{U}^{(t)}_{pq}} \\ &\quad - c_t \sum_{pqr} \left(S^T S\right)_{rp} \widetilde{U}^{(t)}_{pq} \widetilde{U}^{(t)}_{rq} \left(1 + \log \frac{U^{(t)}_{pq} U^{(t)}_{rq}}{\widetilde{U}^{(t)}_{pq} \widetilde{U}^{(t)}_{rq}}\right) \end{aligned} \tag{19}$$

is an auxiliary function for $J(U^{(t)})$, where $A^{(t)}_{rp}$ is the rth row and pth column element of the adjacent matrix $A^{(t)}$ and $U^{(t)}_{pq}$ is the pth row and qth column element of the indicator matrix $U^{(t)}$. It is also a convex function in $U^{(t)}$ and its global minimum is

$$U^{(t)}_{ij} \longleftarrow \widetilde{U}^{(t)}_{ij} \sqrt{\frac{\left(a_t A^{(t)} \widetilde{U}^{(t)} + c_t \widetilde{U}^{(t)} S^T S\right)_{ij}}{(a_t + c_t) \left(\widetilde{U}^{(t)} \left(\widetilde{U}^{(t)}\right)^T \widetilde{U}^{(t)}\right)_{ij}}}. \tag{20}$$

Then we show the convergence of updating $U^{(t)}$ by formula (13); that is, we prove the following statement: when other variables are fixed, updating $U^{(t)}$ according to formula (13) monotonically decreases formula (6) until convergence.

Proof. According to Definition 2, Lemma 3, and the auxiliary function we develop, at any iteration $k \geq 0$ during updating $U^{(t)}$, we have

$$\begin{aligned} J\left(\left(U^{(t)}\right)^{(k)}\right) &= Z\left(\left(U^{(t)}\right)^{(k)}, \left(U^{(t)}\right)^{(k)}\right) \\ &\geq Z\left(\left(U^{(t)}\right)^{(k+1)}, \left(U^{(t)}\right)^{(k)}\right) \\ &\geq J\left(\left(U^{(t)}\right)^{(k+1)}\right), \end{aligned} \tag{21}$$

where $(U^{(t)})^{(k)}$ donates the updated $U^{(t)}$ at kth iteration. Thus, $J(U^{(t)})$ monotonically decreases. Since the objective function formula (6) is bounded below by 0, the updating of $U^{(t)}$ will converge. Then, the convergence of updating rule of formula (13) is proved. □

Similarly, the updating rules of $W^{(t)}, H^{(t)}, S$ can be proved convergent. Therefore, alternately updating $U^{(t)}, W^{(t)}, H^{(t)}, S$ by formulas (13), (14), (15), and (16) monotonically decreases formula (6) until convergence and the stationary point is a KKT fixed point, which guarantees the correctness and convergence of Algorithm 1.

Table 2: Details of the datasets.

Datasets	Users	User lists	Tweets	Communities
Politics-uk	419	3614	539592	5
Politics-ie	348	1047	267488	7
Football	248	7814	351300	20
Olympics	464	4942	725662	28

3.4.3. Complexity Analysis. We now analyze RJNMF's time complexity, using standard NMF as the basis for big O notation. RJNMF is essentially an extension of NMF for multiple data matrices. In terms of the standard NMF, $\|X - WH^T\|_F^2$, where $X \in R^{n \times m}$, $W \in R^{n \times k}$, and $H \in R^{m \times k}$, it is known that the cost for update rules in each iteration is $O(nmk)$. As RJNMF's update rule for each $H^{(t)}$ is the same as the original NMF; its cost is also $O(nmk)$. For each $U^{(t)}$ in formula (13), the additional cost in terms of standard NMF is the second term of the numerator and denominator, whose time complexity is $O(nk^2)$. As k is the number of communities, which is a small constant s.t. $k \ll n$, then the cost of updating each $U^{(t)}$ is $O(n^2k + nk^2) \approx O(n^2k)$. Similarly, the time complexity of updating $W^{(t)}$ is $O(mnk + nk^2) \approx O(mnk)$, and that of updating S is $O(pnk^2 + qnk^2) \approx O(nk^2)$, where p and q are numbers of link views and content views, respectively. Therefore, the time complexity of RJNMF updating rules in each iteration is $O(pn^2k + 2qmnk + nk^2) \approx O(\max(n^2k, mnk))$, making RJNMF a linear extension of NMF.

4. Experimental Results

In this section, we evaluate the RJNMF algorithm on real-world datasets and compare it with some existing community detection algorithms.

4.1. Datasets. Four real-world datasets containing both link data and content data are used in the experiment. The datasets are collected from Twitter (http://mlg.ucd.ie/networks/); the ground truth results of community detection are available, so that we can evaluate the performance of our method according to the ground truth. The details of the datasets are summarized in Table 2. The *politics-uk* dataset is a collection of 419 Members of Parliament in the UK, and it consists of 5 communities, corresponding to the political parties. The *politics-ie* dataset describes the Irish politicians and political organizations which are clustered into 7 communities. The *football* dataset contains 248 English Premier League football players and club active on Twitter; the ground truth corresponds to 20 clubs. The *olympics* dataset contains a collection of 464 users (https://twitter.com/Telegraph2012/london2012) which consists of the athletes and organizations involved in the London 2012 Summer Olympics; they are assigned to 28 communities according to different sports.

A collection of different link and content data is available for each dataset. We choose to use *follows*, *mentions*, *retweets,* and *user list*, *tweets* in the experiment. Particularly, the *follows* describes the follow relationship, the *mentions* contains

Table 3: Demographics of the two datasets.

Datasets	Items	Users	Comments	Descriptions	Communities
Last.fm	9694	131,153	31,172	14076	21
Yelp	2624	17,068	18,067	1779	7

links between users who mentioned each other, the *retweets* describes the retweet interaction, all the three relations are regarded as link information, and three adjacency matrices are constructed from them, respectively. The *user lists* is constructed based on Twitter lists to which each user has most recently been assigned, and *tweets* is constructed from the concatenation of the 500 most recently posted tweets for each user; then two user-content matrices are obtained from the two types of content information.

To further evaluate the performance of our method, we carry out the experiments on *Last.fm* and *Yelp* datasets which contain thousands of items [9]. The *Last.fm* dataset consists of 9694 artists which are clustered into 21 music genres; for each artist, his or her biodescription and user comments are crawled; then 131153 *users*, 31172 *comments*, and 14076 *descriptions* are obtained. *Users*, *comments*, and *descriptions* are all content information used to cluster the artists and three corresponding user-content matrices are gained. The *Yelp* dataset consists of 2,624 items from 7 categories. There are also three types of content information, that is, *users*, *comments*, and *businesses' names* (*descriptions*), from which we obtain three user-content matrices. The summary demographics the datasets are showed in Table 3.

4.2. Baseline Methods. To demonstrate the performance of our method, we also carry out experiments of the following baseline algorithms on the datasets described above and compare the performance of RJNMF with these algorithms'.

(i) Pairwise Coregularized Spectral clustering (PCoSpec) and Center-wise Coregularized Spectral clustering (CCoSpec) [16]: two coregularization schemes are adopted in spectral clustering framework; PCoSpec utilizes a pairwise coregularization to enforce the eigenvectors of each pair to be similar and CCoSpec employs the centroid based coregularization to enforce the eigenvectors to be similar with a common center. In the experiments, all the information contained in each dataset is utilized. The inputting affinity matrices of link information are all the adjacency matrices of each dataset, and affinity matrices of content information are gained by the default Gaussian kernel from all the user-content matrices according to the authors' suggestions. The regularization parameters are set to 0.01 as suggested by the authors.

(ii) Coregularized Graph Clustering (CGC): CGC [12] is based on symmetric NMF with coregularization to deal with multiple link data. The private clustering result of each adjacency matrix is obtained by solving the joint matrix factorization problem. In the experiments, all the information contained in each dataset is utilized. The inputting affinity matrices of link information are all the adjacency matrices of each dataset; the affinity matrices of content information are computed using the RBF kernel from all the user-content matrices with the authors' guidance. We set the regularization parameters to 1, as suggested by the authors.

(iii) Pairwise Coregularized NMF (PCoNMF) clustering and Cluster-wise Coregularized NMF (CCoNMF)clustering: CoNMF is proposed in [9] to combine the link and content information for joint factorization and find out the separate clustering solution to each view; particularly, two kinds of coregularization penalties, pairwise and cluster-wise constraints, are developed to ensure the similarity of each pair of community indicators. In the experiment, all the information contained in each dataset is utilized; the inputting matrices are all the adjacency matrices and user-content matrices. In the experiments on the first four datasets, the parameters for follows, mentions, retweets, user list, and tweets are set to 2, 1, 1, 2, and 1, respectively; because follows and user list are more important according to prior knowledge, the regularization parameters are set to 1 as suggested. In the experiments on *Last.fm* and *Yelp* datasets, the parameter for each view is set to 1, and regularization parameters are set to 2, as suggested in [9].

4.3. Results. In this paper, accuracy [31] and normalized mutual information (NMI) [32] are adopted to evaluate the community detection performances of different methods. Formally, let $G = \{G_1, \ldots, G_k\}$ be the set of communities in the ground truth and $C = \{C_1, \ldots, C_t\}$ be the t communities extracted by different approaches. Both accuracy and NMI adopt the ground truth as a baseline; the values range from 0 to 1 and the higher value means better performance.

Accuracy is used to measure how the extracted community structure approaches the ground truth community structure; it calculates the ratio of the nodes clustered into the correct communities relative to all the nodes contained in the network. To compute the accuracy, each ground truth community is assigned a label, which is also assigned to each node in the community as the true label, denoted as *v.label*. Then, we scan the nodes in $C_i \in C$ to count the occurrences of each true label in each derived community and take the label occurring most frequently in $C_i \in C$ as the label of community. After this process, some communities may have the same labels. For these communities, we keep the community with the largest number of nodes with the same label, and, for each of the other communities, if the nodes

Table 4: Accuracy for different methods on six datasets.

Dataset	PCoSpec	CCoSpec	CGC	PCoNMF	CCoNMF	RJNMF
Politics-uk	0.6682	0.5403	0.6109	0.5775	0.5131	**0.8592**
Politics-ie	0.6034	0.6264	0.7729	0.5574	0.4511	**0.8647**
Football	0.6169	0.5362	0.6733	0.4113	0.2056	**0.7656**
Olympics	0.5991	0.5906	0.6961	0.4267	0.1373	**0.8022**
Last.fm	0.5062	0.5173	0.4376	**0.5197**	0.4971	0.5127
Yelp	0.5732	0.6086	0.3162	0.6763	0.6735	**0.6802**

Table 5: NMI for different methods on six datasets.

Dataset	PCoSpec	CCoSpec	CGC	PCoNMF	CCoNMF	RJNMF
Politics-uk	0.6240	0.5025	0.5531	0.4708	0.3628	**0.7022**
Politics-ie	0.6634	0.6108	0.7732	0.5726	0.3367	**0.8596**
Football	0.7551	0.6960	0.7696	0.5286	0.2948	**0.8403**
Olympics	0.7874	0.7684	0.8397	0.5581	0.2613	**0.8920**
Last.fm	0.6204	0.6433	0.5543	**0.6529**	0.6047	0.6485
Yelp	0.6999	0.7186	0.4267	0.7853	0.7745	**0.7982**

in the community have no other labels, that community is removed from C, all nodes in that community are taken as misclassified nodes; otherwise, we take the next label whose node number is the next-largest in the community as the label of that community. Then, community $C_i \in C$ and community $G_i \in G$ with the same label match with each other, and we assign the label of community $C_i \in C$ to each node $v \in C_i$ as its detected label, denoted as $v.\text{label}^*$. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\sum_{v \in \bigcup_{i=1}^{|C|} C_i} \delta\left(v.\text{label}, v.\text{label}^*\right)}{n}, \tag{22}$$

where v.label donates the community label of node v in the ground truth, $v.\text{label}^*$ is the detected community label of node v, and $\delta(v.\text{label}, v.\text{label}^*)$ is the Kronecker delta function which is 1 if the community labels of node v are same, 0 otherwise. If $C = G$, accuracy equals 1. If C and G are completely different, accuracy equals 0.

NMI is used to measure the partitioning quality based on the ground truth; NMI estimates the similarity between true partitions and the detected. The NMI of our partition C and the ground truth G is defined as follows. Let F be the confusion matrix whose element F_{ij} is the number of nodes of community i of the partition C that are also in the community j of the partition G. NMI is calculated as follows:

$$\text{NMI}(C, G) = \frac{-2\sum_{i=1}^{f_C}\sum_{j=1}^{f_G} F_{ij}\log\left(F_{ij}n/F_{i\cdot}F_{\cdot j}\right)}{\sum_{i=1}^{f_C} F_{i\cdot}\log\left(F_{i\cdot}/n\right) + \sum_{j=1}^{f_G} F_{\cdot j}\log\left(F_{\cdot j}/n\right)}, \tag{23}$$

where f_C (f_G) is the number of groups in the partition C (G), $F_{i\cdot}$ ($F_{\cdot j}$) is the sum of the elements of F in row i (column j), and n is the number of nodes. If $C = G$, $\text{NMI}(C, G) = 1$. If C and G are completely different, $\text{NMI}(C, G) = 0$.

The scores of accuracy and NMI for different methods on the datasets are presented in Tables 4 and 5, respectively. From the results shown in the tables, some conclusions can be drawn. Note that all experiment results are averaged over 50 different runs. The parameters are tuned for optimal performance of all methods.

It can be observed that RJNMF performs better than the other community detection algorithms on most of the datasets. For example, RJNMF algorithm can get about 19% and 8% improvement in accuracy and NMI, respectively, compared with PCoSpec which is the second best method on *politics-uk* dataset and can get about 30% improvement in both accuracy and NMI compared with CCoNMF algorithm. In terms of *Last.fm* and *Yelp* datasets, which contain thousands of nodes, RJNMF also performs well. Although PCoNMF performs better than RJNMF on *Last.fm* dataset, the gap is relatively small. The better performance of RJNMF may due to the following reasons: (1) the similarity regularization contributes to reducing the effect of noise and catching the core community structure. (2) RJNMF drives the common community structure directly by obtaining common community indicator from solving the optimization problem and utilizes the link and content information more effectively than the other methods. Overall, the results demonstrate the effectiveness of RJNMF for community detection in heterogeneous networks.

4.4. RJNMF Parameter Study. There are four sets of parameters in RJNMF to balance the effects of different parts in the optimization: a_t for the link data, b_t for the content data, and c_t and d_t for regularization. The values of a_t and b_t determine the importance of different link and content data in the optimization, respectively, while the values of c_t and d_t determine the weights of the similarity regularization. We carry out experiments on *politics-uk*, *politics-ie*, *football*, *and olympics* datasets to evaluate the performances of our method with varying parameters.

We first discuss the relative values of parameters for different link data and content data. In terms of the first

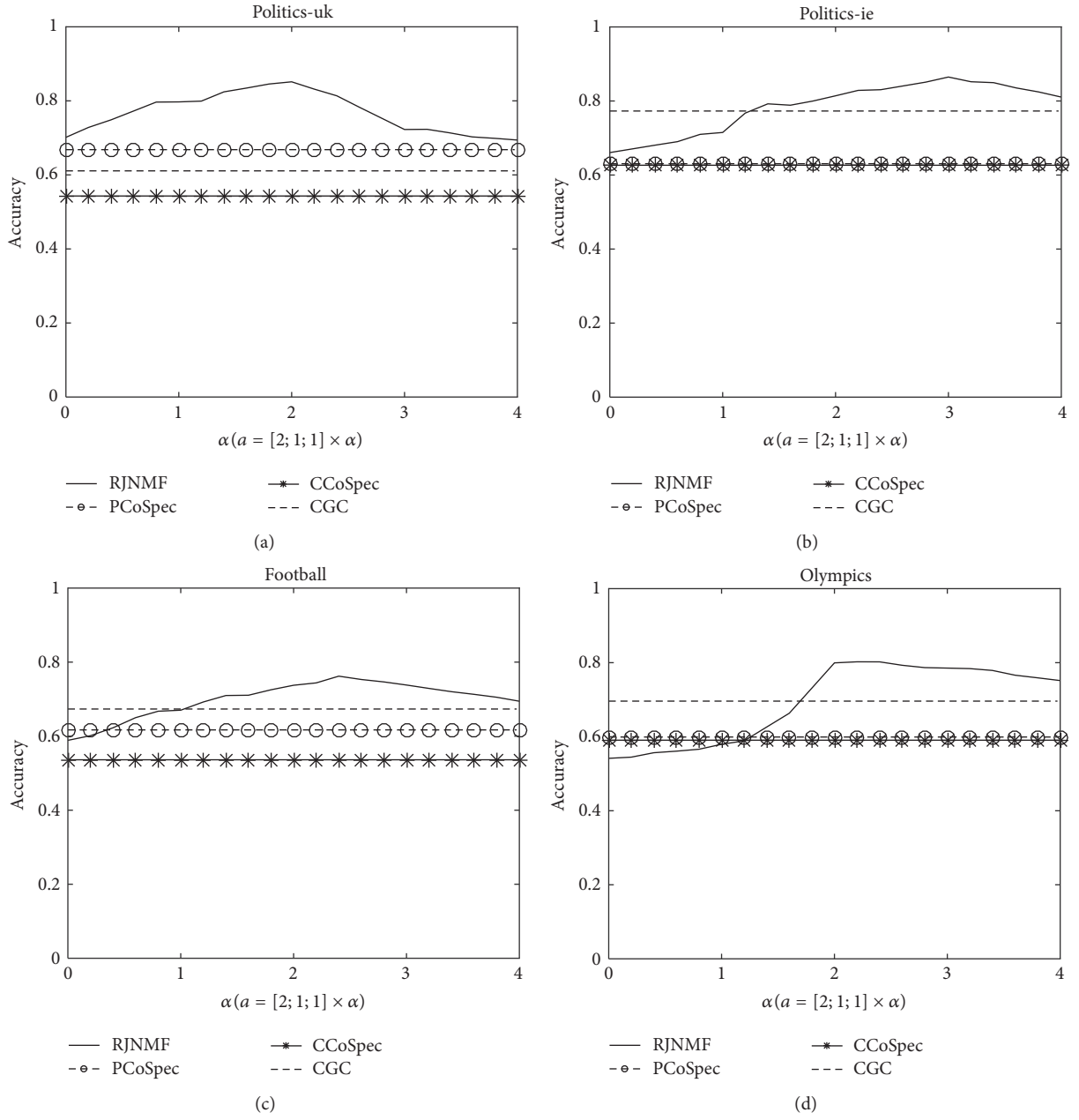

Figure 2: Evaluating accuracy of RJNMF on varying a_t ($t = 1, 2, 3$) when b, c, and d are fixed to [2; 1], [4; 2; 2], and [2; 1], respectively, on four datasets. α is the coefficient determining the rate of change and a_t ($t = 1, 2, 3$) are gained by 2α, α, and α, respectively. The lines of the baseline methods represent the best performance.

four datasets, from the detection results of single view and the ground truth, we know that the *follow* data indicates community structure more clearly than the other two types of link data [15]; then we set the weight of *follow* higher than those of *mention* and *retweet*. Similarly, the weight of *user lists* is set higher than that of *tweets*. However, in most real-world cases, which data is more important and reliable is usually unknown; then parameters a_t ($1 \le t \le p$) and c_t ($1 \le t \le p$) are set to the same value by default, so are b_t ($1 \le t \le q$) and d_t ($1 \le t \le q$). If some prior information is available, one can also choose different values based on the importance of individual view; the view indicating the underlying structure more clearly can be given a higher weight to emphasize its effect.

Then we focus on the performance of RJNMF with varying weights of factorizations relative to regularization. Fixing the other parameters, we change a_t ($1 \le t \le p$) through multiplying by the varying coefficient α, note that the parameters a_t for different domains change in the same pace. Similarly, b_t, c_t, and d_t are varied in the same way. Figures 2–5

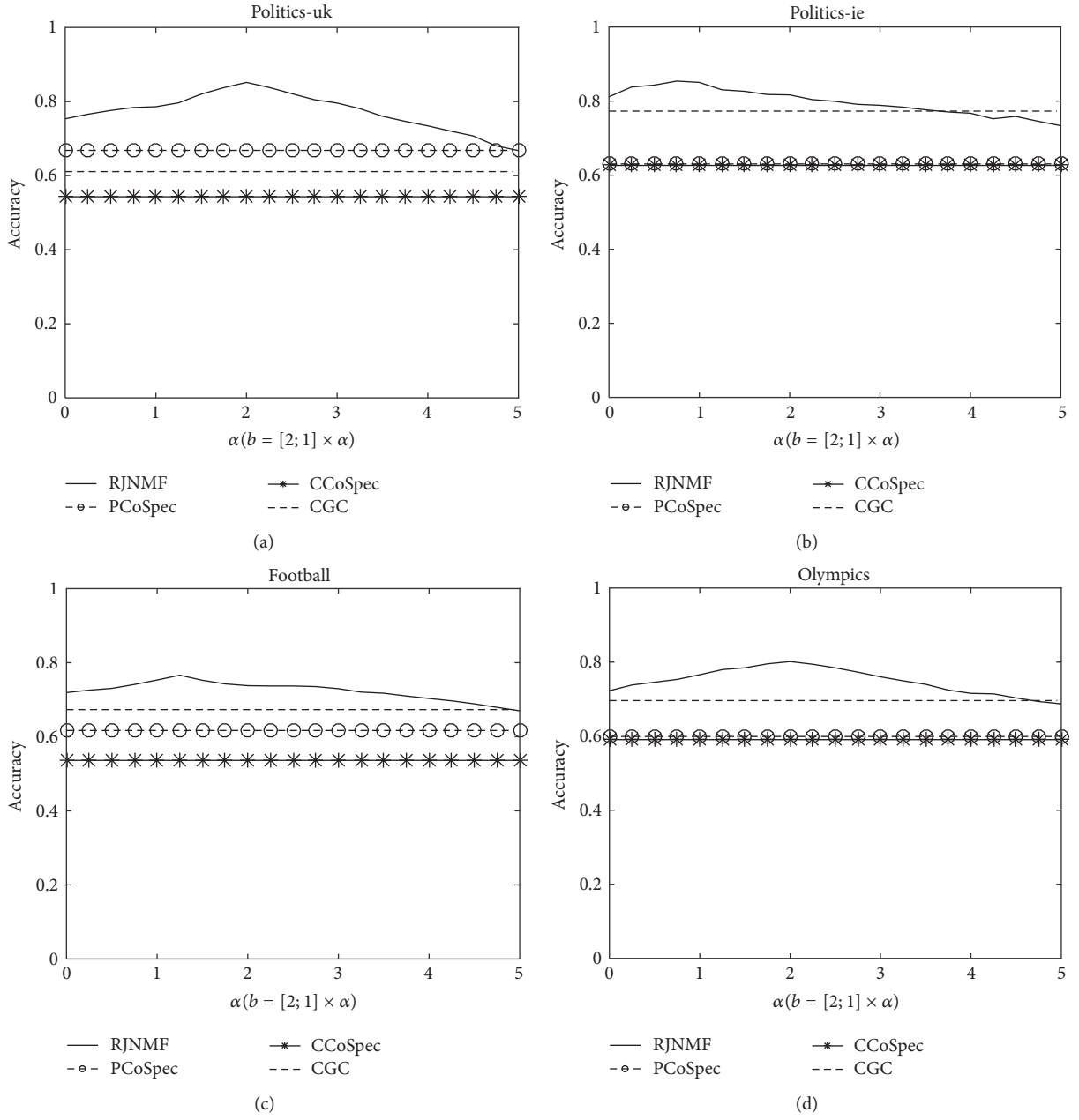

Figure 3: Evaluating accuracy of RJNMF on varying b_t $(t = 1, 2)$ when a, c, and d are fixed to [4; 2; 2], [4; 2; 2], and [2; 1], respectively, on four datasets. α is the coefficient determining the rate of change and b_t $(t = 1, 2)$ are gained by 2α and α, respectively. The lines of the baseline methods represent the best performance.

show the performances of RJNMF on the first four datasets with varying parameters. The lines of PCoSpec, CCoSpec, and CGC represent the best results of these methods; they are used as baselines. As can be seen, for all the four datasets, RJNMF has a relatively stable performance with all the varying parameters. RJNMF remains performing better than the best results of PCoSpec, CCoSpec, and CGC on a, when a is larger than [4; 2; 2]. In terms of c, RJNMF has a better performance across a range of [4; 2; 2] to [6; 3; 3] and performs poorly out of the range. It can be also observed that RJNMF is relatively stable on b and d across a wide range and performs better than the best results of the other algorithms. The results also suggest that the ratio of the parameters for regularization and factorizations can be set to $[0.3, 1]$ to achieve a good performance.

5. Conclusion

In this paper, we investigate community detection in heterogeneous networks systematically and propose a regularized joint NMF community detection framework. The key idea of the framework is to formulate a joint matrix factorization

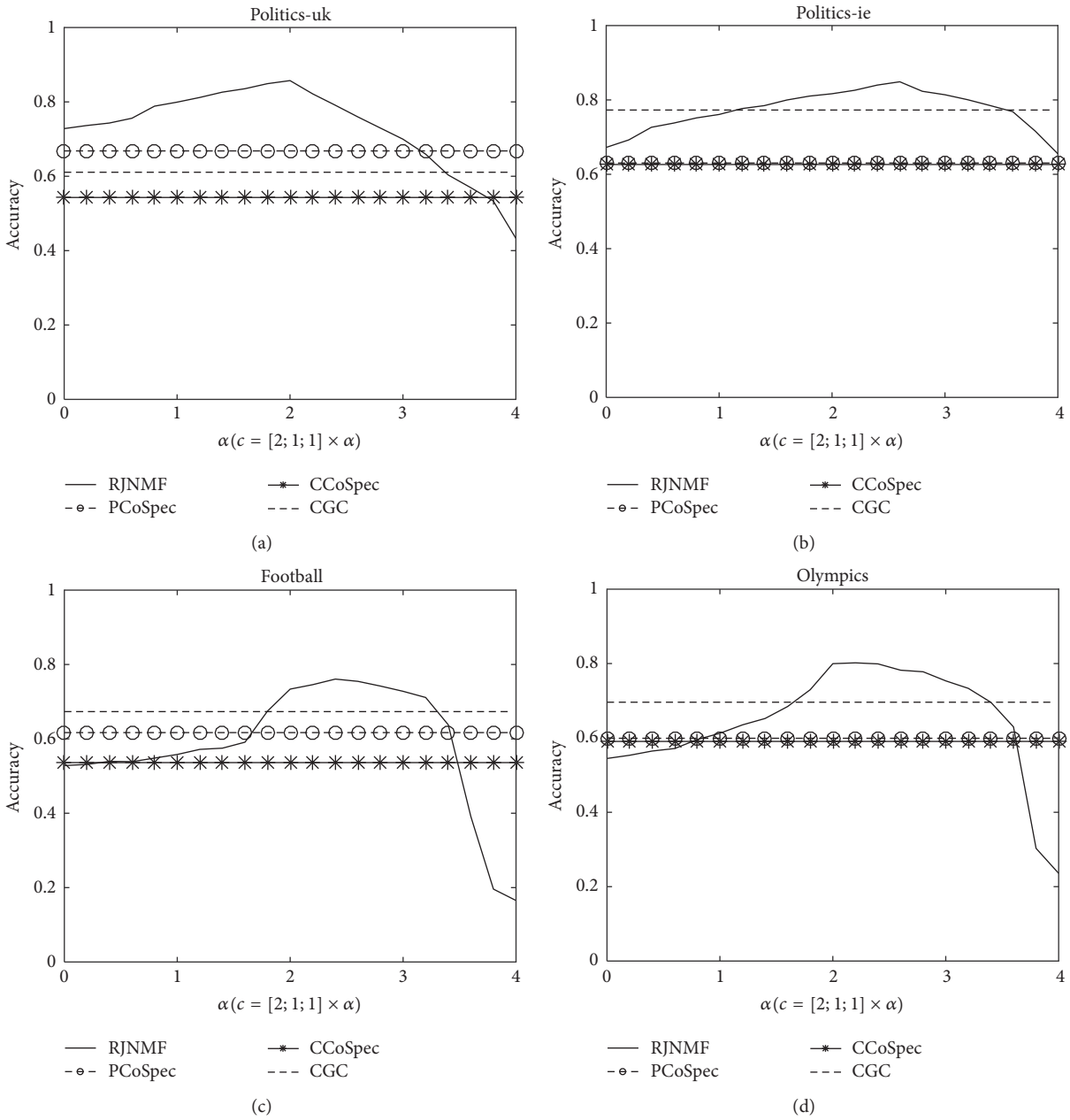

Figure 4: Evaluating accuracy of RJNMF on varying c_t $(t = 1, 2, 3)$ when a, b, and d are fixed to [4; 2; 2], [2; 1], and [2; 1], respectively, on four datasets. α is the coefficient determining the rate of change and c_t $(t = 1, 2, 3)$ are gained by 2α, α, and α, respectively. The lines of the baseline methods represent the best performance.

process with regularization that derive the common community structure directly and more accurately. In the framework, both the link data and content data are considered in the integration strategy and analyzed simultaneously; furthermore, in order to make better use of the two types of information, we develop a consensus similarity regularization to push the individual community detection result of each data source towards a common solution. Experiments on six real-world datasets demonstrate that RJNMF can effectively utilize the compatible and complementary information of link and content data for the joint community detection task.

As, in this paper, we focus on detecting communities with the presupposition that all the data share common underlying structure. However, in some real applications, the assumption may not hold where the community structures of multiple data sources may be different; therefore, in the future, we plan to investigate community detection in such cases.

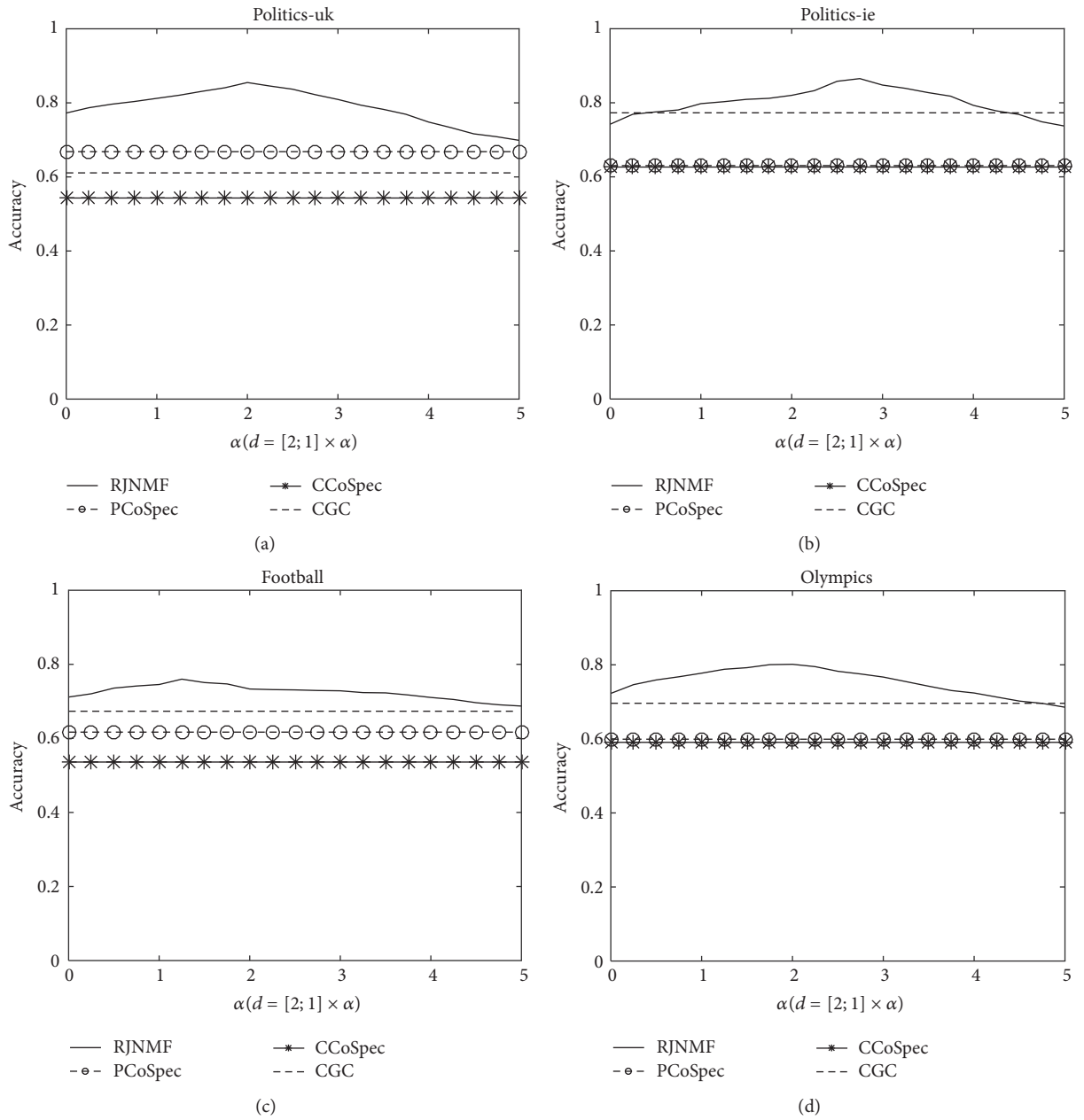

Figure 5: Evaluating accuracy of RJNMF on varying d_t ($t = 1, 2$) when a, b, and c are fixed to [4; 2; 2], [2; 1], and [4; 2; 2], respectively, on four datasets. α is the coefficient determining the rate of change and d_t ($t = 1, 2$) are gained by 2α and α, respectively. The lines of the baseline methods represent the best performance.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61473149) and Natural Science Foundation of Jiangsu Province, China (no. BK20140075).

References

[1] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI '15)*, pp. 2083–2089, July 2015.

[2] D. Kuang, H. Park, and C. H. Ding, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 12th SIAM International Conference on Data Mining (SDM '12)*, pp. 106–117, Anaheim, Calif, USA, April 2012.

[3] S. Van Dongen, "A cluster algorithm for graphs," In Centrum *voor Wiskunde en Informatica* (CWI), 2000.

[4] J. 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.

[5] L. Tang and H. Liu, *Community Detection and Mining in Social Media*, Morgan & Claypool Publishers, 2010.

[6] L. Tang, X. Wang, and H. Liu, "Uncovering groups via heterogeneous interaction analysis," in *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM '09)*, pp. 503–512, Miami, Fla, USA, December 2009.

[7] N. Wang, P. Chen, and X. Li, "Community detection in heterogeneous multi-mode social network via Co-training," in *Foundations of Intelligent Systems: Proceedings of the Eighth International Conference on Intelligent Systems and Knowledge Engineering, Shenzhen, China, Nov 2013 (ISKE 2013)*, vol. 277 of *Advances in Intelligent Systems and Computing*, pp. 531–538, Springer, Berlin, Germany, 2014.

[8] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SigKDD '09)*, pp. 359–368, Paris, France, July 2009.

[9] X. He, M. Y. Kan, P. Xie et al., "Comment-based multi-view clustering of web 2.0 items," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 771–782, International World Wide Web Conferences Steering Committee, 2014.

[10] D. Hidru and A. Goldenberg, "EquiNMF: graph regularized multiview nonnegative matrix factorization," https://arxiv.org/abs/1409.4018.

[11] J. Tang, X. Wang, and H. Liu, "Integrating social media data for community detection," in *Proceedings of the International Conference on Modeling and Mining Ubiquitous Social Media (MSM '11)*, pp. 1–20, Springer, 2011.

[12] W. Cheng, X. Zhang, Z. Guo, Y. Wu, P. F. Sullivan, and W. Wang, "Flexible and robust co-regularized multi-domain graph clustering," in *the 19th ACM SIGKDD international conference*, p. 320, Chicago, Illinois, USA, August 2013.

[13] J. Ni, H. Tong, W. Fan, and X. Zhang, "Flexible and robust multi-network clustering," in *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '15)*, pp. 835–844, Sydney, Australia, August 2015.

[14] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the 22nd International Conference on World Wide Web, International World Wide Web Conferences Steering Committee (WWW '13)*, pp. 1089–1098, Rio de Janeiro, Brazil, May 2013.

[15] D. Greene and P. Cunningham, *Producing a Unified Graph Representation from Multiple Social Fs*, Association for Computing Machinery, 2013.

[16] A. Kumar, P. Rai, and H. Daumé, *Co-Regularized Multi-View Spectral Clustering*, Advances in Neural Information Processing Systems, 2011.

[17] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discovery," *Bioinformatics*, vol. 21, supplement 1, pp. i213–i221, 2005.

[18] M. Lukk, M. Kapushesky, J. Nikkilä et al., "A global map of human gene expression," *Nature Biotechnology*, vol. 28, no. 4, pp. 322–324, 2010.

[19] C. Deng, Z. Lv, W. Liu et al., "Multi-view matrix decomposition: a new scheme for exploring discriminative information," in *Proceedings of the 24th International Conference on Artificial Intelligence*, AAAI Press, 2015.

[20] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2439–2446, IEEE, Barcelona, Spain, November 2011.

[21] X. Guo, D. Liu, B. Jou, M. Zhu, A. Cai, and S.-F. Chang, "Robust object co-detection," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3206–3213, Portland, Ore, USA, June 2013.

[22] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI '14)*, pp. 2149–2155, July 2014.

[23] H. T. Nguyen, T. N. Dinh, and T. Vu, "Community detection in multiplex social networks," in *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS '15)*, pp. 654–659, IEEE, Hong Kong, May 2015.

[24] A. Mahmood and M. Small, "Subspace based network community detection using sparse linear coding," *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 3, pp. 801–812, 2016.

[25] S. Guesmi, C. Trabelsi, and C. Latiri, "Community detection in multi-relational bibliographic networks," in *Database and Expert Systems Applications*, vol. 9828 of *Lecture Notes in Computer Science*, pp. 11–18, Springer International Publishing, Cham, Switzerland, 2016.

[26] X. Zheng, S. Zhu, J. Gao, and H. Mamitsuka, "Instance-wise weighted nonnegative matrix factorization for aggregating partitions with locally reliable clusters," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI '15)*, pp. 4091–4097, Buenos Aires, Argentina, July 2015.

[27] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, "Spectral ensemble clustering," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pp. 715–724, Sydney, Australia, August 2015.

[28] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

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

[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000.

[31] J. Cheng, M. Leng, L. Li, H. Zhou, and X. Chen, "Active semi-supervised community detection based on must-link and cannot-link constraints," *PLoS ONE*, vol. 9, no. 10, Article ID e110088, 2014.

[32] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, 2003.