Received November 4, 2017, accepted December 1, 2017, date of publication December 14, 2017, date of current version May 2, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2783542

Overlapping Community Detection Using Non-Negative Matrix Factorization With Orthogonal and Sparseness Constraints

NAIYUE CHEN[1], YUN LIU[1], AND HAN-CHIEH CHAO[2,3,4,5], (Senior Member, IEEE)

[1]Key Laboratory of Communication and Information Systems, School of Electronic and Information Engineering, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
[2]School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China
[3]School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430023, China
[4]Department of Electrical Engineering, National Dong Hwa University, Hualien 97401, Taiwan
[5]Department of Computer Science and Information Engineering, National Ilan University, Yilan City 26041, Taiwan

Corresponding author: Yun Liu (liuyun@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61401015. Academic Discipline and Postgraduate Education Project of Beijing Municipal Commission of Education.

ABSTRACT Network is an abstract expression of subjects and the relationships among them in the real-world system. Research on community detection can help people understand complex systems and identify network functionality. In this paper, we present a novel approach to community detection that utilizes a nonnegative matrix factorization (NMF) model to divide overlapping community from networks. The study is based on the different physical meanings of the pair of matrices W and H to optimize the constraint condition. Many community detection algorithms based on NMF require the number of known communities as a prior condition, which limits the field of application of the algorithms. This paper handled the problem by feature matrix preprocessing and ranking optimization, so that the proposed algorithm can divide the network structure with unknown community number. Experiments demonstrated that the proposed algorithm can effectively divide the community structure, and identify network overlay communities and overlapping nodes.

INDEX TERMS Community detection, non-negative matrix factorization, orthogonal constraint, sparse constraint, ranking optimization.

I. INTRODUCTION

Community structure is an important property of network, which can reflect the functionality and features of network. In a real-world network structure, each community can reflect different functions, interests, behaviors, and so on. Communities can not only reveal the potential characteristic of complex networks, but also uncover the underlying correlations among their components [1]. An Accurate community structure is very important for improving the relevance of search engine results and the accuracy of recommender systems.

In a protein network, individual proteins in the same community have the same or similar functions [2]. It can detect community belonging to the unknown protein to identify the function of the unknown protein. In an academic cooperation network, the scholars in the same community have the same research field [3]. In a social network, the users in the same community have the same interests or focus [4]. On the World Wide Web, websites in the same community may have the similar themes [5]. Therefore, community detecting plays an important role in network analysis, users behavior recommendation, and the organizational behavior of the real-word system.

Cyber-Physical-Social Computing and Network has been adapted to very different domains such as online social networks. The community detection is an important part of social networks analysis. Hence, the research of community detection closely related to Cyber-Physical-Social Computing and Network.

Many community detection algorithms have been proposed in the literatures to identity complex community structures in complex network. A prevalent measure, namely modularity, has been extensively used for community detection, which is rooted from the seminal work of community structure analysis by Girvan and Newman [6]. Modularity is often used to measure the quality of community detection, and quantify how much difference between the density of the edges inside identified communities and the expected edge density of a network with the same number of nodes and edges incident to each node, but randomly connected [7]. Thus, the higher the value of modularity, the more precise communities detected in the network. In addition, numerous techniques have been proposed for community detection in bipartite network networks. Guimer *et al.* [8] proposed a new modularity concept based on the number of heterogeneous vertices that are joined together by homogeneous vertices. Barber [9] proposed the BRIM algorithm which allowed the two kinds of nodes to join in the same community. Based on this point, Mutara [10] proposed another modularity formula to reflect the relationship between the two kinds of communities. Suzuki and Wakita [11] proposed an algorithm that can quantify the degree of connectivity between any two heterogeneous communities. Lancichinettiet et al. proposed another algorithm which seeks a local maximum of the community "fitness" function (based on internal link density) by modifying nodes' community "appropriateness" scores through a series of inclusion–exclusion moves. Some researchers have introduced label propagation as bipartite network network community detection, improving the performance of the algorithm by optimizing leap attenuation, penalty factors and so on [12]–[14].

Communities in networks often overlap such that nodes simultaneously belong to several groups [15], [16]. In fact, that many real networks have communities with pervasive overlap, where allow node belongs to more than one group, has the consequence that a global hierarchy of nodes cannot capture the relationships between overlapping groups. The work on detecting overlapping communities was previously proposed by Palla *et al.* [15] with the clique percolation algorithm (CPM). CPM is based on the assumption that a community consists of fully connected subgraphs and detects overlapping communities by searching each subgraph for adjacent cliques that share at least a certain number of nodes with it. Kumpula *et al.* [17] have developed a fast implementation of the CPM, called the Sequential Clique Percolation algorithm (SCP). It consists in detecting k-clique communities by sequentially inserting the edges of the graph at study, one by one, starting from an initial empty graph. LFM [18] expands a community from a random seed node until the fitness function is locally maximal. LFM depends significantly on a parameter of the fitness function that controls the size of the communities. Gregory [19] proposed an overlapping community method called GONGA, and it extended Girvan and Newman's well-known algorithm based on the betweeness centrality measure. The algorithm performed hierarchical clustering — partitioning a network into any desired number of clusters — but allows them to overlap. COPRA [20] is an extension of the label propagation algorithm for overlapping community detection. Each node updates its belonging coefficients by averaging the coefficients over all its neighbors.

With the development of Internet technology, the scale of users and networks has increased. Most belong to sparse networks and satisfy the condition of a sparse network. We can consider the network as users' relationship matrix. If a connection exists between users, the factor in the matrix is 1; otherwise, it is 0. As the network has the sparse property, the matrix will be a sparse matrix. Therefore, the matrix shows high dimensional and sparse phenomena. The traditional algorithms have drawbacks for dealing with the phenomenon. In this paper, we introduce (non-negative matrix factorization) NMF to determine the based and membership matrices using dimension reduction decomposition.

In this work, we propose a novel approach to community detection based on NMF. The advantages of this methodology are: i) the orthogonal and sparseness optimization strategy, which can make an approximate factorization of the matrix V into a pair of matrices W and H; ii) ranking optimization allows the community to share members and optimize the community number; and iii) the method does not suffer from the drawbacks of prior conditions, such as the community number.

In the following section we present the theoretical foundations of our approach in related work. In section III, we propose the community detection algorithm using non-negative matrix factorization with orthogonal and sparseness constraint along with an illustrative example to provide intuition behind the algorithm. Following the section, we test our algorithm on a variety of artificial and real-world benchmark problems and present our experimental results.

II. RELATED WORK

This research is motivated by the NMF technique, a machine-learning algorithm based on decomposition by parts that can uncover localized features in feature space [21].

The algorithm is a matrix factorization method that all the elements in the matrix are under the non-negative constraints condition. NMF can focus on the relationship among different parts of the data, reduce the dimensionality and extract the features efficiently, and the decomposition form and the decomposition result are interpretable.

A. NON-NEGATIVE MATRIX FACTORIZATION MODEL

D.D.Lee and H.S.Seung presented the results of the study of nonnegative matrix factorization on "Nature" in 1999. Non-negative Matrix Factorization is a feature extraction and dimensionality reduction technique in machine learning, which has been adapted to community detection recently. The technique decomposes the feature matrix into two matrices with non-negativity constraints.

The network can be represented as a single matrix V of size $n \times m$. In community study the matrix is symmetric to show the nodes connection with each other as $n \times n$. In this matrix, column and row both correspond to the similarities from one node to all nodes because of the symmetry of V. The major analytical method application of NMF is an approximate factorization of the matrix V into a pair of matrices W and H.

$$V \simeq W \times H \tag{1}$$

The NMF is only an approximate factorization, not an exact one [21]. The unique feature of the NMF algorithm is every element in each is non-negative. The factorization is carried out with a particular rank k so that W is of dimension $n \times k$ and H is $k \times m$. Moreover, the factorization could be viewed as a representation of the data in a new space of lower dimensionality k. Most experiments show the matrix W can reflect the final clustering partition and matrix H means the membership correspondingly. In our research, we combined the clustering partition and ranking membership to determine the community structure finally.

The key of matrix factorization is iteratively updating matrices W and H to improve the approximation to V while maintaining non-negative matrix entries throughout [21]. Thus, the algorithm need a given value of the NMF dimensionality k which is an important factor in matrix factorization. It is also a technical challenge in most improved algorithms for community discovery because the number of communities is unknown in advance. NMF starts with random initial matrices W and H which are chosen from a normal distribution with mean 0, variance 1, and standard deviation 1. The two matrices are iteratively updated using the following rules:

$$W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{2}$$

$$H_{au} \leftarrow H_{au}\frac{(W^T V)_{au}}{(W^T WH)_{au}} \tag{3}$$

NMF algorithm minimize the root-mean-square error between the actual data V and the reduced dimension reconstruction of the data WH. For a given k, the algorithm run iteratively updating procedure and stop criterion until find a good approximate factorization.

B. NMF APPLICATION IN COMMUNITY DETECTION

Many community detection algorithms based on NMF have been proposed in the literatures to identity complex community structures in complex network. After ten years of development, non-negative matrix decomposition has made great progress, in addition to the standard NMF. In addition, it is mainly divided into constrained NMF, structured NMF and generalized NMF [22]. Liu *et al.* [23] proposed sparse NMF by Lagrangian multiplier method. Hoyer [24] proposed sparse NMF which can make the basis matrix and membership matrix sparseness. Li *et al.* [25] proposed an NMF algorithm with orthogonal constraint to make the basis matrix more orthogonal for reflecting the network feature. In addition to the standard NMF, Ding et al. Proposed a spectral clustering algorithm based on nonnegative matrix, proposed an extended form of NMF - symmetric NMF and weighted symmetric NMF. The symmetric matrix is decomposed to obtain the mesh data of spectral clustering. In addition, matrix operations can effectively solve multi-label classification problems [26].

Many researchers proposed community detection algorithms based on NMF, like Bayesian NMF, bounded NMF, symmetric NMF. Jin utilized the NMF to preserve the expected node degrees and enhance applicability to real-world networks. Zarei found a novel NMF-based algorithm to divide fuzzy communities. It is worth noting that the mentioned above algorithms based on NMF can detect overlapping communities structure.

III. NON-NEGATIVE MATRIX FACTORIZATION WITH ORTHOGONAL AND SPARSENESS CONSTRAINT

A. FEATURE MATRIX PREPROCESSING

Markov Cluster (MCL) is a fast graph clustering algorithm proposed by Dongen. The advantage of MCL is fewer prior conditions which do not need clustering numbers in advance [27]. We use the random-walk way to supplement sparse matrix information. This approach can reflect the possibility of a connection between users, which can consider the users' relationship weight. MCL not only simply simulates the random walkway as reality, but also constantly modifies the transition probability matrix. It repeats ''expansion'' and ''inflation'' until the state of the matrix reaches a definite value. In this paper, we just need the matrix to cluster to fuzzy classes, which will be defined as the initial community number. We use a sample model network as shown in Figure 1 to introduce the improved MCL algorithm applications in this paper.

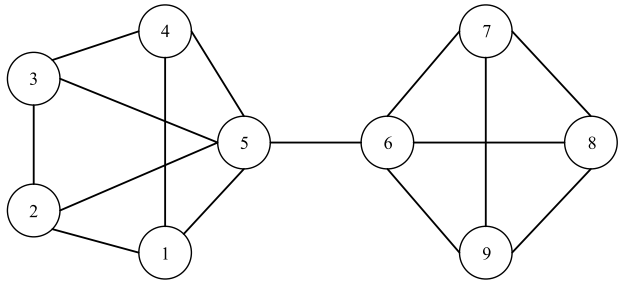

FIGURE 1. Sample model network.

We can get the initial adjacency matrix as X

$$X = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

First, we calculate the initial adjacency matrix through self-loop to reduce the effect from the null diagonal value when expansion occurs with an odd and even power. The self-loop adds an edge for the node self. Then, we can get the matrix

$$\inf X^{*}=\begin{bmatrix} 0.1659 & 0.1366 & 0.1366 & 0.1366 & 0.1393 & 0.0006 & 0.0000 & 0.0000 & 0.0000\\ 0.1366 & 0.1659 & 0.1366 & 0.1366 & 0.1393 & 0.0006 & 0.0000 & 0.0000 & 0.0000\\ 0.1366 & 0.1366 & 0.1659 & 0.1366 & 0.1393 & 0.0006 & 0.0000 & 0.0000 & 0.0000\\ 0.1366 & 0.1366 & 0.1366 & 0.1659 & 0.1393 & 0.0006 & 0.0000 & 0.0000 & 0.0000\\ 0.4229 & 0.4229 & 0.4229 & 0.4229 & 0.4361 & 0.0057 & 0.0017 & 0.0017 & 0.0017\\ 0.0013 & 0.0013 & 0.0013 & 0.0013 & 0.0038 & 0.2897 & 0.2819 & 0.2819 & 0.2819\\ 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0009 & 0.2340 & 0.2388 & 0.2388 & 0.2388\\ 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0009 & 0.2340 & 0.2388 & 0.2388 & 0.2388\\ 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0009 & 0.2340 & 0.2388 & 0.2388 & 0.2388 \end{bmatrix}$$

$X^{'}=X+I_n$, where $I_n=diag(1,1,\cdots,1)$. We normalize the matrix $X^{'}$ as X^{*}.

$$X^{*}=\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0\\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0\\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0\\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0\\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4}\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4}\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

The matrix is no longer made up of 0 and 1 elements, but is formed by the close relationship between nodes through 'expansion' and 'inflation'. The "expansion" and "inflation" process can result in prominent tight node relationships and enlarge weak node relations. We consider the clustered result as the initial community number used in NMF. Therefore, the MCL clustering process should not cluster until reaching a steady state. In this research, we set the MCL proceedings containing one "expansion" and two "inflation" factors. The matrix $powX^{*}$ is the "expansion" result of the simple network.

$$powX^{*}$$
$$=\begin{bmatrix} 0.23 & 0.17 & 0.17 & 0.17 & 0.15 & 0.03 & 0 & 0 & 0\\ 0.17 & 0.23 & 0.17 & 0.17 & 0.15 & 0.03 & 0 & 0 & 0\\ 0.17 & 0.17 & 0.23 & 0.17 & 0.15 & 0.03 & 0 & 0 & 0\\ 0.17 & 0.17 & 0.17 & 0.23 & 0.15 & 0.03 & 0 & 0 & 0\\ 0.23 & 0.23 & 0.23 & 0.23 & 0.22 & 0.07 & 0.05 & 0.05 & 0.05\\ 0.04 & 0.04 & 0.04 & 0.04 & 0.06 & 0.22 & 0.24 & 0.24 & 0.24\\ 0 & 0 & 0 & 0 & 0.03 & 0.19 & 0.24 & 0.24 & 0.24\\ 0 & 0 & 0 & 0 & 0.03 & 0.19 & 0.24 & 0.24 & 0.24\\ 0 & 0 & 0 & 0 & 0.03 & 0.19 & 0.24 & 0.24 & 0.24 \end{bmatrix}$$

In the matrix $powX^{*}$, the nodes 5 and 6 can connect every node in the network. In the real network, nodes 5 and 6 are indirect neighbor nodes with each node in Figure 1, which influence the community structure. Therefore, MCL can make the indirect nodes relationship abundant in a sparse matrix. With the inflation process, we can get matrix $\inf X^{*}$, as shown at the top of this page.

Analyzing the matrices above, we find the probability always changes realistically. The probability between node 6 and nodes 1 to 5 is reduced because they belong to different communities. The phenomenon shows that the nodes in different communities have less influence than the nodes in the same community that is fit for defining the community structure.

B. NON-NEGATIVE MATRIX FACTORIZATION WITH ORTHOGONAL AND SPARSENESS CONSTRAINT

We aim to improve the accuracy of NMF based on the physical meaning of matrix factorization. The factorization can be considered that each data vector v (the row of V) is approximated by a linear combination of the rows of H weighted by the components of w (the row of W): $v=wH$. We can see that relatively few basis vectors are used to represent many data vectors and the entries of w represent the weight of every basis vector to produce the data vector v.

We can use matrix W to group the n objects into k clusters. The entries in W can be viewed as the memberships of every node to each community. The matrix W is the base matrix reflecting the feature in the network. The more orthogonal the vectors in the matrix, the clearer the matrix can express the network characteristics. On the one hand, the orthogonal constraint can normalize the elements of the base matrix; on the other hand, it can make $w_i^T w_j=0, i\neq j$ to ensure the factor in matrix W reflects the network characteristics.

Matrix H is a basis that is optimized for the linear approximation of the feature data in V. The matrix H can be regarded as membership weight matrix which indicates the probability of the node belonging to the community. Sparse constraint refers to the use of a small number of elements in a collection representing all elements of the data. The application of sparse constraints in the NMF algorithm not only reduces the interference of noise and the running time of the algorithm, but also improves the recognition rate in the classification.

We used $\inf X^{*}$ as the matrix X for subsequent NMF decomposition. In this research, we set the MCL proceedings containing one "expansion" and two "inflation" factors, because we need the outputs of MCL containing the communities number cursorily. We utilized the initial communities number as the dimension of NMF decomposition. It is significance to employ NMF with constrains to utilize the MCL outputs because it can divide network structure without prior community number.

$$\begin{aligned}\frac{\partial D}{\partial W_{ab}} &= \frac{\partial(\frac{1}{2}\sum_{ij}(V_{ij}^2 - 2V_{ij}(WH)_{ij} + (WH)_{ij}^2) + \frac{1}{4}\alpha\sum_i\sum_j\left[(WW^T)_{ij} - I_{ij}\right]^2)}{\partial W_{ab}} \\ &= \frac{\partial(\frac{1}{2}\sum_{ij}(WH)_{ij}^2 - 2V_{ij}(WH)_{ij} + \frac{1}{4}\alpha\sum_i\sum_j\left[(WW^T)_{ij} - I_{ij}\right]^2}{\partial W_{ab}} \end{aligned} \tag{7}$$

NMFOSC modifies the existing NMF methods and adapts them to large networks. Although NMF methods use l_2 norm as an objective function, l_2 norm is not suitable for modeling binary adjacency matrices. Thus, we combined l_1 norm and l_2 norm and proposed the objective function as follows:

$$\begin{aligned} D = \frac{1}{2}\sum_{ij}(V_{ij} - (WH)_{ij})^2 - \frac{\alpha}{4}\sum_{ij}\left\|\left(W^TW\right)_{ij} - I_{ij}\right\|^2 \\ - \frac{\beta}{2}\sum_{ij}H_{ij}^2, \\ s.t.\ \alpha \geq 0, \quad \beta \geq 0,\ W \geq 0,\ H \geq 0, I_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \end{aligned} \tag{4}$$

where α and β are positive numbers, meaning the orthogonal constraint and sparseness constraint are separate. We introduced the *sparseness* function proposed by PO Hoyer to balance the sparseness in matrix H. Based on the requirement that the sum of non-negativity H elements vectors equal to 1, maximizing $\|H\|_2^2$ can make vector h get higher sparsity. Under the constraint with $\sum_i H_{ij} = 1$, we make the decomposition results as sparse as possible in order to get the main features. The spares objective function is maximizing $\|H\|_{ij}^2$.

$$sparseness(x) = \frac{\sqrt{n} - (\sum|x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{5}$$

where n is the dimension of x. We set the sparseness to 0.5 to avoid conforming a large sparseness matrix. To solve this convex problem, we use a projected gradient ascent. The gradient can be computed straightforwardly.

$$W_{ik}^{n+1} \leftarrow W_{ik}^n + \lambda_1\frac{\partial D}{\partial W}, \quad H_{kj}^{n+1} \leftarrow H_{kj}^n + \lambda_2\frac{\partial D}{\partial H} \tag{6}$$

where $\frac{\partial D}{\partial W}$ and $\frac{\partial D}{\partial H}$ are the partial derivative of matrix W and H separately. The calculation process is as (7), as shown at the top of this page, where

$$\begin{aligned} &\frac{\partial}{\partial W_{ab}}\sum_{ij}(WH)_{ij}^2 \\ &= \frac{\partial}{\partial W_{ab}}\sum_j((W_{ab}H_{bj})^2 + 2W_{ab}H_{bj}\sum_{s\neq b}(W_{as}H_{sj})) \\ &= \sum_j(2W_{ab}H_{bj}^2 + 2H_{bj}\sum_{s\neq b}(W_{as}H_{sj})) \\ &= 2\sum_j(H_{bj}\sum_s(W_{as}H_{sj})) \end{aligned} \tag{8}$$

$$\begin{aligned} &\frac{\partial}{\partial W_{ab}}\sum_{i,j}(-2V_{ij}(WH)_{ij}) \\ &= -\sum_j 2V_{aj}H_{bj} \end{aligned} \tag{9}$$

$$\begin{aligned} &\frac{\partial}{\partial W_{ab}}(\sum_i\sum_j\left[(WW^T)_{ij} - I_{ij}\right]^2) \\ &= \frac{\partial}{\partial W_{ab}}\sum_{ij}\left[(WW^T)_{ij}^2 - 2I_{ij}(WW^T)_{ij}\right] \end{aligned} \tag{10}$$

$$\begin{aligned} &\frac{\partial}{\partial W_{ab}}\sum_{ij}(WW^T)_{ij}^2 \\ &= \frac{\partial}{\partial W_{ab}}\sum_{ij}(W_{ib}W_{jb} + \sum_{t\neq b}W_{it}W_{jt})^2 \\ &= \frac{\partial}{\partial W_{ab}}\sum_{ij}(W_{ab}^4 + 2W_{ab}^2\sum_{t\neq b}W_{at}^2) + \sum_{j\neq a}(W_{ab}^2W_{jb}^2 \\ &\quad + 2W_{ab}W_{jb}\sum_{t\neq b}W_{at}W_{jt}) + \sum_{i\neq a}(W_{ab}^2W_{ib}^2 \\ &\quad + 2W_{ab}W_{ib}\sum_{t\neq b}W_{it}W_{at}) \\ &= 4\sum_{i=1}\sum_{t=1}W_{at}W_{it}W_{ib} \end{aligned} \tag{11}$$

$$\begin{aligned} &\frac{\partial}{\partial W_{ab}}(\sum_{ij}-2I_{ij}(WW^T)_{ij}) \\ &= \frac{\partial}{\partial W_{ab}}(\sum_{ij}-2I_{ij}W_{ib}W_{jb}) \\ &= -4I_{ab}W_{ab} - \sum_{i\neq a}^m(2I_{ia}W_{ib}) - \sum_{j\neq a}^m(2I_{aj}W_{jb}) = -4W_{ab} \end{aligned} \tag{12}$$

$$\begin{aligned} &\frac{\partial D}{\partial W_{ab}} \\ &= (WHH^T)_{ab} - (VH^T)_{ab} + \alpha(WW^TW - W)_{ab} \end{aligned} \tag{13}$$

$$\begin{aligned} &\frac{\partial D}{\partial H_{xy}} \\ &= \frac{\partial(\frac{1}{2}\sum_{i,y}(V_{iy} - (WH)_{iy})^2 + \frac{1}{2}\beta\sum_{i,y}H_{iy}^2)}{\partial H_{xy}} \\ &= \frac{\partial(\frac{1}{2}\sum_{i,y}(-2V_{iy}(WH)_{iy}) + (WH)_{iy}^2 + \frac{1}{2}\beta\sum_{i,y}H_{iy}^2)}{\partial H_{xy}} \\ &= \frac{-\partial\sum_{i,y}V_{iy}(WH)_{iy}}{\partial H_{xy}} + \frac{\frac{1}{2}\partial\sum_{i,y}(WH)_{iy}^2}{\partial H_{xy}} + \frac{\partial\frac{1}{2}\beta\sum_{i,y}H_{iy}^2}{\partial H_{xy}} \\ &= \sum_i W_{ix}(WH)_{iy} - \sum_i W_{ia}V_{iy} + \beta H_{iy} \\ &= (W^TWH)_{xy} - (W^TV)_{xy} + \beta H_{xy} \end{aligned} \tag{14}$$

Thus, we set the step size in the negative gradient direction to λ_1 and λ_2:

$$\lambda_1 = \frac{W_{ab}}{(WHH^T)_{ab.} + \alpha(WW^TW)_{ab}} \tag{15}$$

$$\lambda_2 = \frac{H_{xy}}{(W^TWH)_{xy} + \beta H_{xy}} \tag{16}$$

For NMFOSC algorithm, the computational cost is governed chiefly by MCL process and formula (4). Notice that, the time to calculate the processing of MCL, W_{ik}^{n+1} and H_{kj}^{n+1} are $O(n^3)$, $O(n^2k)$ and $O(n^2k)$, respectively. Hence, the total computational cost of NMFOSC is $O(n^3 + n^2k)$, where k is the number of communities.

C. FILE FORMATS FOR GRAPHICS

In network preprocessing, we can get the initial community number by MCL. In this section, we update the number and attribution of network communities based on the results of NMFOSC. The matrix H analyzes the probability of each node belonging to every community, shown as h_{ci}. We rank the membership probability to optimize the community structure.

$$R(i,c) = \frac{h_{ci} - \min h_{ci}}{\max h_{ci} - \min h_{ci}} \tag{17}$$

where, $\min h_{ci}$ is the minimum non-zero membership probability of node i, and $\max h_{ci}$ is the maximum non-zero membership probability of node i. We compared the ranking with the threshold and found that the community c occupies the weight of node i. The threshold defined as σ.

$$\sigma = \frac{h_{ci}}{\sum_{c=1}^{k} h_{ci}} \tag{18}$$

If more than one $R(i,c)$ is greater than the threshold σ, the node i is an overlapping node. Therefore, NMFOSC is an overlapping community detection algorithm. If the R value is less than the threshold σ, we delete the node i in community c. This optimal algorithm can show the relationship among the various communities for the node independently and division of the network community accurately.

IV. EXPERIMENT

We test the performance of the method proposed here by applying it to a class of artificial networks and some real-world networks. We demonstrate the effectiveness of our method on a range of networks from large different domains and research areas. And if there is no special mention, we choose $\alpha = 0.2$ and $\beta = 0.1$ in the feature matrices in our study. The language of choice for all implementations is Java according to the JDK 1.6 standard, allowing us to use object-oriented and functional programming concepts while also compiling to native code. The experimental environment showed as Table 1.

TABLE 1. Experimental environment.

Environmental category	Conversion from Gaussian and CGS EMU to SI [a]
Hardware *CPU*	Intel (R) Xeon () CPU (R), 4G memory Intel(R) Xeon(R) E5-2620v3 @ 2.40GHz, 64 threads
Development environment	Eclipse 32, 64bit java version 1.6.0_02

A. DATASETS AND EVALUATION CRITERIA

The following list outlines the different types of graphics published in IEEE journals. They are categorized based on their construction, and use of color / shades of gray:

1) DATESETS

The following list shows the different kinds of datasets in Table 2.

TABLE 2. Datasets.

Network	Vertices	Edges	Description
Karate	34	78	Zachary's karate club
Dolphins	62	159	Dolphins Dolphin social network
Football	115	613	Football American College football
Southern Women Data	32	89	Southern Women bipartite nnetwork
Polblogs	1490	16718	Blogs about politics
Netsci	1589	2742	Network scientists

2) EVALUATION CRITERIA

There are various standard measures that can be used to measure the performance of community structure delivered by the algorithm. This paper used average conductance (AC) of communities with weights, which extends the conductance used by Leskovec, mapping the weighted value of conductance for all the communities in a cover [28]. The conductance can be simply thought of as the ratio between the number of edges inside the community and those leaving it. More formally, the conductance is defined as follows:

$$\phi(S) = C_S \big/ \min(Vol(S), Vol(V \backslash S)) \tag{19}$$

where $C_S = |\{(u,v) : u \in S, V \notin S\}|$, $Vol(S) = \sum_{u \in S} d_u$, and d_u is the degree of vertex u. Thus, more community-like sets of vertices have lower conductance. Consequently, the AC can be defined as

$$AC = \frac{\sum_{i=1}^{K} N(C_i)\phi(C_i)}{\sum_{i=1}^{K} N(C_i)} \tag{20}$$

where K denotes the number of communities, C_i denotes the ith community, and $N(C_i)$ denotes the number of vertices in C_i.

Another commonly used measure is normalized mutual information (NMI) which has become a de facto standard for the networks with known communities [29].

$$NMI(A,B) = \frac{2I(A,B)}{H(A) + H(B)} \tag{21}$$

where A and B denote the two partitions of the network. If the found communities are identical to the real communities, then $NMI(A,B)$ takes its maximum value of 1. If the found communities are totally independent of the real partition, for example when the entire network is classified to be one community, $NMI(A,B) = 0$.

In the experimental datasets, the Southern Women network is a bipartite network. We used the modularity fit for the bipartite network community structure. Baber considered the two types of nodes and proposed an improved modularity called Q_B [9].

$$Q_B = \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{m}(A(i,j) - P(i,j))\delta(g_i, g_{j+n}) \qquad (22)$$

where $P(i,j)$ is the expected value of the edges between node i and node j. g_i means the community to which node i belongs. The bipartite network has two kinds of nodes, which one contains n nodes and another contains m nodes. Hence, node i means the node belongs to the one kind of node set. Node $j+n$ means the node belong to another kind of node set. The function δ is used to judge whether node i and node $j+n$ are in the same community.

Murata proposed another modularity formula that considered communities' relationships from a different perspective [10].

$$Q_M = \sum\nolimits_l (e_{lm} - a_l a_m), m = \arg\max(e_{lk}) \qquad (23)$$

where m represents the closest connection between heterogeneous communities. e_{lm} means the actual number of edges connecting community l with community m. a_l means the expected value of the edges between community l and community m. The higher Q_M is, the more clearly divided the community structure is.

B. ARTIFICIAL NETWORKS

The artificial network consists 128 nodes and 1024 edges. It contains 4 communities, and each community contains 32 nodes. We can adjust the fuzzy parameter μ and overlap degree O_M to change the network community structure. We found that the accuracy of algorithms decreases as μ increases. When $\mu \in [0.1, 0.2]$, the structure of the community is clear, and all algorithms can effectively divide the community. In Figure 2, as μ increases until 0.5, our algorithm becomes $NMI = 1$ on the network, but others begin to decay. Even the mixing parameter increases to 0.7, and the value of NMI is still greater than 0.8, which means a good match with the community structure of the original network. When $\mu = 0.7$, the community structures are difficult to successfully detect, because they have become indistinct in these networks. Generally speaking, the performance of the NMFOSC is still better than other baseline algorithms.

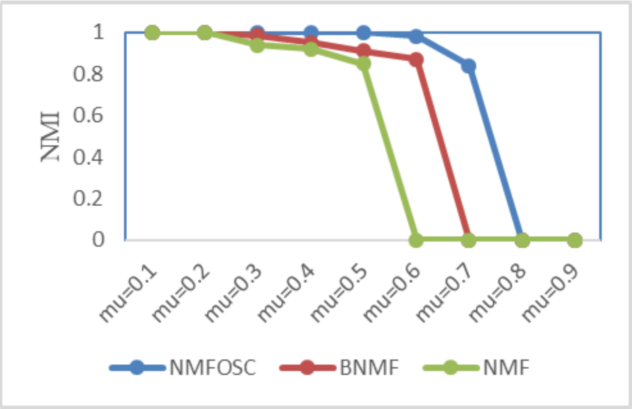

FIGURE 2. NMI with different mixing parameter in benchmark networks.

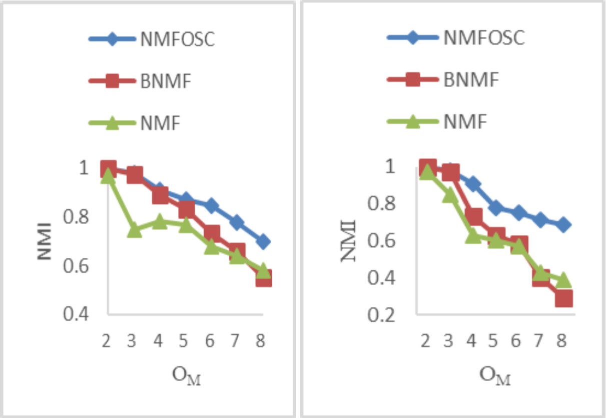

FIGURE 3. NMI with a function of number of members in Om benchmark network (left: μ = 0.1 , right: μ = 0.3).

In order to assess the performance of detecting overlapping communities, we compared the accuracy of various algorithms on two types of overlapping benchmark networks with different O_m. The results are shown in Figure 3. It can be observed that the NMFOSC can accurately identify the overlapping communities in networks. With the increase in O_m, the accuracy decreases especially when $O_m > 3$. However, our algorithm can decay more slowly than others and keeps the highest NMI in these three algorithms. In addition, the results of our algorithm are very stable.

The bipartite artificial network contains 8 communities, and in each community, there are 32 users nodes and 14 target nodes. We can adjust the homogeneity coefficient ρ to change the network community detection. When $\rho = 1$, the users connect only to the target nodes in the same community. There is no connection between different communities. When $\rho = 0$, the nodes connect with each other randomly, it is hard to identify the community structure. Figure 4 shows the performance of different algorithms in different networks. With the increase of the ρ value, the NMI value increases gradually. We can find the NMI value is low when ρ is less than 0.55. However, NMFOSC can make the community divided MNI value reach 0.6, which is greater than others. Meanwhile, the NMI value raised by this algorithm is faster than other algorithms. When ρ is bigger than 0.55, the NMI value caused by NMFOSC is greater than 0.8 which means the community structure detected is close to the real networks. The NMFOSC considered the community relationship weight ranking to optimize the matrix factorization result to networks' reality.

C. REAL-WORLD NETWORKS

We evaluated the performance of the algorithm NMFOSC with other classical community detection algorithms using some real-world networks. The methods compared include the Louvain method, which is regarded as one of the

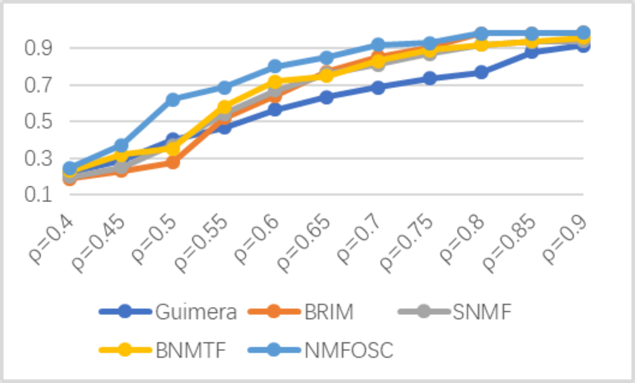

FIGURE 4. NMI value in an artificial network.

best options for vertex partition; the clique percolation method (CPM), which is the most prominent algorithm for overlapping community detection; and BNMF and BNMTF, which are both community detection methods based on NMF.

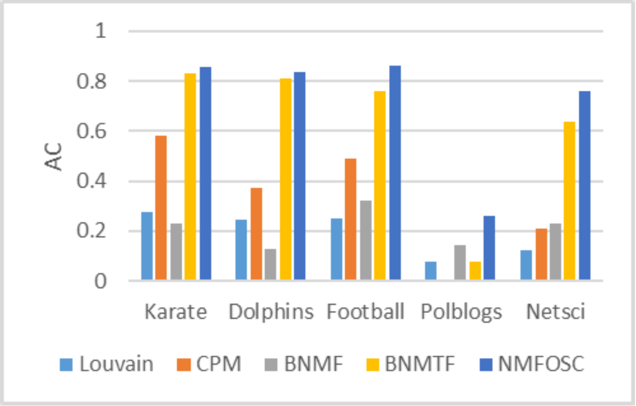

FIGURE 5. AC value in real-world network.

Figure 5 shows the results of different algorithms in terms of average conductance. Generally, the performance of NMFOSC is still better than the other four algorithms in terms of the AC quality. To sum up, our algorithm is very effective on real-world networks in terms of both accuracy and quality. Therefore, as we can see, NMFOSC can not only detect three types of vertices roles, providing richer information from networks, but also find community results with high accuracy and quality.

We experiment with NMFOSC in real-world bipartite networks at the same time. The Southern Women Data network contains 18 Southern women and 14 social activities used to verify the performance of community discovery algorithms for bipartite networks widely. Figure 6 shows the Southern Women activity relationship network, where nodes 1-18 mean Southern Women and nodes 19–32 mean activities.

From the perspective of ethnology, Davis et al. divided the user nodes in two sub networks into two communities: {1-9} and {9-18}, separately. BRIM algorithm divided the network into 4 communities: {1-6,19-24}, {7,9,10,25,26}, {8,16-18, 27,29}, and {11-15,28,30-32}, separately. In Figure 6, nodes 8 and 16 joined activities 26 and 27, but did not join activity 29. However, the BRIM divided nodes 8 and 16 into community {8,16-18,27,29}, which is obviously unreasonable. We used NMFOSC to divide the network community structure shown in Figure 7.

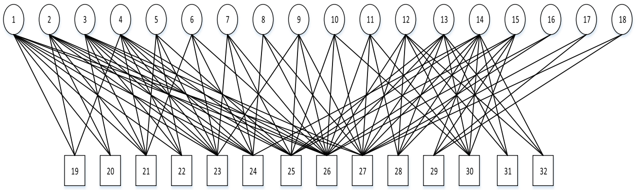

FIGURE 6. Numbered southern women bipartite network.

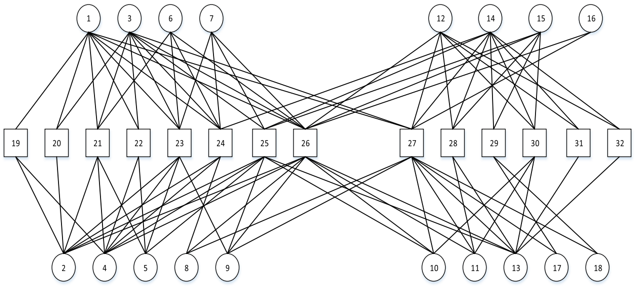

FIGURE 7. Community structure of southern women bipartite network.

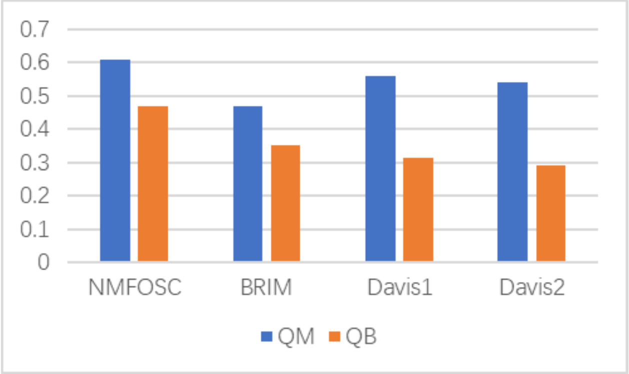

FIGURE 8. Modularity of southern women bipartite network.

We also compared the communities' results with Q_B and Q_M in Figure 8. As node 9 is an overlapping node in Davis research, we computed the results in two modes, such as Davis 1 {1-9} and Davis 2 {9-18}.

To sum up, our method with iterative bipartition not only has a higher clustering quality compared with other methods, but can also determine the number of communities automatically. Thus, it may be more suitable for use when detecting communities on unexplored real networks.

V. CONCLUSION

In this paper, we present a method based on the NMF technique to divide community structure in complex networks. The proposed algorithm combined MCL pretreatment and ranking optimization to solve the unknown community number in an advance problem. The NMFOSC algorithm used orthogonal and sparseness constraints to optimize matrix decomposition results. It makes the membership matrix more orthogonal to reflect the network feature and the membership weight matrix sparser to show the main relationship. As in a real-world network, it is natural that some nodes belong to more than one community. The ranking optimization processes allows nodes to belong to different communities at the same time. The experimental results demonstrate that the proposed method has good performance on both the artificial benchmark networks and some real-world networks.

REFERENCES

[1] A. Lancichinetti and S. Fortunato, "Consensus clustering in complex networks," *Sci. Rep.*, vol. 2, no. 13, 2012, Art. no. 336.

[2] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22073–22078,2009.

[3] K. Yamada and K. Knight, "A syntax-based statistical translation model," in *Proc. Meeting Assoc. Comput. Linguistics*, 2001, pp. 523–530.

[4] K. Jahanbakhsh, V. King, and G. C. Shoja, "Predicting missing contacts in mobile social networks," *Pervasive Mobile Comput.*, vol. 8, no. 5, pp. 698–716, 2012.

[5] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[6] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[7] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, p. 026113, Feb. 2004.

[8] R. Guimer, M. Sales-Pardo, and L. A. N. Amaral, "Module identification in bipartite and directed networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, p. 036102, Sep. 2007.

[9] M. J. Barber, "Modularity and community detection in bipartite networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, p. 066102, Dec. 2007.

[10] T. Murata, "Detecting communities from bipartite networks based on bipartite modularities," in *Proc. Int. Conf. Comput. Sci. Eng.*, Aug. 2009, pp. 50–57.

[11] K. Suzuki and K. Wakita, "Extracting multi-facet community structure from bipartite networks," in *Proc. Int. Conf. Comput. Sci. Eng.*, Aug. 2009, pp. 312–319.

[12] L. Šubelj and M. Bajec. (Mar. 2011). "Unfolding network communities by combining defensive and offensive label propagation." [Online]. Available: https://arxiv.org/abs/1103.2596

[13] M. J. Barber and J. W. Clark, "Detecting network communities by propagating labels under constraints," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, p. 026129, Aug. 2009.

[14] X. Liu and T. Murata, "Advanced modularity-specialized label propagation algorithm for detecting communities in networks," *Phys. A, Stat. Mech. Appl.*, vol. 389, no. 7, pp. 1493–1500, 2010.

[15] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[16] E. Ravasz and A. L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 67, p. 026112, Feb. 2004.

[17] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki, "Sequential algorithm for fast clique percolation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 2, p. 026109, 2008.

[18] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, pp. 19–44, 2008.

[19] S. Gregory, "An algorithm to find overlapping community structure in networks," in *Proc. Eur. Conf. Principles Data Mining Knowl. Discovery*, 2007, pp. 91–102.

[20] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, pp. 2011–2024, 2009.

[21] D. Lee, H. Sohn, G. V. Kalpana, and J. Choe, "Interaction of E1 and hSNF5 proteins stimulates replication of human papillomavirus DNA," *Nature*, vol. 399, no. 6735, pp. 487–491, 1999.

[22] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[23] W. Liu, N. Zheng, and X. Lu, "Non-negative matrix factorization for visual coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3. Apr. 2003, p. III-293-6.

[24] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, no. 1, pp. 1457–1469, 2004.

[25] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Dec. 2001, pp. I-207–I-212,

[26] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for weakly-supervised multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 121–135, Jan. 2015.

[27] S. M. van Dongen, "Graph clustering by flow simulation," Ph.D. dissertation, Dept. Math. Comput., Utrecht Univ. Repository, Utrecht, The Netherlands, 2000.

[28] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 631–640.

[29] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mech., Theory Experim.*, vol. 2005, no. 9, p. 09008, 2005.

NAIYUE CHEN received the master's degree in electronics and information engineering from Beijing Jiaotong University, China, in 2013, where she is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering. Her current research fields include community detection, trend prediction, and network analysis.

YUN LIU received the Ph.D. degree in electronic engineering from Beijing Jiaotong University in China. She is currently an IET Fellow, a Professor, and the Director of the Department of Communication and Information System, and the Director of the Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University. She has edited many books and published over 200 papers and book chapters, and participated in many international academic activities, including the organization of several international conferences. Her current research interests include telecommunication, computer networks, network security, intelligent transportation system, and social dynamics.

HAN-CHIEH CHAO (SM'16) received the M.S. and Ph.D. degrees in electrical engineering from Purdue University, in 1989 and 1993, respectively. He is currently a joint appointed Distinguished Professor of the Department Computer Science and Information Engineering and Electronic Engineering, National Ilan University (NIU), I-Lan, Taiwan. He has been serving as the President with NIU since 2010. He has authored or co-authored five books and has published about 400 refereed professional research papers. His research interests include high speed networks, wireless networks, IPv6 based networks, digital creative arts, and e-government and digital divide. He was an officer of Award and Recognition for IEEE Taipei Section from 2010 to 2012. He is a Fellow of IET (IEE).