Overlapping Community Detection via Self-constrained Symmetric Non-negative Matrix Factorization

Yu Liu, Bin Wu, Yunlei Zhang, Bai Wang
Beijing Key Laboratory of Intelligence Telecommunication Software and Multimedia
Beijing University of Posts and Telecommunications
Beijing, 100876, China
Email: liuyu@bupt.edu.cn, wubin@bupt.edu.cn, yunlei0518@126.com, wangbai@bupt.edu.cn

***Abstract*—A number of approaches based on symmetric non-negative matrix factorization (SNMF) have been proposed to improve the performance and the interpretability of community detection. Due to the nature of NMF, the partition results obtained by conventional NMF without post processing are soft assignments of nodes w.r.t. communities, which demonstrates overlapping of communities.**

Based on the traditional SNMF method, we propose a self-constrained symmetric non-negative matrix factorization (SCSNMF) with tuning ability to control the degree of community overlapping, which controls if the community partition result is "most overlapping", "nearly overlapping" or "nearly non-overlapping". We use both traditional and overlapping version of modularity and partition density to investigate community overlapping on five real-world social network datasets. The experimental results show that SCSNMF has the ability of interpretation for overlapping degree of communities.

I. Introduction

In complex network science, community detection for social network, collaborative network, biological network, etc., is an important means to understand and analysis network clustering problem. According to node similarity or other information, nodes in networks can be grouped into communities, that are sharing common attributes, features or functions. With the help of identified communities in network, numerous data mining tasks can be improved, such as social-based recommendations for both individuals [1] and groups [2], combinatorial clustering [3]. It is a common understanding that nodes are densely connected within same community, while sparsely connected in different communities [4].

It is a common phenomenon that individuals in a social network may be assigned with multiple community memberships. For example, an individual can have different roles when he/she has connections with different kinds of group of people, e.x. family, coworkers, friends. Previous studies showed that the overlap in communities is a significant feature of numerous real-work networks.

Algorithms for overlapping community detection can be categorized into different classes that distinguish the way communities are discovered [5]. Clique Percolation based methods (CPM) [6] considers that a community is a combination of fully connected subgraphs and communities are discovered by finding adjacent cliques, which has great performance on graphs having densely connected component. Link Partitioning based methods (LP) [7] groups links instead of nodes when communities are identified in a line graph, which seems conceptually natural. However, LP methods still rely on an ambiguous definition of community as node-based community detection approaches do. Agent-based methods, such as label propagation algorithms, SLPA [8], COPRA [9] can identify community structure in nearly linear time.

Another class of approaches for overlapping community detection bases on non-negative matrix factorization (NMF). Matrix factorization is an algorithm for feature extraction, dimension reduction and clustering. NMF has the advantage in graph mining, since the adjacent matrix of a graph is naturally non-negative, which makes the use of NMF method will be interpretable. An NMF algorithm approximately factorizes the non-negative matrix $\mathbf{V}$ into two matrices with the non-negativity constraint as $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{V}$ is $n \times m$, $\mathbf{W}$ is $n \times k$, $\mathbf{H}$ is $k \times m$, and k is the number of communities provided by users. $\mathbf{W}$ represents the data in the reduced feature space. Each element $\mathbf{W}_{ic}$ in the normalized $\mathbf{W}$ quantifies the dependence of node i with respect to community c. In addition to directly using the adjacent matrix, diffusion kernel matrix is used as the matrix to be factorized in [10], node similar and node neighborhood ratio matrix is also exploited in [11] and [12] respectively. A Bayesian based method is proposed in [13] to automatically identify the number of communities.

NMF based methods have the ability to assign nodes with soft membership that will demonstrate how much a node will belong to a community and how many communities will a node belong to. However, to the best of our knowledge, there exists no method that could investigate the degree of overlapping. Leveraging community detected with different degree of overlapping will have a great impact on several tasks of social computing, such as group recommendation [2]. To tackle this problem, we investigate this problem and make come contributions in this paper:

- We propose a self-constrained symmetric non-negative matrix factorization (SCSNMF) based on [14] for community detection in undirected and unweighted networks.

- The proposed method has the ability to tune the degree of overlapping. It means that the overlaps of communities detected can be tuned for a peculiar network for different oriented community based research, such as the study of group recommendation in [2].
- Elementary experiments on the effectiveness of the proposed method are conducted to show the impact of community overlapping tuning ability on network modularity [15] and partition density [7].

The rest of the paper is organized as follows. Section II reviews related work on non-negative matrix factorization and NMF based community detection. In Section III, the proposed SCSNMF model is introduced. Experiments on five real-world network is carried out to demonstrate the effectiveness of overlapping degree tuning in Section IV. We conclude our work and provide some possible future work in Section V.

II. Related Work

A. Non-negative Matrix Factorization

Traditional matrix factorization decomposes a matrix into two or three matrices that have no any additional constraints. These methods embrace singular value decomposition (SVD) [16], probabilistic matrix factorization (PMF) [17], etc. However, analyses of images [18], graph networks, user-item ratings, etc. usually require non-negativity in the results of matrix factorization in consideration of interpretations, which can be solved through non-negative matrix factorization.

Given a series of observed data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ that form a data matrix denoted by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^\top \in \mathbf{R}^{m \times n}$, where the data point x_i is an *n*-dimensional vector whose elements are non-negative, a NMF approach aims to decompose $\mathbf{X}$ into a product of two or three non-negative matrices. The procedure can be modeled as follows,

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top,\ \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}, \tag{1}$$

where matrices $\mathbf{U}^{m \times d}$ and $\mathbf{V}^{n \times d}$ are factorized results, d is an integer. For the purpose of solving the matrix approximation problem, an error function (or optimization function) is used to quantify the approximation errors. For methods using Frobenius norm, the error function to be minimized is

$$\mathcal{J} = \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}^2,\ \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}, \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm. The sizes of $\mathbf{U}$ and $\mathbf{V}$ are respectively $m \times d$ and $n \times d$. In order to represent expressions of $\mathbf{X}$ in the space of latent factors with reduced dimensions, d is usually set such that $d \ll m$ and $d \ll n$. Algorithms to solve the objective function (2) include Bayesian NMF [13], multiplicative updating method [19].

B. NMF based Community Detection

The technology of NMF is exploited to discover group (community) partition in graph networks that are usually modeled as non-negative matrices. An unweighted graph containing n nodes with m edges can be denoted by an adjacent matrix $\mathbf{G}^{n \times n} = [g_{ij}]$, where g_{ij} represents the relationship between node i and node j. $g_{ij} = 1$, if there is an edge from node i to node j; $g_{ij} = 0$, otherwise. Thus, $m = \sum_{ij} g_{ij}$ for an directed graph, and $m = \frac{1}{2}\sum_{ij} g_{ij}$ for undirected one. The task of community detection for a given graph $\mathbf{G}$ can be modeled using NMF method as follows,

$$\mathcal{J}_{\mathrm{NMF}} = \frac{1}{2}\|\mathbf{G} - \mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}^2,\ \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}, \tag{3}$$

where the matrices $\mathbf{U}$ or $\mathbf{V}$ represents relationships between nodes and a partition, and the latent space number d is the number of communities.

For an undirected network, the adjacent matrix $\mathbf{G}$ is symmetric. The community detection tasks reduce to symmetric non-negative matrix factorization (SNMF):

$$\mathcal{J}_{\mathrm{SNMF}} = \frac{1}{2}\|\mathbf{G} - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2,\ \mathbf{U} \geq \mathbf{0}. \tag{4}$$

Each row vector contained in the resulting matrix $\mathbf{U}$, denoted by $\mathbf{U}_i$ ($i = 1, 2, \ldots, n$), indicates the partition for node i. An element u_{ic} in matrix $\mathbf{U}$ suggests whether node i belongs to community c; or how much (membership coefficient) node i belongs to community c, since the results of NMF bring soft membership assignments, resulting overlapping communities. Therefore, NMF based methods are usually exploited to identify overlapping communities.

Many NMF based community detection methods have been proposed to deal with numerous problems. Wang et al. [14] proposed three NMF techniques, i.e., Symmetric NMF, Asymmetric NMF and Joint NMF, to solve the community detection in undirected, directed and compound networks, respectively.

Psorakis et al. [13] leveraged Bayesian approach to automatically identify the number of communities and the partition of groups in networks so that the partition can achieve the highest Newman modularity Q [15].

The membership coefficients assigned by an NMF algorithm has the ability to express overlapping communities. However, a membership coefficient does fail to determine whether a node belongs to a community or not. Zhang et al. [20] proposed a method based on symmetric binary NMF to deal with this kind of inability.

Original NMF based methods factorize the adjacent matrix. Some work uses different matrices. A Laplacian kernel matrix is used in NMF to achieve higher modularity in [10]; A neighborhood ration matrix is factorized instead of directly using of the adjacent matrix by Eustace [12]; A nonnegative similarity matrix is also used in graph clustering in [11].

Traditional NMF or SNMF methods using Equations (3) and (4) reveal limited improvement in performance. Additional information is leveraged to enhance the ability to identify more accurate partition of overlapping or non-overlapping communities through NMF. For example, label data as additional input is used in semi-supervised methods to identify overlapping communities in [21] and [22], respectively. Content data, such as topics and messages, in online social networks is exploited to achieve better partition result in [3].

III. Community Detection via Self-constrained Symmetric NMF

In this section, we propose a community detection model via self-constrained symmetric non-negative matrix factorization. Here the constraints will not exploit any additional data other than the given adjacent matrix of a graph.

A. The Self-constrained Symmetric NMF model

Given an undirected unweighted network $\mathbf{G}$ containing n nodes with m edges, the task of community detection can be modeled through Equation (4), i.e., SNMF. Begin with non-overlapping community detection, the resulting matrix $\mathbf{U}$ should be orthogonal in rows. We denote element in matrix $\mathbf{U}$ by $\mathbf{U}_{ic}$, where $i = 1, 2, \ldots, n$ represents node ID, $c = 1, 2, \ldots, d$ indicates community ID.

For the task of non-overlapping community detection, the non-negative matrix factorization shows:

- For any node i, the inner product of $\mathbf{U}_{i\cdot}$ and $\mathbf{U}_{i\cdot}$ always equals to 1, i.e., $\mathbf{U}_{i\cdot}\mathbf{U}_{i\cdot}^{\top} = 1$.
- For node i and node j that $i \neq j$, $\sum_{c=1}^{d} \mathbf{U}_{ic} = 1$ and $\sum_{c=1}^{d} \mathbf{U}_{jc} = 1$. If the two nodes are in the same community, the inner product of $\mathbf{U}_{i\cdot}$ and $\mathbf{U}_{j\cdot}$ equals to 1, i.e., $\mathbf{U}_{i\cdot}\mathbf{U}_{j\cdot}^{\top} = 1$ which also indicates an edge between node i and j. If the two nodes are in different communities, the corresponding inner product equals to 0, i.e., $\mathbf{U}_{i\cdot}\mathbf{U}_{j\cdot}^{\top} = 0$ which implies that there should not be an edge between them.

Now we define an indicating matrix $\mathbf{E}_{hk}$ of size $d \times d$ with only the h^{th} row, k^{th} column element is 1 and the others 0. Taking into account another product of row vector in matrix $\mathbf{U}$:

- For any node i that belongs to community h, the product $\mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{i\cdot} = \mathbf{E}_{hh}$.
- For node i and node j with $i \neq j$, the product $\mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{j\cdot} = \mathbf{E}_{hh}$, if the two nodes are in the same community h; the product $\mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{j\cdot} = \mathbf{0}$, if the two nodes are in different communities.

It follows that

$$
\begin{aligned}
\mathbf{U}^{\top}\mathbf{U} &= \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{21} & \cdots & \mathbf{U}_{n1} \\ \mathbf{U}_{12} & \mathbf{U}_{22} & \cdots & \mathbf{U}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{1d} & \mathbf{U}_{2d} & \cdots & \mathbf{U}_{nd} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} & \cdots & \mathbf{U}_{1d} \\ \mathbf{U}_{21} & \mathbf{U}_{22} & \cdots & \mathbf{U}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{n1} & \mathbf{U}_{n2} & \cdots & \mathbf{U}_{nd} \end{bmatrix} \\
&= \begin{bmatrix} \sum_i \mathbf{U}_{i1}\mathbf{U}_{i1} & \sum_i \mathbf{U}_{i1}\mathbf{U}_{i2} & \cdots & \sum_i \mathbf{U}_{i1}\mathbf{U}_{id} \\ \sum_i \mathbf{U}_{i2}\mathbf{U}_{i1} & \sum_i \mathbf{U}_{i2}\mathbf{U}_{i2} & \cdots & \sum_i \mathbf{U}_{i2}\mathbf{U}_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i \mathbf{U}_{id}\mathbf{U}_{i1} & \sum_i \mathbf{U}_{id}\mathbf{U}_{i2} & \cdots & \sum_i \mathbf{U}_{id}\mathbf{U}_{id} \end{bmatrix} \\
&= \sum_{i=1}^{n} \begin{bmatrix} \mathbf{U}_{i1}\mathbf{U}_{i1} & \mathbf{U}_{i1}\mathbf{U}_{i2} & \cdots & \mathbf{U}_{i1}\mathbf{U}_{id} \\ \mathbf{U}_{i2}\mathbf{U}_{i1} & \mathbf{U}_{i2}\mathbf{U}_{i2} & \cdots & \mathbf{U}_{i2}\mathbf{U}_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{id}\mathbf{U}_{i1} & \mathbf{U}_{id}\mathbf{U}_{i2} & \cdots & \mathbf{U}_{id}\mathbf{U}_{id} \end{bmatrix} \\
&= \sum_{i=1}^{n} \mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{i\cdot} = \mathbf{B}, \qquad (5)
\end{aligned}
$$

where the matrix $\mathbf{B}$ is a diagonal matrix and all the diagonal entries in it are positive numbers, in the schema of non-overlapping community.

For the task of overlapping community detection, the obtained matrix $\mathbf{U}$ through SNMF contains the membership coefficients for each node by row vectors. Slightly different from the aforementioned analysis, the facts demonstrate as:

- For any node i, the inner product $\mathbf{U}_{i\cdot}\mathbf{U}_{i\cdot}^{\top}$ should be a positive number, the product $\mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{i\cdot}$ is not necessarily a diagonal matrix.

The derivation shown in Equation (5) still holds in the scenario of overlapping community detection problem. However, the resulting matrix $\mathbf{B}$ has different styles. Due to the fuzzy community membership of each node assigned by a traditional SNMF algorithm, $\mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{i\cdot}$ might not be a non-negative diagonal dominant matrix, and thus the summation over i, i.e., $\sum_{i=1}^{n} \mathbf{U}_{i\cdot}^{\top}\mathbf{U}_{j\cdot}$ may not always be a non-negative diagonal dominant matrix.

Based on above depiction and in order to make the model to be precisely controllable, the proposed self-constrained symmetric non-negative matrix factorization is modeled as following cost function to optimize,

$$
\begin{aligned}
\mathcal{J}_{\text{SCSNMF}} &= \frac{1}{2}\|\mathbf{G} - \mathbf{U}\mathbf{U}^{\top}\|_{\text{F}}^2 + \frac{\lambda}{2}\|\mathbf{U}^{\top}\mathbf{U} - \alpha\mathbf{I}\|_{\text{F}}^2, \\
&\quad \mathbf{U} \geq \mathbf{0}, \qquad (6)
\end{aligned}
$$

where $\mathbf{I}$ is a $n \times n$ identity matrix whose diagonal entries are all ones and zeros for the others. The parameter λ controls the second regularization, and the parameter α controls the overlapping degree.

Soft membership assignments by any traditional NMF methods are unavoidable, since mature methods [19] [23] [13] to solve NMF problem always begin with a randomized matrix of $\mathbf{U}$. Thus, many NMF based community detection approaches reassign nodes with on-demand membership to hand out clear partitions.

In Equation (6), we use a parameter, i.e., α, to balance the overlapping degree of a partition. If we set α to a larger number, the model enforces $\mathbf{U}$ to conform the product $\mathbf{U}^{\top}\mathbf{U}$ to be diagonal dominant so that the resulting partition will be "more non-overlapping", and vice versa.

B. The Algorithm to Solve SCSNMF

The cost function $\mathcal{J}$ of **SCSNMF** in Equation (6) is not convex in $\mathbf{U}$. Thus, it is probable to find a local minima of $\mathcal{J}$ by using multiplicative updating rules proposed by Ding el al. in [24].

Taking into account the property of matrix trace, the Lagrangian function of Equation (6) can be rewritten as follows,

$$
\begin{aligned}
&\mathcal{J}_{\text{SCSNMF}} \\
&= \text{tr}(\mathbf{G}^{\top}\mathbf{G} - \mathbf{U}\mathbf{U}^{\top}\mathbf{G} - \mathbf{G}^{\top}\mathbf{U}\mathbf{U}^{\top} + \mathbf{U}\mathbf{U}^{\top}\mathbf{U}\mathbf{U}^{\top}) \\
&+\lambda\text{tr}(\mathbf{U}^{\top}\mathbf{U}\mathbf{U}^{\top}\mathbf{U} - 4\mathbf{U}^{\top}\mathbf{U} + \alpha^2\mathbf{I}) - \text{tr}(\Lambda\mathbf{U}), \qquad (7)
\end{aligned}
$$

Algorithm 1 commDetSCSCNMF(): Updating Procedure for Community Detection

Input:

graph network adjacent matrix $\mathbf{G}$, number of communities d, regularization parameter λ, overlapping degree parameter α

Output:

user-community membership indicator matrix $\mathbf{U}$

initialize elements of $\mathbf{U}$ with non-negative random numbers ranged in $[0, 1]$
while not convergent **do**
    update $\mathbf{U}$ according to Equation 12
end while
return $\mathbf{U}$

where Λ is the Lagrangian multipliers for non-negativity of $\mathbf{U}$. It follows that,

$$\begin{aligned}\frac{\partial \mathcal{J}_{\text{SCSNMF}}}{\partial \mathbf{U}} &= 4(\mathbf{U}\mathbf{U}^\top\mathbf{U} + \lambda\mathbf{U}\mathbf{U}^\top\mathbf{U}) \\ &- 4(\mathbf{G}^\top\mathbf{U} + \lambda\alpha\mathbf{U}) - \Lambda^\top. \end{aligned} \tag{8}$$

With the KKT complementary condition, which is,

$$\mathbf{U}(i,c)\Lambda(i,c) = 0, \ \forall i \in [1,n], c \in [1,d], \tag{9}$$

and let Equation (8) be 0, i.e., $\frac{\partial \mathcal{J}_{\text{SCSNMF}}}{\partial \mathbf{U}} = 0$, we have,

$$\Lambda = (\mathbf{U}\mathbf{U}^\top\mathbf{U} + \lambda\mathbf{U}\mathbf{U}^\top\mathbf{U}) - (\mathbf{G}^\top\mathbf{U} + \lambda\alpha\mathbf{U}). \tag{10}$$

Using the KKT complementary condition in Equation (9), we have,

$$\left[(\mathbf{U}\mathbf{U}^\top\mathbf{U} + \lambda\mathbf{U}\mathbf{U}^\top\mathbf{U}) - (\mathbf{G}^\top\mathbf{U} + \lambda\alpha\mathbf{U})\right](i,c)\mathbf{U}(i,c) = 0. \tag{11}$$

Thus we get the following updating rule for $\mathbf{U}$ that satisfies the above KKT condition:

$$\mathbf{U}_{ic} \leftarrow \mathbf{U}_{ic}\sqrt{\frac{[\mathbf{G}\mathbf{U} + \lambda\alpha\mathbf{U}]_{ic}}{[(1+\lambda)\mathbf{U}\mathbf{U}^\top\mathbf{U}]_{ic}}}. \tag{12}$$

The overall algorithm for self-constrained symmetric NMF community detection is listed in Algorithm (1).

IV. Experiments

In this section, the task of overlapping community detection is carried out on 5 real-world social networks for the performance evaluations of the proposed SCSNMF method.

A. Datasets

We use 5 classic real-world social networks for experiments: (1) Zachary's karate club (Karate) [25], (2) Dolphin social network (Dolphin) [26], (3) E-mail network URV (Email) [27], (4) Books about US politics (Polbooks) [1], and (5) College football teams (Football) [4]. The specifications of these datasets are listed in Table I.

[1]http://www-personal.umich.edu/~mejn/netdata/

TABLE I: Specifications of datasets.

Dataset	# of nodes	# of edges
Karate	34	78
Dolphin	62	159
Email	1133	5451
Polbooks	105	441
Football	115	613

B. Evaluation Metrics

The commonly used modularity Q [15] is one of the metrics we used in experiments. Modularity is a measure of network structure, which is also a evaluation of performance of community detection. For non-overlapping community detection of a undirected unweighted network, the modularity Q is defined as,

$$\mathrm{Q} = \frac{1}{2m}\sum_{ij}\left(\mathbf{G}_{ij} - \frac{k_i k_j}{2m}\right)\delta(i,j), \tag{13}$$

where m is the number of edges in graph $\mathbf{G}$, k_i is the degree of node i, and $\delta(i,j)$ indicates if node i and node j belong to the same community, i.e., $\delta(i,j) = 1$ if they belong to the same community, and 0 otherwise. As defined by Equation (13), a network with high modularity Q should have dense connections among nodes within communities, but sparse connections between nodes in different communities. However, for soft assigned membership coefficients in overlapping community detection, the community indicator $\delta(i,j)$ should be in the other form, and the modularity Q_{ov} for overlapping community can be defined as,

$$\mathrm{Q}_{\text{ov}} = \frac{1}{2m}\sum_{c}\sum_{ij}\left(\mathbf{G}_{ij} - \frac{k_i k_j}{2m}\right)P_{ic}P_{jc}, \tag{14}$$

where P_{ic} is the membership coefficient of node i in community c.

Another metric we use in experiments is the partition density D [7]. The partition density D is defined as follows,

$$\mathrm{D} = \frac{2}{m}\sum_{c} m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}, \tag{15}$$

where m_c and n_c are the number of edges and the number of nodes embraced in community c, respectively. An edge that connects node i and node j is in community c, if nodes i and j both are in community c.

C. Experiments Setup

Since the SNMF method is basis the proposed SCSNMF progresses on, we choose the community numbers identified by SNMF in the following experiments, which are 4, 6, 12, 4 and 10 for datasets Karate, Dolphin, Email, Polbooks and Football, respectively. In the experiments, the parameter λ varies in $[0, 15]$ with step size 1, and the parameter α rages in $[1, 15]$ with step size 1. Experiments on the effectiveness of parameter λ are carried out when α is fixed at 1, and vice versa. The maximum iteration number for SCSNMF is set as 1000. Nevertheless, the algorithm will converge early. Due to

the randomness of matrix factorization, experiments with the same parameter setting are carried out 100 times.

D. Experimental Results

The result of experiments on the effectiveness of parameters λ and α are shown in Figure 1 and Figure 2, respectively. From the results we can figure that the parameter λ controls the second regularization term in the Equation (6) of the SCSNMF model, and α controls the overlapping degree of identified communities. As λ and α increase, the non-overlapping modularity Q varies little, while the overlapping modularity Q_{ov} increase. This implies that with the enforcement of matrix $\mathbf{B}$ (c.f. Equation (5)) to be diagonal, i.e., enforcing matrix $\mathbf{U}$ to be orthogonal, the community partition becomes "less overlapping", resulting more densely connected nodes within community. Thus, the shrink of overlapping communities result in less nodes and edges in communities, and the vanishing of edges is much faster than that of nodes, which leads to a smaller partition density D.

V. Conclusion

In this paper we propose a self-constrained symmetric non-negative matrix factorization with tuning ability to control the overlapping degree of detected communities in a graph network for undirected networks. We use both traditional and overlapping version of modularity and partition density to investigate community overlapping on five real-world social network datasets. The experimental results show that the proposed method has the ability of interpretation for overlapping degree of communities.

For possible future work, an asymmetric NMF version should be considered to investigate the overlapping community in directed network.

Acknowledgment

This research is supported in part by the National Key Basic Research and Department (973) Program of China (No.2013CB329606), the National Natural Science Foundation of China (No. 71231002), and the Special Fund for Beijing Common Construction Project.

References

[1] H. Li, D. Wu, W. Tang, and N. Mamoulis, "Overlapping community regularization for rating prediction in social recommender systems," in *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 2015, pp. 27–34.

[2] Y. Liu, B. Wang, B. Wu, X. Zeng, J. Shi, and Y. Zhang, "Cogrec: A community-oriented group recommendation framework," in *International Conference of Young Computer Scientists, Engineers and Educators*. Springer, 2016, pp. 258–271.

[3] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," 2015.

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

[5] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *Acm computing surveys (csur)*, vol. 45, no. 4, p. 43, 2013.

[6] 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.

[7] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

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

[9] S. Gregory, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, p. 103018, 2010.

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

[11] D. Kuang, S. Yun, and H. Park, "Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering," *Journal of Global Optimization*, vol. 62, no. 3, pp. 545–574, 2014.

[12] J. Eustace, X. Wang, and Y. Cui, "Overlapping community detection using neighborhood ratio matrix," *Physica A: Statistical Mechanics and its Applications*, vol. 421, pp. 510–521, 2015.

[13] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using bayesian non-negative matrix factorization," *Physical Review E*, vol. 83, p. 066114, Jun 2011.

[14] 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, 2010.

[15] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[16] "Singular value decomposition," https://en.wikipedia.org/wiki/Singular_value_decomposition, accessed 29-April-2016.

[17] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems*, Conference Proceedings, pp. 1257–1264.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[20] Z.-Y. Zhang, Y. Wang, and Y.-Y. Ahn, "Overlapping community detection in complex networks using symmetric binary matrix factorization," *Physical Review E*, vol. 87, no. 6, p. 062803, 2013.

[21] Z. Wang, W. Wang, G. Xue, P. Jiao, and X. Li, "Semi-supervised community detection framework based on non-negative factorization using individual labels," *Advances in Swarm and Computational Intelligence: 6th International Conference, ICSI 2015 held in conjunction with the Second BRICS Congress, CCI 2015, Beijing, June 25-28, 2015, Proceedings, Part II*, pp. 349–359, 2015.

[22] X. Shi, H. Lu, Y. He, and S. He, "Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 541–546.

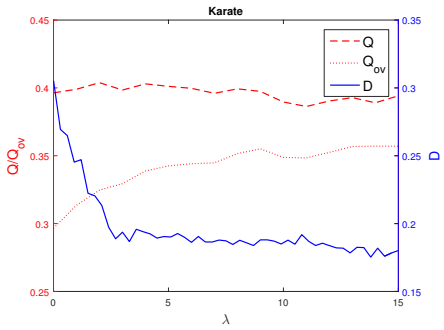
[23] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[24] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 126–135.

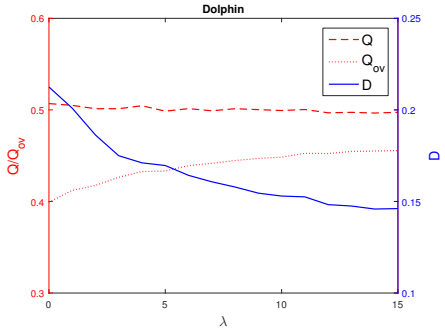
[25] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, pp. 452–473, 1977.

[26] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

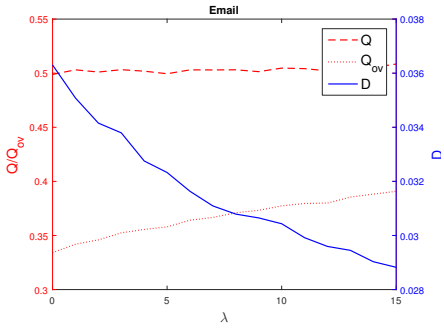
[27] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical review E*, vol. 68, no. 6, p. 065103, 2003.



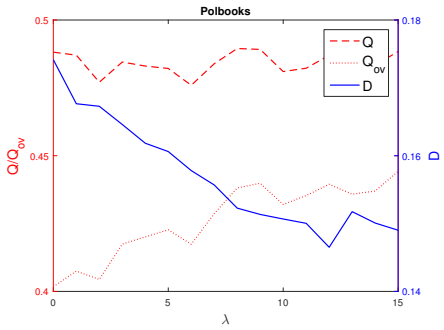
(a) Effectiveness of λ on Karate.



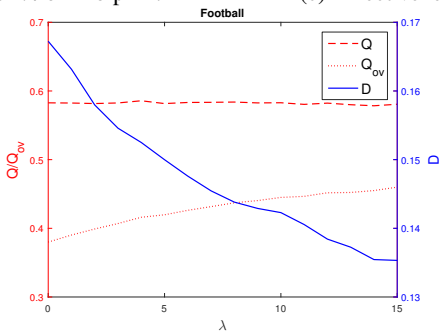
(b) Effectiveness of λ on Dolphin.



(c) Effectiveness of λ on Email.

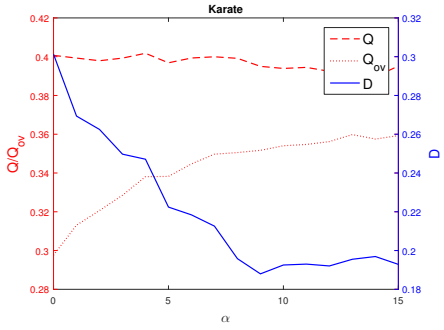


(d) Effectiveness of λ on Polbooks.

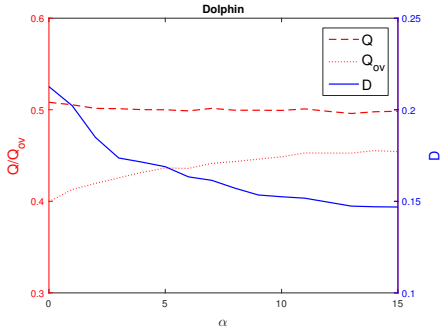


(e) Effectiveness of λ on Football.

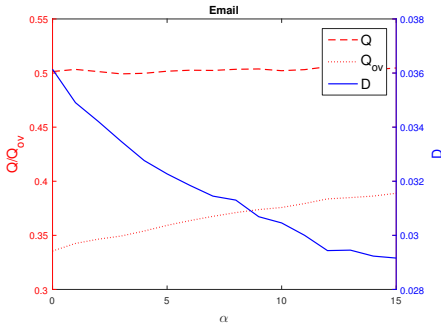
Fig. 1: Effectiveness of λ



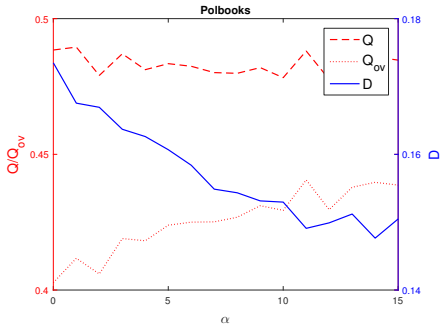
(a) Effectiveness of α on Karate.



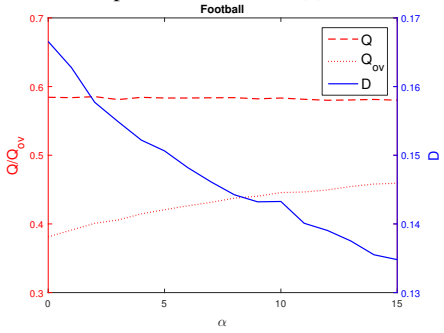
(b) Effectiveness of α on Dolphin.



(c) Effectiveness of α on Email.



(d) Effectiveness of α on Polbooks.



(e) Effectiveness of α on Football.

Fig. 2: Effectiveness of α