Statistica Sinica Preprint No: SS-2023-0061

Title	Orthogonal Symmetric Non-Negative Matrix Factorization Under the Stochastic Block Model
Manuscript ID	SS-2023-0061
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0061
Complete List of Authors	Subhadeep Paul and Yuguo Chen
Corresponding Authors	Yuguo Chen
E-mails	yuguo@illinois.edu

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

ORTHOGONAL SYMMETRIC NON-NEGATIVE MATRIX FACTORIZATION UNDER THE STOCHASTIC BLOCK MODEL

Subhadeep Paul and Yuguo Chen

The Ohio State University and University of Illinois at Urbana-Champaign

Abstract: We present a method based on the Orthogonal Symmetric Non-negative matrix Tri-Factorization (OSNTF) of the adjacency and the normalized Laplacian matrices for community detection in networks. We establish the connection of the factors obtained through this factorization to a non-negative basis of an invariant subspace of the approximating matrix, drawing parallel with the spectral clustering. Since the exact OSNTF may not exist or may not be computable for a given matrix like many non-negative matrix factorization methods, we study the approximate OSNTF that solves an optimization problem. We show that the global optimizer of the OSNTF objective function is consistent for community detection in networks generated from the stochastic block model as well as its degree corrected version. We compare the method with several state-of-the-art methods for community detection, including regularized spectral clustering, SCORE and SCOREplus, and spectral clustering followed by likelihood-based refinement, in both simulations and real datasets with known ground truth community assignments. These results show the excellent performance of the OSNTF under a wide variety of simulation setups and for real datasets obtained from disparate fields.

Key words and phrases: Community detection, degree corrected stochastic block

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

model, invariant subspace, network data, non-negative matrix factorization.

1. Introduction

Over the last two decades, there has been a surge in interest in the statistical inference of network data motivated by their applications in information sciences, biology, social sciences, and economics. A network consists of a set of entities called nodes or vertices and connections among them called edges or relations. The problem of community detection in networks has received considerable attention in the literature. A community is often defined as a group of nodes that are more "structurally similar" to each other than the rest of the network. Therefore nodes that belong to a community have similar patterns of connection to the rest of the network.

Several methods have been proposed in the literature for the efficient detection of network communities. These methods include modularity maximization (Newman and Girvan, 2004; Blondel et al., 2008), spectral clustering (Ng et al., 2002; McSherry, 2001; Rohe et al., 2011; Lei and Rinaldo, 2015; Qin and Rohe, 2013), semidefinite programming (Montanari and Sen, 2016; Hajek et al., 2016), and model based clustering using maximum likelihood (Choi et al., 2012; Zhao et al., 2012), variational EM (Daudin et al., 2008) and MCMC algorithms (Snijders, 2001; McDaid et al., 2013). Several

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

of these methods have been studied theoretically under the stochastic block model (SBM) and the degree corrected stochastic block model (DCSBM). In particular, spectral clustering and its variants including the regularized spectral clustering, the SCORE method have been studied extensively under both SBM and DCSBM (Rohe et al., 2011; Qin and Rohe, 2013; Jin, 2015; Lei and Rinaldo, 2015; Gao et al., 2017). A spectral clustering followed by likelihood based refinement scheme was shown to be mininax optimal under the SBM in Gao et al. (2017), and under the DCSBM in Gao et al. (2018).

In this paper, we consider methods for community detection in networks based on the non-negative matrix factorization of the adjacency and the Laplacian matrices of the network. Non-negative matrix factorization (NMF) has received strong attention in the machine learning and data mining literature since it was first introduced in Lee and Seung (1999). The method has many good properties in terms of performance and interpretability. It is quite popular in many applications, including image and signal processing, information retrieval, document clustering, neuroscience, and bioinformatics. A matrix X is said to be non-negative if all its elements are non-negative, i.e., $X_{ij} \geq 0$ for all i, j. The general NMF of order K decomposes a non-negative matrix $X \in \mathcal{R}_+^{N \times M}$ into two non-negative factor matrices $W \in \mathcal{R}_+^{N \times K}$ and $H \in \mathcal{R}_+^{K \times M}$, i.e., $X = WH$. When

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

$K \leq \min\{M, N\}$, NMF can also be looked upon as a dimension reduction technique that "decomposes a matrix into parts" that generate it (Lee and Seung, 1999).

However, an exact NMF of order K may not exist for any given non-negative matrix. Even if one does, finding the exact NMF in general settings is a computationally difficult problem (Vavasis, 2009). In fact it was shown that not just finding an exact order K NMF but also verifying the existence of the same is NP-hard. Several algorithms for an approximate solution have been proposed in the literature to remedy this (Lee and Seung, 2001; Lin, 2007; Cichocki et al., 2009). Popular optimization-based algorithms aim to minimize the difference between X and WH in the Frobenius norm under the non-negativity constraints. However, a natural question arises that given the matrix X is generated by exact multiplication of non-negative matrices (the "parts"), whether the decomposition can uniquely identify those parts of the generative model. A number of researchers have tackled this problem both geometrically and empirically (Donoho and Stodden, 2004; Hoyer, 2004; Laurberg et al., 2008; Huang et al., 2014).

NMF has also been applied in the context of clustering (Xu et al., 2003; Ding et al., 2005, 2006; Kim and Park, 2008). The "low-rank" NMF, where $K \leq \min\{M, N\}$, can be used to obtain a low-dimensional factor matrix,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

which can subsequently be used for clustering. Ding et al. (2005) showed interesting connections of NMF with other clustering algorithms, such as kernel k-means and spectral clustering. For applications in graph clustering where we generally have a symmetric adjacency matrix or a Laplacian matrix as the non-negative matrix, a symmetric version of the factorization was proposed in Wang et al. (2011). This factorization, called the symmetric non-negative matrix factorization (SNMF), has been empirically shown to yield good results in various clustering scenarios, including community detection in networks (Wang et al., 2011; Kuang et al., 2012). Arora et al. (2011) used a special case of SNMF, the left stochastic matrix factorization, for clustering and derived perturbation bounds. Yang et al. (2012) used a regularized version of the SNMF algorithm for clustering, while Psorakis et al. (2011) used a Bayesian NMF for overlapping community detection.

In this paper, we consider another non-negative matrix factorization designed to factorize symmetric matrices, the orthogonal symmetric non-negative matrix (tri) factorization (OSNTF) (Ding et al., 2006; Pompili et al., 2014). In contrast with earlier approaches, the requirement of being orthogonal in OSNTF adds another layer of extra constraints but generates sparse factors which are good for clustering. It also performs well in our simulation experiments. We prove that OSNTF is consistent under both

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

the SBM and its degree corrected variant. Through simulations and real data examples, we demonstrate the efficacy of OSNTF in community detection. In particular, our simulations show the proposed methods either outperform or are competitive with the spectral clustering and its modifications including regularization and projection, SCORE and SCOREplus, and likelihood based refinement under a variety of scenarios including high degree of heterogeneity and sparsity. We also explore various issues associated with the methods including interpretation of the estimated factors, numerical algorithms, and initialization.

Main contributions: We have two main contributions in this paper. First, we provide motivation and theoretical justification for using OSNTF for the problem of community detection in network data. We interpret the OSNTF factors as a nonnegative basis for an invariant subspace of the closest (in Frobenius norm) approximating matrix. We establish an upper bound on the mis-clustering rate using this method under the SBM and its degree corrected variant. Second, we demonstrate the consistently excellent performance of the proposed method in comparison to state-of-the-art methods for community detection in extensive simulation studies as well as real data experiments.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

2. Methods

We consider an undirected graph G on a set of N vertices. The adjacency matrix A associated with the graph is defined as a binary symmetric matrix with $A_{ij} = 1$, if node i and j are connected and $A_{ij} = 0$, if they are not. Throughout this paper we do not allow the graphs to have self loops. In this context we define degree of a node as the number of nodes it is connected to, i.e., $d_i = \sum_j A_{ij}$. The corresponding normalized graph Laplacian matrix can be defined as $L = D^{-1/2}AD^{-1/2}$, where D is a diagonal matrix with the degrees of the nodes as elements, i.e., $D_{ii} = d_i$. For a matrix H, the notation $H \geq 0$ means H is non-negative, i.e., all its elements are non-negative. We denote the Frobenius norm as $\| \cdot \|_F$ and the spectral norm as $\| \cdot \|_2$. We use $\| \cdot \|$ to denote the L^2 norm (Euclidean norm) of a vector.

We first describe SNMF which was previously used for community detection in networks using adjacency matrix by Wang et al. (2011) and Kuang et al. (2012). Given a symmetric positive semi-definite matrix A, the *exact* SNMF of order K for the matrix can be written as

$$A_{N\times N} = HH^T, \qquad H_{N\times K} \geq 0. \tag{2.1}$$

However since finding or even verifying the existence of the exact SNMF is NP-hard, an approximate solution is obtained instead by solving the following optimization problem, which seeks to minimize the distance in Frobenius norm between A and HH^T, i.e., we find, $\hat{H} = \arg\min_{H_{N\times K}\geq 0} \|A - HH^T\|_F$. Denoting $\hat{A} = \hat{H}\hat{H}^T$, it is easy to see that $\hat{H}$ is an exact SNMF factor of $\hat{A}$. We will refer to the solution of this optimization problem as SNMF. Clearly if A has an exact factorization as in Equation (2.1), that factorization will be the solution to this optimization problem and then SNMF will refer to that exact factorization. However since HH^T is necessarily positive semi-definite, the *exact* factorization in Equations (2.1) can not exist for matrices A or L that are not positive semi-definite. Moreover, being positive semi-definite is not a sufficient condition for the non-negative matrix A to have a decomposition of the form HH^T with $H \geq 0$. A non-negative positive semi-definite matrix that can be factorized into an SNMF is called a completely positive matrix (Berman, 2003; Gray and Wilson, 1980).

In an attempt to remedy this situation, we consider another symmetric non-negative matrix factorization where the matrix A is not required to be completely positive. Given a matrix A, this factorization, called the orthogonal symmetric non-negative matrix tri-factorization (OSNTF) of order K (Ding et al., 2006), can be written as

$$A_{N\times N} = HSH^T, \qquad H_{N\times K} \geq 0,\ S_{K\times K} \geq 0,\ H^T H = I. \tag{2.2}$$

The matrix S is symmetric but not necessarily diagonal and can have both positive and negative eigenvalues. Note that having the S matrix gives the added flexibility of factorizing matrices which are not positive semi-definite and hence has negative eigenvalues. In this connection it is worth mentioning that another symmetric tri-factorization was defined in Ding et al. (2005) without the orthogonality condition on the columns of H. However we keep this orthogonality condition as it leads to sparse factors and our experiments indicate that it leads to better performance for non-overlapping community detection both in simulations and in real networks.

As before, in practice it is difficult to obtain or verify the existence of the *exact* OSNTF in Equation (2.2) for any given adjacency matrix. Hence to obtain an approximate decomposition, we minimize the distance in Frobenius norm between A and HSH^T, i.e., we find

$$[\hat{H}, \hat{S}] = \underset{H_{N\times K}\geq 0,\, S_{K\times K}\geq 0,\, H^T H = I}{\arg\min} \|A - HSH^T\|_F. \tag{2.3}$$

The solution to this optimization problem will be referred to as OSNTF of A. If an exact OSNTF of A exists then this solution will coincide with the exact OSNTF.

Once we obtain $\hat{H}$, the community label for the ith node, z_i, is obtained by assigning the ith row of $\hat{H}$ to the column corresponding to its largest

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

element, i.e.,

$$z_i = \arg\max_{j \in \{1,\ldots,K\}} \hat{H}_{ij}. \tag{2.4}$$

Here the rows of $\hat{H}$ represent the nodes and the columns represent the communities. This way each node is assigned to one of the K communities.

Uniqueness While finding if an exact OSNTF of order K exists is NP-hard, it is worth investigating that given such a factorization exists, whether it is even possible to uniquely recover the parts or factors through non-negative matrix factorization. This issue has been investigated in detail in Donoho and Stodden (2004), Laurberg et al. (2008), Ding et al. (2006), and Huang et al. (2014). We describe two observations which together are sufficient for identification in our application. Let us denote the set $\mathcal{H}_+^{N \times K} = \{H \geq 0, H^T H = I\}$. We notice the following two propositions.

Proposition 1. *For any $H \in \mathcal{H}_+^{N \times K}$, each row of H contains at most one non-zero (positive) element.*

Proposition 2. *For any $N \times N$ symmetric matrix A, if $rank(A) = K \leq N$, then the order K exact OSNTF of A is unique up to a permutation matrix, provided the exact factorization exist.*

The proof of these two propositions along with those of all other lemmas and theorems are given in the Supplementary Material. Note while the second proposition is a result of the *exact* version of OSNTF, the global optimizer of the optimization problem for OSNTF leads to the same solution as the exact version when the matrix has an exact OSNTF.

Connections to invariant subspaces, projections, and spectral clustering We now connect OSNTF to invariant subspaces of a linear transformation on a finite dimensional vector space. Suppose $[\hat{H}, \hat{S}]$ is an OSNTF of order K of the matrix A. Then $\hat{A} = \hat{H}\hat{S}\hat{H}^T$ is an at most rank K matrix approximating A. By definition $\hat{A}$ has an exact OSNTF of order K. Focusing on the *exact* OSNTF, we note that the factorization in (2.2) of order $K \leq N$ can be equivalently written as

$$AH = HSH^T H = HS, \qquad H_{N\times K} \geq 0, S_{K\times K} \geq 0, H^T H = I. \tag{2.5}$$

This implies that the columns of H span a K-dimensional invariant subspace, $\mathcal{R}(H)$, of A. Further, Since H has K orthonormal columns, $rank(H) = K$. Consequently, the columns of H form an orthogonal basis for the subspace $\mathcal{R}(H)$. Every eigenvalue of S is an eigenvalue of A and the corresponding eigenvector is in $\mathcal{R}(H)$. To see this, note that if x is an eigenvector of S corresponding to the eigenvalue λ, then, $Sx = \lambda x$. Now, $AHx = HSx = \lambda Hx$. Hence λ is an eigenvalue of A and the corresponding eigenvector is Hx, which is in $\mathcal{R}(H)$. Finally, since in this case,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

$rank(A) = rank(S)$, S contains all the non-zero eigenvalues of A as its eigenvalues.

Note that the projection matrix onto the column space of H, i.e., $\mathcal{R}(H)$, is given by $P = HH^T$. From Equation (2.5) we have $AP = AHH^T = HSH^T = HH^THSH^T = PA$. Therefore, $\mathcal{R}(H)$ is also a reducing subspace of the column space of A (Radjavi and Rosenthal, 2003; Stewart and Sun, 1990). Hence the following decomposition holds (called the spectral resolution of A): $\begin{pmatrix} H_1^T \\ H_2^T \end{pmatrix} A(H_1\, H_2) = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$, where H_1 and H_2 are matrices whose columns span $\mathcal{R}(H)$ and its orthogonal complement respectively (Stewart and Sun, 1990).

Reverting back to the approximate factorization, we notice that the optimization problem in Equation (2.3) is to find the best projection of A into an at most rank K matrix $\hat{A}$ which has a non-negative invariant subspace. Note, here and subsequently, the "best" approximation implies a matrix which minimizes the distance in Frobenius norm. The difference of this projection with the projection in spectral clustering through singular value decomposition (McSherry, 2001; Von Luxburg, 2007) is that the projection in singular value decomposition ensures that the result is the best at most rank K matrix approximating A, however it does not necessarily have

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

an invariant subspace with a non-negative basis. In that sense the OSNTF projection adds an additional constraint on the projection and consequently the resultant matrix is no longer the best at most rank K approximating matrix. In OSNTF, the non-negative invariant subspace $\mathcal{R}(H)$ is used for community detection. Hence in general, the discriminating subspace in OSNTF is different from the one used in spectral clustering. We can make a similar observation for SNMF.

An equivalent objective function for OSNTF We characterize the optimization problem of OSNTF in (2.3) as an equivalent maximization problem. Given a feasible $H \in \mathcal{H}_+^{N\times K}$, the square of the objective function in the optimization problem in (2.3) can be written as

$$J = tr[(A - HSH^T)^T(A - HSH^T)] = tr(AA - 2SH^TAH + SS).$$

Solving for S (without nonnegativity constraint), we obtain $\hat{S} = H^TAH$. We note that this $\hat{S}$ is automatically non-negative since both H and A are non-negative matrices. Therefore, given an $H \in \mathcal{H}_+^{N\times K}$, the solution obtained for S is a feasible solution. Replacing S by $\hat{S}$ in the objective function J, we get the concentrated objective function as

$$\arg\min_{H\in\mathcal{H}_+^{N\times K}} tr(AA - H^TAHH^TAH) \equiv \arg\max_{H\in\mathcal{H}_+^{N\times K}} \|H^TAH\|_F^2. \qquad (2.6)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

We denote the (positive) square root of the concentrated objective function as $F(A, H) = \|H^T AH\|_F$.

The OSNTF procedure seeks to solve the optimization problem in Equation (2.6) to estimate an $H \in \mathcal{H}_+^{N\times K}$. This can be compared with the optimization viewpoint of spectral clustering, which seeks to optimize the same objective function as in (2.6), keeping the constrain $H^T H = I_K$ but removing the non-negativity constraint $H \geq 0$. Therefore, intuitively, OSNTF solves a more complex problem than the spectral clustering objective function in the first step. However, the advantage of the OSNTF method is in the second stage, where Equation (2.4) suggests that communities can be assigned simply by comparing the entries in each row of H and does not require a k-means algorithm to assign communities. In the next section, we show that OSNTF is consistent for community detection in graphs sampled from both the SBM and the DCSBM. As part of the proofs in the Supplementary Material, we show that OSNTF is able to correctly recover the community structure from the expected probability matrices (noiseless case) for both models. Since spectral clustering relies on k-means clustering in the second step, for DCSBM, either the eigenvectors need to be projected in a unit sphere or ratios of eigenvectors should be taken in order to correctly identify the community structure (Qin and Rohe, 2013; Jin, 2015).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

However, the same OSNTF algorithm works for both SBM and DCSBM graphs given the method is unaffected by degree heterogeneity. In the simulations, we observe that the OSNTF may have some advantages over the spectral methods (including those designed for degree heterogeneity) when the graph is sparse or has substantial degree heterogeneity.

3. Consistency of OSNTF for community detection

We now turn our attention to more general adjacency and Laplacian matrices. The SBM is a well studied statistical model of a network with community structure (Holland et al., 1983; Snijders, 2001; Rohe et al., 2011; Lei and Rinaldo, 2015; Choi et al., 2012). The K block SBM assigns to each node of a network, a K dimensional community label vector which takes the value of 1 in exactly one position and 0's everywhere else. Let Z be a matrix whose ith row is the community label vector for the ith node. Given the community labels of the nodes, the edges between them are formed independently following a Bernoulli distribution with a probability that depends only on the community assignments. We further assume that there is at least one non-zero element in each column, i.e., each community has at least one node. The SBM can be written in the matrix form as

$$E(A) = \mathcal{A} = ZBZ^T, \quad B \in [0,1]^{K\times K},\ Z \in \{0,1\}^{N\times K}, \tag{3.1}$$

where the matrix $B \geq 0$ is a $K \times K$ symmetric matrix of probabilities. We assume the matrix B is of full rank, i.e., of rank K. We will refer to the matrix $\mathcal{A}$ as the population adjacency matrix. The population Laplacian matrix is defined from this adjacency matrix as $\mathcal{L} = \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}$, where $\mathcal{D}$ is a diagonal matrix with the elements being $\mathcal{D}_{ii} = \sum_j \mathcal{A}_{ij}$. The matrix $\mathcal{L}$ under the K class SBM defined above can be written as $\mathcal{L} = ZD_B^{-1/2}BD_B^{-1/2}Z^T = ZB_LZ^T$, where $D_B = diag(BZ^T\mathbf{1}_N) \in R^{K\times K}$ with $\mathbf{1}_N$ being the vector of all ones in $\mathcal{R}^N$, is a diagonal matrix and $B_L = D_B^{-1/2}BD_B^{-1/2}$ (Rohe et al., 2011).

Although the SBM is a well-studied model, it is not very flexible in terms of modeling real world networks. Many real world networks exhibit heterogeneity in the degrees of the nodes which the SBM fails to model. To remedy this, an extension of SBM for general degree distributions was proposed in Karrer and Newman (2011), called the DCSBM. In our matrix terms the model can be written as

$$\mathcal{A} = \Theta ZB'Z^T\Theta, \quad B' \in \mathcal{R}_+^{K\times K},\ Z \in \{0,1\}^{N\times K},\ \Theta \in \mathcal{R}_+^{N\times N}, \qquad (3.2)$$

where B' is a symmetric full rank matrix and Θ is an $N \times N$ diagonal matrix containing the degree parameters θ_i for the nodes as elements. Following Karrer and Newman (2011) we impose identifiability constraints $\sum_{\{i:Z_{iq}=1\}} \theta_i = 1$ for each $q \in \{1, \ldots, K\}$. The interpretation of B' is that

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

each entry B'_{ql} represents the expected number of edges between communities q and l.

We define regularized versions of the adjacency matrix A_τ and the Laplacian matrix L_τ as follows. Let $\Delta = n \max_{i,j} \mathcal{A}_{ij}$. We reduce the weights of the edges incident on vertices whose degrees are larger than 2Δ such that all degrees of the new weighted graph is bounded by 2Δ (Le et al., 2017). For the regularized Laplacian matrix, we first define $A_\tau = A + \frac{\Delta}{n}\mathbf{1}\mathbf{1}^T$. Then define the Laplacian as $L_\tau = D_\tau^{-1/2} A_\tau D_\tau^{-1/2}$, where D_τ is the diagonal matrix of degrees of A_τ such that $(D_\tau)_{ii} = \sum_j (A_\tau)_{ij}$. We derive nonasymptotic upper bounds on the error rates of community detection by applying OSNTF to the regularized adjacency matrix A_τ and the regularized normalized Laplacian matrix L_τ as defined above. The results for the non-regularized matrices follow similarly, albeit, with weaker bounds.

Lemmas 1 and 2 in the Supplementary Materials show that (i) the population adjacency matrix $\mathcal{A}$ and (regularized) Laplacian matrix $\mathcal{L}_\tau$ have exact OSNTFs and (ii) the community assignment vector z_i can be recovered from these OSNTF factor matrices $\bar{H}$ and $\bar{H}_L$ (explicitly defined in the lemmas) under the models SBM and DCSBM respectively.

However, the sample regularized adjacency matrix A_τ and sample Laplacian matrix L_τ may not have exact OSNTFs. In that case, let the opti-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

mization problem in (2.3) or equivalently in (2.6), obtain a solution $[\hat{H}, \hat{S}]$ as OSNTF of A_τ. The matrix approximating A_τ is then $\hat{A} = \hat{H}\hat{S}\hat{H}^T$ and we assign the nodes to the communities using the matrix $\hat{H}$. We denote the objective function in the optimization problem of (2.6) as $F(A_\tau, H) = \|H^T A_\tau H\|_F$. This is a function of the regularized adjacency matrix A_τ and the factor matrix H. We can define a corresponding "population" version of this objective function with the population adjacency matrix as $F(\mathcal{A}, H) = \|H^T \mathcal{A} H\|_F$. The next lemma, which is an intermediate result, shows that for any $H \in \mathcal{H}_+^{N\times K}$, the difference between $F(A_\tau, H)$ and $F(\mathcal{A}, H)$ is bounded in high probability.

Lemma 1. *For any $H \in \mathcal{H}_+^{N\times K}$, (i) there exists a constant $c_1(r_1) > 0$, such that with probability at least $1-n^{-r_1}$, we have $|F(A_\tau, H) - F(\mathcal{A}, H)| \leq c_1 K \Delta^{3/2}$ for some $r_1 > 0$, and (ii) there exists a constant $c_2 > 0$, such that with probability at least $1 - o(1)$, we have $|F(L_\tau, H) - F(\mathcal{L}_\tau, H)| \leq \frac{c_2 K}{\sqrt{\Delta}}$.*

We quantify the error in community detection through a measure called mis-clustering rate which, given the true community assignment and a candidate community assignment, computes the proportion of nodes for which the assignments do not agree. Let $\bar{z}$ denote the true assignment and $\hat{z}$ denote a candidate assignment. Then we define the mis-clustering rate $r = \frac{1}{N} \inf_\Pi d_H(\bar{z}, \Pi(\hat{z}))$, where Π is a permutation of the labels and $d_H(\cdot,\cdot)$

is the Hamming distance between two vectors.

Lemmas 4 and 5 in the Supplementary Material relate this mis-clustering rate with the difference between $\hat{H}$ and $\bar{H}P$ for some arbitrary permutation matrix P in Frobenius norm. The next two theorems obtain upper bounds on the mis-clustering rate of OSNTF under SBM and DCSBM. Define r_A and r_L as the mis-clustering rate for community detection through OSNTF of A_τ and L_τ respectively.

Theorem 1. *Let G be a graph generated from an N-node K-community SBM with parameters (Z, B) as in Equation (3.1). Define $\lambda^{\mathcal{A}}$ and $\lambda^{\mathcal{L}_\Delta}$ as the smallest non-zero (in absolute value) eigenvalues of the population adjacency matrix $\mathcal{A}$ and normalized Laplacian matrix with regularization parameter $\tau = \Delta$, i.e., $\mathcal{L}_\Delta$, respectively. Further, define $N_{max} = \max_{k\in\{1,\dots,K\}}(Z^TZ)_{kk}$, i.e., the population of the largest block. Then (i) with probability at least $1-n^{-r_1}$, we have $r_A \le \frac{N_{\max}c_1K\Delta^{3/2}}{N(\lambda^{\mathcal{A}})^2}$, and (ii) with probability at least $1-o(1)$, we have $r_L \le \frac{c_2N_{\max}K}{N(\lambda^{\mathcal{L}_\Delta})^2\sqrt{\Delta}}$, for some constants $r_1, c_1, c_2 > 0$.*

Theorem 2. *Let G be a graph generated from an N-node K-community DCSBM with parameters (Θ, Z, B) as in Equation (3.2). Define $\lambda^{\mathcal{A}}$ and $\lambda^{\mathcal{L}_\Delta}$ as the smallest non-zero (in absolute value) eigenvalues of the population adjacency matrix $\mathcal{A}$ and normalized Laplacian matrix with regularization parameter $\tau = \Delta$, i.e., $\mathcal{L}_\Delta$, respectively. Further, define $m =$*

$\min_{i\in\{1,\ldots,N\}} \theta_i/\sqrt{(Z^T\Theta^2 Z)_{kk}}$ *with* k *being the community to which node* i *truly belongs. Then (i) with probability at least* $1 - n^{-r_1}$*, we have* $r_A \leq \frac{4c_1 K\Delta^{3/2}}{Nm^2(\lambda^{\mathcal{A}})^2}$*, and (ii) with probability at least* $1-o(1)$*, we have* $r_L \leq \frac{4c_2 K}{Nm^2(\lambda^{\mathcal{L}}\Delta)^2\sqrt{\Delta}}$*, for some constants* $r_1, c_1, c_2 > 0$.

We end this section with several theoretical remarks on the results.

Remark 1 (Simplified setting SBM)**.** We apply Theorem 1 to a simplified special case of the SBM (Rohe et al., 2011; Qin and Rohe, 2013). Let all the probabilities of connection within blocks be p for all blocks, and the probability of connection between nodes from different blocks be q for all block pairs. The number of nodes within each block is $\frac{N}{K}$ (hence all blocks are of the same size), and K is the number of blocks. Then we have $N_{\max} = N/K$. Further let $p = a\frac{\Delta}{N}$ and $q = b\frac{\Delta}{N}$ with a and b being constants. Consequently, $p \asymp q$. Then we can write the matrix B as

$$B = (p-q)\mathbf{I}_K + q\mathbf{1}_K\mathbf{1}_K^T,$$

where $\mathbf{I}_K$ is the K-dimensional identity matrix and $\mathbf{1}_K$ is the K-dimensional vector of all 1s. Clearly the non-zero eigenvalues of $\mathcal{A}$ correspond to the eigenvalues of $(Z^TZ)^{1/2}B(Z^TZ)^{1/2} = \frac{N}{K}B$. Using this we obtain $\lambda^{\mathcal{A}} = \frac{N}{K}(p-q) = \frac{\Delta}{K}(a-b)$. Then from the result of Theorem 1, we have

$$r_A \lesssim \frac{K^2}{\sqrt{\Delta}(a-b)^2}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

This result indicates that the mis-clustering rate increases with increasing number of communities, decreases with increasing average degree of nodes (which can happen either by increasing the density of the network or the number of nodes in the network), and decreases with increasing separation between intra and inter community probabilities of connection. Further, the mis-clustering rate goes to 0 as long as $K = o(\Delta^{1/4})$.

We can obtain a similar result for the OSNTF with regularized normalized Laplacian matrix as well. We first define the following matrix:

$$B_\Delta = (p-q)\mathbf{I}_K + \Big(q + \frac{\Delta}{N}\Big)\mathbf{1}_K\mathbf{1}_K^T.$$

Then we can write the normalized Laplacian matrix with regularization parameter Δ as $\mathcal{L}_\Delta = ZB_{\Delta,L}Z^T$, where

$$B_{\Delta,L} = \frac{1}{N(q+\frac{\Delta}{N}) + \frac{N}{K}(p-q)}\left((p-q)\mathbf{I}_K + \Big(q + \frac{\Delta}{N}\Big)\mathbf{1}_K\mathbf{1}_K^T\right).$$

The non-zero eigenvalues of $\mathcal{L}_\Delta$ are same as the eigenvalues of $B_{\Delta,L}$, we compute the smallest eigenvalue of $\frac{N}{K}B_{\Delta,L}$. This implies

$$\lambda^{\mathcal{L}_\Delta} = \frac{1}{1 + K\frac{q+\Delta/N}{(p-q)}} = \frac{1}{1+K\frac{b+1}{a-b}} \asymp \frac{(a-b)}{(b+1)K}.$$

Then using the result in Theorem 1, we obtain

$$r_L \lesssim \frac{K^2}{\sqrt{\Delta}(a-b)^2}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

Therefore the asymptotic upper bound for the mis-clustering rate is the same using the (regularized versions of) adjacency matrix or the normalized Laplacian matrix. This is not surprising since several authors have observed that for spectral clustering, one cannot differentiate between using adjacency or the normalized Laplacian matrix by upper bounds on error rates (Sarkar and Bickel, 2015; Tang and Priebe, 2018).

Remark 2 (Simplified setting DCSBM)**.** We next analyze the asymptotic upper bounds in Theorem 2 similarly in terms of simplified settings of DCSBM that will elucidate the role of degree heterogeneity as well. As in the previous remark, we let the number of nodes within each block be $\frac{N}{K}$. Note the $K \times K$ matrix B' has the interpretation that its (q, l)th element B'_{ql} is the expected number of edges from community q to community l. Then we assume

$$B' = ((p-q)\mathbf{I}_K + q\mathbf{1}_K\mathbf{1}_K^T)\frac{N^2}{K^2}.$$

Suppose $\bar{\Delta}$ is a quantity (different from Δ defined in the theorem) such that $p = a\frac{\bar{\Delta}}{N}$ and $q = b\frac{\bar{\Delta}}{N}$ with a and b being constants $(a > b)$. Note that the entries of the matrix B'_{ql} are $\frac{N^2 p}{K^2}$ on the diagonal and $\frac{N^2 q}{K^2}$ off the diagonal. Therefore, the total number of links in the network is $c\bar{\Delta}N$ for some constant $c > 0$. This gives $\bar{\Delta}$ the interpretation of being the average degree of the network. Then notice, $\max_{q,l} B'_{ql} = \frac{N^2 p}{K^2}$. Now we compute

Statistica Sinica

from Theorem 2:

$$\Delta = N \max_{i,j} P_{ij} = N \max_{i,j} \theta_i \theta_j B'_{Z_i,Z_j} \le N\theta^2_{\max} \max_{q,l} B'_{ql} = N\left(\frac{N}{K}\right)^2 p\theta^2_{\max}.$$

On the other hand, for computing m, we make the assumption that for any q, we have $(Z^T\Theta^2 Z)_{qq} = \sum_{i:Z_{iq}=1} \theta_i^2 = \frac{K}{N}$. This implies that the groups themselves are not heterogeneous and the heterogeneity is only within a group. This can be thought of as a similar assumption to the constraint $\sum_{i:Z_{iq}=1} \theta_i = 1$. Then we compute

$$m = \min \frac{\theta_i}{\sqrt{(Z^T\Theta^2 Z)_{qq}}} = \frac{\theta_{\min}}{\sqrt{\frac{K}{N}}}.$$

Next we need to compute $\lambda^{\mathcal{A}}$. For this we note that the non-zero eigenvalues of $\mathcal{A}$ are the same as the eigenvalues of

$$(Z^T\Theta^2 Z)^{1/2} B' (Z^T\Theta^2 Z)^{1/2} = \frac{K}{N} B' = \frac{N}{K}((p-q)\mathbf{I}_K + q\mathbf{1}_K\mathbf{1}_K^T).$$

Therefore, we obtain $\lambda^{\mathcal{A}} = \frac{N}{K}(p-q) = \frac{\bar{\Delta}}{K}(a-b)$. Then combining these results, we have

$$r_A \lesssim \left(\frac{K}{N}\right)^2 \frac{N^{3/2}(\frac{N}{K})^3(\frac{\bar{\Delta}}{N})^{3/2}\theta^3_{\max}}{\theta^2_{\min}(\frac{\bar{\Delta}}{K})^2(a-b)^2} = \frac{NK}{\sqrt{\bar{\Delta}}(a-b)^2}\frac{\theta^3_{\max}}{\theta^2_{\min}}.$$

Therefore, the asymptotic upper bound in this case depends on the extent of degree heterogeneity through a function of $\theta_{\min}$ and $\theta_{\max}$. This behavior is similar to that observed for the spectral algorithm SCORE in (Jin et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

2022). It is easy to see that if all θ_is were equal, the DCSBM boils down to the SBM and then by the constraint $\theta_{\max} = \theta_{\min} = \frac{K}{N}$. Then this asymptotic upper bound becomes the same as what we found for the case of SBM.

Further consider a scenario, where all θ_is are $\frac{K}{N}$, except for a finite number of them which are all $\frac{K}{f(N)N}$, where $f(N) > 1$ is a function of N. Then $\theta_{\max} = \frac{K}{N}$ and $\theta_{\min} = \frac{K}{f(N)N}$. Then the upper bound becomes $r_A \lesssim \frac{K^2 f(N)^2}{\sqrt{\bar{\Delta}}(a-b)^2}$. Therefore, consistent community detection will still be possible as long as $f(N) = o(\frac{\bar{\Delta}^{1/4}}{K})$. Hence we could have $\theta_{\min}$ an order of magnitude smaller than $\theta_{\max}$, and yet, consistent community detection will be possible with OSNTF for appropriate growth rates on K and $\bar{\Delta}$.

Remark 3. Note a seemingly simpler technique can be used to obtain a bound if the adjacency matrix concentrates in the Frobenius norm (Arora et al., 2011). Note that from Lemmas 1 and 2 in the Supplementary Material, $\mathcal{A} = \bar{H}\bar{S}\bar{H}^T$ for both SBM and DCSBM. Now let $\hat{H}, \hat{S}$ be the solution of the OSNTF problem applied to the matrix A_τ. Then define $A_1 = \hat{H}\hat{S}\hat{H}^T$. Clearly both $\mathcal{A}$ and A_τ are matrices that have exact OSNTFs. Then we note the following relationship:

$$\|A_1 - \mathcal{A}\|_F \leq \|A_\tau - A_1\|_F + \|A_\tau - \mathcal{A}\|_F \leq 2\|A_\tau - \mathcal{A}\|_F,$$

since $\|A_\tau - A_1\|_F \leq \|A_\tau - \mathcal{A}\|_F$ as A_1 is the closest matrix in Frobenius norm to A_τ that has a rank K ONSTF. Now $\hat{H}$ and $\bar{H}$ are invariant subspaces of A_1 and $\mathcal{A}$ respectively. Therefore a perturbation theorem for invariant subspaces will provide a bound provided $\|A_\tau - \mathcal{A}\|_F$ can be bounded. However, the adjacency matrix is known to not concentrate well in Frobenius norm (Rohe et al., 2011) and the crude bound of $\sqrt{n}\|A_\tau - \mathcal{A}\|_2$ is too loose for our purpose.

4. Algorithm for OSNTF: convergence and implementation

In this section we discuss algorithms and implementation details for the OSNTF method. There are several algorithms proposed in the literature to solve the OSNTF optimization problem in Equation (2.3). The algorithm given by Ding et al. (2006) is a multiplicative update rule (MUR) which alternates with the following update rules:

$$S_{ik} \leftarrow S_{ik}\sqrt{\left(\frac{(H^T A H)_{ik}}{(H^T H S H^T H)_{ik}}\right)}, \tag{4.1}$$

$$H_{ik} \leftarrow H_{ik}\sqrt{\left(\frac{(AHS)_{ik}}{(HH^T AHS)_{ik}}\right)}. \tag{4.2}$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, K$. The matrix H is used for community detection in OSNTF. The algorithm needs a starting solution H_0, S_0. In this paper we call this procedure applied to the Laplacian matrix as the OSNTF method and applied to the regularized Laplacian matrix as the regularized OSNTF method.

Statistica Sinica

```
Algorithm 1: OSNTF- Convergent
    Input: $A, K$, tuning parameter $\alpha$
    Result: Clustering solution $z \in \{1, \dots, K\}^N$.
  Compute $L$ or $L_\tau$. Set $\delta = 10^{-10}$, and $\sigma = 10^{-6}$
  Initialize $H \geq 0$, $S \geq 0$ randomly, or with regularized Spectral
  $\nabla_S J(H, S) = H^T HSH^T H - H^T AH$
        $\tilde{S}_{ij} = \max(S_{ij}, \sigma)$ if $\nabla_S J(H, S) < 0,$ and $\tilde{S}_{ij} = S_{ij}$ otherwise
        $S_{ij}^{(t+1)} = S_{ij}^t - \tilde{S}_{ij}^t \frac{(\nabla_S J(H, S^t))_{ij}}{(H^T H\tilde{S}^t H^T H)_{ij} + \delta}$
  $\nabla_H J(H, S) = HSH^T HS + \alpha HH^T H - AHS - \alpha H$
        $\tilde{H}_{ij} = \max(H_{ij}, \sigma)$ if $\nabla_H J(H, S) < 0,$ and $\tilde{H}_{ij} = H_{ij}$ otherwise
        $H_{ij}^{(t+1)} = H_{ij}^t - \tilde{H}_{ij}^t \frac{(\nabla_H J(H^t, S))_{ij}}{\tilde{H}S\tilde{H}^T \tilde{H}S + \alpha\tilde{H}\tilde{H}^T \tilde{H})_{ij} + \delta}$
  Assign community: $\hat{z}_i = \arg\max_{j \in \{1, \dots, K\}} H_{ij}$.
```

While Ding et al. (2006) showed that the multiplicative algorithm converges to a stationary point, the guarantee applies to the algorithm which contains an unknown Lagrange multiplier. The authors obtained the multiplier parameter with approximations. The algorithm also suffers from a zero-locking problem. It is easy to note that if any entry of the parameter matrices H, S is zero at any point during the updates, then the algorithm will stop updating that entry irrespective of a stationary point is reached

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

or not (Mirzal, 2014).

The zero-locking problem can be avoided by providing a starting solution that is strictly positive. It can be shown that under the assumptions that the Laplacian matrix does not contain any zero rows or columns (no isolated nodes) and the starting solution is strictly positive, all subsequent updates will keep the solution strictly positive. However, the algorithm will not generally converge to a stationary point in that case since a stationary point can lie on the boundary of the feasible space. In order to provide a convergent algorithm, Mirzal (2014) developed an algorithm which provably converges to a stationary point. An algorithm based on gradients in the Stiefel manifold was developed in Choi (2008). However, the author did not investigate convergence properties of the algorithm. The convergence of algorithms for orthogonal tri-factorization is even less studied. The only paper we found that has studied the problem is Mirzal (2017) where a convergent algorithm was developed following the ideas in Mirzal (2014).

Using the approach of Mirzal (2014, 2017), we develop a convergent algorithm for the OSNTF problem described in Algorithm 1. The method requires a tuning parameter in the input which balances the relative importance of orthogonality constraint. The algorithm is an additive update rule (AUR) method that alternates between updating H and S with modifications in each stage to avoid the zero locking problem. We call the convergent algorithms in Algorithm 1 applied to L and L_τ as OSNTF Convergent and Regularized OSNTF Convergent respectively.

For both the OSNTF algorithms and the OSNTF-Convergent algorithms, we obtain the starting solution H_0 as the matrix whose rows contain $0.01/(K-1)$ in each position, except for the assigned community according to a community assignment methods where it contains 0.99. For the community assignment method of the initial solution, we use regularized spectral clustering with spherical k-means applied to the Laplacian matrix (Qin and Rohe, 2013) to obtain a community assignment. The matrix S is initialized as $S_0 = 0.08\mathbf{I}_K + 0.02\mathbf{1}_K\mathbf{1}_K^T$. In our simulation study, we found that OSNTF is reasonably agnostic to the starting solution if we run OSNTF multiple times with multiple starting values and choose the solution that minimizes the OSNTF objective function over these different runs.

5. Simulation Results

In this section we generate networks from both the SBM and the DCSBM and evaluate the performance of OSNTF approaches along with a few other methods applied to the normalized Laplacian matrix of the networks. In this section and the next section on real data analysis, we consider four OSNTF algorithms, namely, OSNTF (algorithm by Ding et al. (2006)), Regularized OSNTF (Reg. OSNTF), OSNTF-convergent (OSNTF-conv, Algorithm 1), Regularized OSNTF-convergent (Reg. OSNTF-conv). The methods we compare the OSNTF procedures against are the spectral clustering (Spectral) (Rohe et al., 2011; Lei and Rinaldo, 2015), the regularized spectral clustering (Reg. Spectral) (Qin and Rohe, 2013), the SCORE method (Jin, 2015), the SCOREplus method (Jin et al., 2022), the spectral clustering followed by likelihood refinement scheme (SBM refine) in Gao et al. (2017), and the algorithm in Gao et al. (2018) which we call DCBM refine. The prefix "regularized" before a method implies that the method is applied to the regularized Laplacian matrix. For brevity, in the simulations we only present results from OSNTF and Reg. OSNTF while omitting the results from OSNTF-conv and Reg. OSNTF-conv. For the real data analysis, we present results from all four OSNTF algorithms. We also omit results from the SCORE procedure since the procedure fails to execute for some graphs with high degree heterogeneity and low density. As discussed in this paper, the OSNTF method conceptually is similar to spectral clustering, and therefore it is natural to compare it with spectral methods. However, the algorithms used for computing the solution remind the reader of the likelihood refinement-based schemes starting from a suitable initial solution in

Statistica Sinica

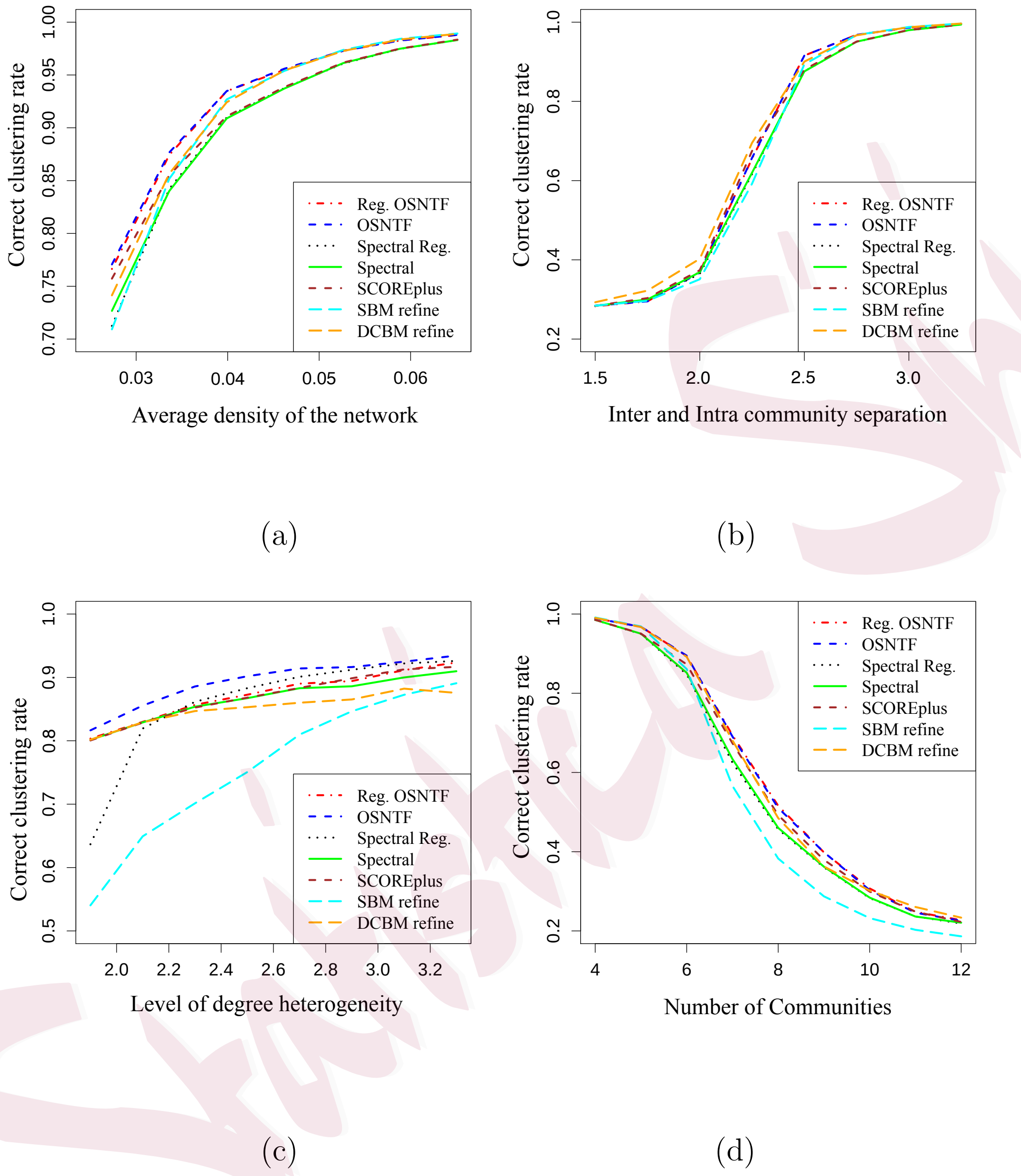
Gao et al. (2017, 2018). Therefore we compare the performance of OSNTF against those two methods as well.

We conduct four experiments, generating data from the SBM for the first three and from the DCSBM for the last one. The clustering quality of a partition is evaluated by measuring its agreement with the known community structure, i.e., the fraction of nodes correctly classified (correct classification rate). Clearly the correct classification rate takes a value between 0 and 1, with higher values indicating better agreement between the partitions. Further note that a random assignment is expected to have a correct classification rate of $1/K$, where K is the number of communities. All results are averaged over 100 simulations.

SBM: increasing density of the network: We generate data from the SBM with 4 clusters and 600 nodes. For the SBM, the signal to noise ratio, defined as the diagonal to off-diagonal elements' ratio, is fixed at around 3. We increase the network's average density (defined as the fraction of the pairs of nodes that are connected by an edge) from 0.025 to 0.08. This simulation is designed to test the robustness of the methods for sparse graphs where node degrees are relatively low. The results are presented in Figure 1(a). We notice that OSNTF and Reg. OSNTF are the two best performing methods for sparser graphs. The usual spectral clustering with-

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061



(a) (b)

(c) (d)

Figure 1: Comparison of the performance of various methods for three simulation settings: (a) SBM with $K = 4, N = 600$ and increasing average density, (b) SBM with $K = 4, N = 600$ and increasing separation between inter and intra community probabilities, (c) DCSBM with $K = 4, N = 600$ and decreasing degree heterogeneity, and (d) SBM with increasing K and $N = 500$.

Statistica Sinica

out any regularization and SBM refine method underperform among the methods compared in sparse graphs. The SCORE method (figure omitted) is not able to improve upon the spectral clustering's performance, while regularized spectral clustering slightly improves it. The SCOREplus and DCBM refine methods perform better than regularized spectral clustering, but underperform OSNTF and regularized OSNTF.

SBM: Increasing number of communities We increase the number of communities K from 4 to 12 while holding $N = 500$. As expected, the performance of all community detection methods deteriorates with increasing K (see Figure 1(d)). Among the methods, we observe that SCORE (figure omitted), DCBM refine, and the two OSNTF methods perform well.

SBM: Increasing difference between intra and inter community parameters: As predicted by the theoretical result, the performance of the OSNTF methods improves with increasing separation between the inter and intra community parameters. The theoretical results for the other methods also indicate a similar phenomenon. Our simulation confirms this observation and we notice that the performance of all methods improves similarly with increasing separation (see Figure 1(b)).

DCSBM: varying degree heterogeneity parameter: In our last experiment, we generate data from a DCSBM with 4 communities and 600

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

nodes. The degree parameter is generated from a power law distribution with lower bound parameter $x_{min} = 1$ and shape parameter β. We increase the shape parameter from 1.9 to 3.3 in steps of 0.2. A smaller β leads to greater degree heterogeneity, and hence increasing the parameter gradually makes the DCSBM more similar to an SBM. We again keep the signal to noise ratio at 3 and the average density of the networks generated is around 0.05. The results are presented in Figure 1(c). Here we see that the (unregularized) spectral clustering and the SBM refine completely break down in the presence of degree heterogeneity and recover slowly as the parameter β increases. The two OSNTF methods, regularized spectral clustering, DCBM refine and SCOREplus are robust against degree heterogeneity with the OSNTF method consistently outperforming all other methods. We dropped SCORE from this simulation since the method gives an error for the case of lowest β. This simulation study indicates that ONSTF performs well under severe degree heterogeneity.

5.1 Advantages of OSNTF and computational cost

From the above simulation results, we see that the OSNTF method generally has some advantages in terms of better performance than other existing methods in the literature when the network is sparse or has a severe degree

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

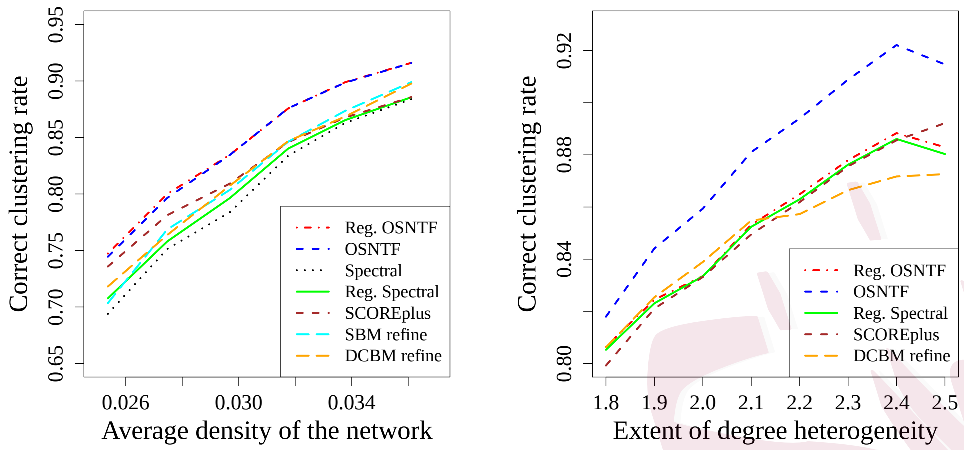
of heterogeneity. In most cases, the closest a method comes to OSNTF is SCOREplus, which was designed keeping in mind sparse and high degree-heterogeneity cases (Jin et al., 2022). Therefore, to elucidate the advantages of OSNTF specifically, we explore these cases more and conduct another simulation. In Figure 2(a) we focus on very sparse graphs with low average density of the network, and in Figure 2(b) we focus on graphs with higher degree heterogeneity. These two figures confirm that in some of those cases, OSNTF has substantially better performance compared to the other existing methods.

We further confirm this by checking the number of cases different algorithms returned the best correct clustering rate out of the 90 repetitions in each scenario. For the simulation with low average density graphs, we note from Table S1 in the Supplementary Material that either OSNTF and Reg. OSNTF performed the best most of the times for the cases investigated. For the simulation with high degree heterogeneity, we notice that OSNTF performed the best in vast majority of repetitions for all scenarios (Table S2 in the Supplementary Material). These confirm the strong performance of OSNTF in these scenarios.

However, we find that OSNTF has a higher computational cost compared to non-iterative methods like SCOREplus and regularized spectral

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061



(a) (b)

Figure 2: Scenarios where OSNTF has advantage over other methods. (a) SBM with $K = 4, N = 600$ and increasing average density, (b) DCSBM with $K = 4, N = 600$ and decreasing degree heterogeneity

clustering, and the computational cost is comparable to the iterative methods like SBM refine and DCBM refine. In Figure S1 in the Supplementary Material, we plot the computational cost with increasing number of nodes N. As N increases, the computational cost for the iterative methods, including that of OSNTF, increases quite a bit.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

6. Real data analysis

In this section, we apply the OSNTF methods to five real network datasets with known ground truth community structures and compare their performance with competing methods. All methods are applied to the normalized Laplacian matrix of the networks with their respective regularization techniques when applicable. We briefly describe the datasets below.

We analyze the **political blogs dataset** collected by Adamic and Glance (2005). The dataset comprises of 1490 political blogs during US presidential election with the directed edges indicating hyperlinks. We consider the largest connected component of the graph comprising of 1222 nodes and convert it into an undirected graph by assigning an edge between two nodes if there is an edge between the two in any direction. This dataset with the above mentioned preprocessing was also analyzed by Karrer and Newman (2011); Amini et al. (2013); Qin and Rohe (2013); Jin (2015); Zhao et al. (2012); Gao et al. (2017), etc. for community detection, and is generally considered as a benchmark for evaluating algorithms. The ground truth community assignments partitions this network into two groups, liberal and conservative, according to the political affiliations or leanings of the blogs.

The second dataset is the **Dolphins data** which is an undirected social

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

Table 1: Comparison of OSNTF and Regularized OSNTF with other methods in terms of number of nodes mis-clustered with respect to the ground truth communities. The best two algorithms (least number of nodes mis-clustered) are highlighted in bold for all datasets.

Dataset	Polblogs	Dolphins	Football	Simmons	Emails
N	1222	62	110	1137	1005
K	2	2	11	4	42
Exp. Rand $(N(1-\frac{1}{K}))$	611	31	100	853	981
OSNTF	**55**	**1**	5	**159**	**437**
Reg. OSNTF	66	**1**	5	197	461
OSNTF Conv.	56	2	5	174	**454**
Reg. OSNTF Conv.	64	2	**4**	210	458
Spectral	600	**1**	6	384	531
Reg. Spectral	63	**1**	5	241	467
SCORE	58	5	5	268	730
SCOREplus	**51**	**1**	6	**127**	536
SBM refine	581	2	**4**	278	818
DCBM refine	59	2	**4**	227	480

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

network of associations among 62 dolphins living in Doubtful Sound, New Zealand, curated by Lusseau et al. (2003). During the course of the study, it was observed that a well connected dolphin coded as SN100 left the group, and this resulted into a split of the group into two subgroups. These subgroups consisting of the remaining 61 dolphins constitute our ground truth communities.

The third dataset is the US college **Football** network data from Girvan and Newman (2002), which is a network representation of all Division I games for the season 2000. The ground truth communities for comparison are the conferences the teams belong to.

The fourth dataset is the **Simmons College Facebook** network data that consists of 24257 friendship links among 1137 students (Traud et al., 2012; Jin et al., 2022). The ground truth community labels are the years of graduation for the members.

The final dataset is the **Email-EU-Core** data which is a network of emails exchanged among members of a large European research institute (Yin et al., 2017; Leskovec et al., 2007). The network has 25571 edges among 1005 members. We consider an undirected network, and therefore an edge is recorded between two members if either of them has sent an email to the other member. The ground truth community structure assigns each

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

member to the department they belong to, resulting in 42 communities.

Results: Table 1 summarizes the performance of the proposed methods, along with a number of methods that have appeared in the literature, in terms of the number of nodes mis-clustered for these 5 datasets. The row N shows the number of nodes in the networks, the row K is the number of communities in the ground truth community labels (which is assumed to be known for all methods), and the row "Exp. Rand" provides the expected number of nodes that a random community assignment will mis-cluster and is given by $N(1-\frac{1}{K})$. All methods perform reasonably well in the Dolphins and the Football datasets. In the political blogs data, the spectral clustering and the SBM refine method perform poorly, while the remaining methods perform well. In the Simmons dataset, OSNTF and SCOREplus outperform the others. Finally, in the Emails dataset, OSNTF substantially outperforms all other methods. In all datasets, OSNTF delivers either the best or the second-best performance. Further note that in all cases, the OSNTF method either matches or improves the clustering performance of its initialization, namely, the regularized spectral clustering. Overall, the comparison in terms of performance in real data analysis reveals that the OSNTF methods are competitive with other state-of-the-art methods proposed in the literature.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

7. Discussions and Conclusions

In this paper, we proposed a factorization of the symmetric Laplacian matrix with non-negativity and orthogonality constraints (OSNTF) for community detection in network data. The factorization approximates the Laplacian matrix (or a regularized version of it) with a matrix that has an exact OSNTF by solving an optimization problem. We derived nonasymptotic upper bounds on the error rate of community detection using the OSNTF method (assuming global optimizer can be found for the approximation optimization problem) in graphs generated from the SBM and DCSBM.

This method is quite similar to spectral clustering, and attempts to estimate the same discriminating subspace as spectral clustering for a block-diagonal Laplacian matrix that corresponds to a graph with K connected components. However, for more general graphs the two methods obtain very different invariant subspaces for discrimination. Our simulations show that this method outperforms spectral clustering in a wide variety of situations. In particular, for sparse graphs and for graphs with high degree heterogeneity, this method does not suffer from some of the issues spectral clustering faces. While it is clear from Eckart-Young theorem that spectral clustering uses the best K dimensional subspace that represents the data, the subspace may not be the best discriminating subspace for clustering.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

Supplementary Material

The online Supplementary Material contains proofs of theoretical results and additional tables and figures.

Acknowledgments

This work was supported in part by NSF grants DMS-1830547, CCF-1934986, DMS-2015561 and DMS-2413610.

References

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43. ACM.

Amini, A. A., A. Chen, P. J. Bickel, E. Levina, et al. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics 41*(4), 2097–2122.

Arora, R., M. Gupta, A. Kapila, and M. Fazel (2011). Clustering by left-stochastic matrix factorization. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 761–768.

Berman, A. (2003). *Completely Positive Matrices.* World Scientific.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*(10), P10008.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012). Stochastic blockmodels with a growing number of classes. *Biometrika 99*, 273–284.

Choi, S. (2008). Algorithms for orthogonal nonnegative matrix factorization. In *2008 IEEE In-*

ternational Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1828–1832. IEEE.

Cichocki, A., R. Zdunek, A. H. Phan, and S.-i. Amari (2009). *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons.

Daudin, J. J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and Computing 18*, 173–183.

Ding, C., T. Li, W. Peng, and H. Park (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 126–135.

Ding, C. H., X. He, and H. D. Simon (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM International Conference on Data Mining*, Volume 5, pp. 606–610.

Donoho, D. and V. Stodden (2004). When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16*, pp. 1141–1148. MIT Press.

Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou (2017). Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research 18*(1), 1980–2024.

Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou (2018). Community detection in degree-corrected block models. *The Annals of Statistics 46*(5), 2153–2185.

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

Gray, L. J. and D. G. Wilson (1980). Nonnegative factorization of positive semidefinite nonnegative matrices. *Linear Algebra and Its Applications 31*, 119–127.

Hajek, B., Y. Wu, and J. Xu (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory 62*(5), 2788–2797.

Statistica Sinica

Holland, P., K. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: some first steps. *Social Networks 5*, 109–137.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research 5*, 1457–1469.

Huang, K., N. Sidiropoulos, and A. Swami (2014). Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing 62*(1), 211–224.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics 43*(1), 57–89.

Jin, J., Z. T. Ke, and S. Luo (2022). Improvements on score, especially for weak signals. *Sankhya A 84*(1), 127–162.

Karrer, B. and M. E. J. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(1), 016107.

Kim, J. and H. Park (2008). Sparse nonnegative matrix factorization for clustering. technical report, georgia institute of technology.

Kuang, D., H. Park, and C. H. Ding (2012). Symmetric nonnegative matrix factorization for graph clustering. In *SIAM International Conference on Data Mining*, Volume 12, pp. 106–117.

Laurberg, H., M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen (2008). Theorems on positive data: On the uniqueness of nmf. *Computational Intelligence and Neuroscience*, article 764206.

Le, C. M., E. Levina, and R. Vershynin (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms 51*(3), 538–561.

Lee, D. D. and H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing*

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

Systems 13, pp. 556–562. MIT Press.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics 43*(1), 215–237.

Leskovec, J., J. Kleinberg, and C. Faloutsos (2007). Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD) 1*(1), 2–es.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Computation 19*(10), 2756–2779.

Lusseau, D., K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology 54*(4), 396–405.

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

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537.

Mirzal, A. (2014). A convergent algorithm for orthogonal nonnegative matrix factorization. *Journal of Computational and Applied Mathematics 260*, 149–166.

Mirzal, A. (2017). A convergent algorithm for bi-orthogonal nonnegative matrix tri-factorization. *arXiv preprint arXiv:1710.11478*.

Montanari, A. and S. Sen (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 814–827.

Newman, M. E. J. and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E 69*(2), 026113.

Ng, A. Y., M. I. Jordan, Y. Weiss, et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 2*, 849–856.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

Pompili, F., N. Gillis, P.-A. Absil, and F. Glineur (2014). Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. *Neurocomputing 141*, 15–25.

Psorakis, I., S. Roberts, M. Ebden, and B. Sheldon (2011). Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E 83*(6), 066114.

Qin, T. and K. Rohe (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3120–3128. Curran Associates, Inc.

Radjavi, H. and P. Rosenthal (2003). *Invariant Subspaces*. Dover Publications.

Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*(4), 1878–1915.

Sarkar, P. and P. J. Bickel (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics 43*(3), 962–990.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology 31*(1), 361–395.

Stewart, G. W. and J.-g. Sun (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA.

Tang, M. and C. E. Priebe (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics 46*(5), 2360–2415.

Traud, A. L., P. J. Mucha, and M. A. Porter (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications 391*(16), 4165–4180.

Vavasis, S. A. (2009). On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization 20*(3), 1364–1377.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*(4), 395–416.

Wang, F., T. Li, X. Wang, S. Zhu, and C. Ding (2011). Community discovery using nonnegative

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0061

matrix factorization. *Data Mining and Knowledge Discovery 22*(3), 493–521.

Xu, W., X. Liu, and Y. Gong (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 267–273. ACM.

Yang, Z., T. Hao, O. Dikmen, X. Chen, and E. Oja (2012). Clustering by nonnegative matrix factorization using graph random walk. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, pp. 1079–1087. Curran Associates, Inc.

Yin, H., A. R. Benson, J. Leskovec, and D. F. Gleich (2017). Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 555–564.

Zhao, Y., E. Levina, and J. Zhu (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics 40*, 2266–2292.

Department of Statistics, The Ohio State University, Columbus, OH 43210

E-mail: paul.963@osu.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820

E-mail: yuguo@illinois.edu

Statistica Sinica