Adaptive Overlapping Community Detection with Bayesian NonNegative Matrix Factorization

Xiaohua Shi[1,2(✉)], Hongtao Lu[1], and Guanbo Jia[3]

[1] MOE-Microsoft Laboratory for Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering,
Shanghai JiaoTong University, Shanghai, China
xhshi@sjtu.edu.cn

[2] Library, Shanghai Jiaotong University, Shanghai, China

[3] University of Birmingham, Birmingham, UK

Abstract. Overlapping Community Detection from a real network is unsupervised, and it is hard to know the exact community number or quantized strength of every node related to each community. Using Nonnegative Matrix Factorization (NMF) for Community Detection, we can find two non-negative matrices from whole network adjacent matrix, and the product of two matrices approximates the original matrix well. With Bayesian explanation in factorizing process, we can not only catch most appropriate count of communities in a large network with Shrinkage method, but also verify good threshold how a node should be assigned to a community in fuzzy situation.

We apply our approach in some real networks and a synthetic network with benchmark. Experimental results for overlapping community detection show that our method is effective to find the communities number and overlapping degree, and achieve better performance than other existing overlapping community detection methods.

Keywords: Overlapping community detection · Non-negative matrix factorization · Bayesian inference · Automatic relevance determination

1 Introduction

Overlapping Community Detection is an important approach in complex networks to understand and analysis large network character [3,50], such as social network [30,49], collaborative network [39], and biological network [1]. We can find most correlated overlapping sub-communities to simplify global structure to understand the network topology, and keep original network with overlapping structure especially in density network.

It is a recognition with community detection that nodes in same community are densely connected, and nodes in different communities are sparsely connected. A node can be allocated into different communities in overlapping situation [55]. We can find overlapping communities with methods as clique percolation techniques [23], random walk [18], label propagation [12,51], seed

S. Candan et al. (Eds.): DASFAA 2017, Part II, LNCS 10178, pp. 339–353, 2017.
DOI: 10.1007/978-3-319-55699-4_21

expansion [47], objective function optimization (modularity or other function) [35], or statistical inference [11,40,48]. Overlapping communities can also be detected based on the graph partitioning approach, which tries to find underling clusters from minimize the number of edges between communities [8,43].

Macropol *et al.* [29] propose a biologically sensitive algorithm based on repeated random walks (RRW) for discovering functional modules, e.g., complexes or pathways, within large-scale protein networks. RRW considers the element of network topology, edge weights, and long range interactions between proteins. Zhang *et al.* [53] propose a learning algorithm which can learn a node-community membership matrix via stochastic gradient descent with bootstrap sampling. Lee *et al.* [25] introduce a community assignment algorithm named Greedy Clique Expansion (GCE). GCE algorithm identifies distinct cliques as seeds and expands these seeds by greedily optimizing a local fitness function.

In many clustering applications, object data is nonnegative due to their physical nature, e.g., images are described by pixel intensities and texts are represented by vectors of word counts. As to a graph-based network, the adjacency matrix (or weighted adjacency matrix) $\mathbf{A}$ as well as the Laplacian matrix completely represents the structure of network, and $\mathbf{A}$ is non-negative naturally. Meanwhile, Nonnegative Matrix Factorization (NMF) was originally proposed as a method for dimension reduction and finding matrix factors with parts-of-whole interpretations [15,27]. Based on the consideration that there is no any physical meaning to reconstruct a network with negative adjacency matrix, using NMF to obtain new representations of network with non-negativity constraints can achieve much productive effect in overlapping community analysis [52,53]. It is likely an efficient network partition tool to find the communities because of its powerful interpretability and close relationship with other clustering methods. Overlapping community detection with NMF can capture the underlying structure of network in the low dimensional data space with its community-based representations [41]. Zhang *et al.* [54] propose a method called bounded nonnegative matrix tri-factorization (BNMTF) with three factors in the factorization, and explicitly model and learn overlapping community membership of each node as well as the interaction among communities.

NMF decomposes a given nonnegative data matrix $\mathbf{X}$ as $\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathbf{T}}$ where $\mathbf{U} \geq \mathbf{0}$ and $\mathbf{V} \geq \mathbf{0}$ (meaning that U and V are *component-wise nonnegative*). Tan *et al.* [45] addresses the estimation of the latent dimensionality in nonnegative matrix factorization (NMF) with the β-divergence, and proposes for maximum a posteriori (MAP) estimation with majorization-minimization (MM) algorithms. Psorakis *et al.* [40] presents a novel approach to community detection that utilizes the Bayesian non-negative matrix factorization model to extract overlapping modules from a network.

In this paper, we propose an adaptive Bayesian non-negative matrix factorization (ABNMF) method for overlapping community detection. In a Bayesian framework, ABNMF assumpts that original matrix $\mathbf{X}$ with object matrix $\mathbf{U}$ and $\mathbf{V}$ follow a certain probability distribution. In this way, we expect that ABNMF can obtain a relevant count of communities and quantized strength of each node related to every community from original network data. To achieve this, we design a new non-negative matrix factorization objective function by incorporating Bayesian Detection, and suggest an adaptive node-based threshold for different communities. Our experiments show that the proposed approach can validly estimate relevant dimension in lower space, find suitable overlapping communities, and also achieve better performance than the state-of-arts overlapping methods.

2 Related Works

Let $\mathbf{X}$ be a $m \times n$ non-negative matrix, and NMF consists in finding an approximation:

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $m \times k$ and $n \times k$ non-negative matrices. The factorization rank $\mathbf{k}$ is often chosen such that $k \ll \min(m, n)$. The objective behind this choice is to summarize and split the information contained in $\mathbf{U}$ into $\mathbf{k}$ factors (the columns of $\mathbf{U}$). Depending on the application field, these factors are given different names: basis images, metagenes or source signals. In community detection, we equivalently and alternatively use the terms primary communities to refer to matrix $\mathbf{U}$, and mixture coefficient matrix or communities assignment profiles to refer with matrix $\mathbf{V}$. We examine each row of $\mathbf{V}$, and assign node $\mathbf{x}_j$ to community c if $c = \arg\max_c v_{jc}$ [44] in non-overlapping community detection like crisp clustering. If we define a proper threshold set δ, a node j can be assigned into community c if $v_{jc} \geq \delta_c$ in overlapping situation like fuzzy clustering [37].

The main approach of NMF is to estimate matrices $\mathbf{U}$ and $\mathbf{V}$ as a local minimum with a cost function in some distance metric. Generally we use β-Divergence $\mathbf{D}_\beta(\mathbf{X}; \mathbf{U}\mathbf{V}^\mathbf{T})$ [7]. When $\beta = 0, 1, 2, \mathbf{D}_\beta(\mathbf{X}; \mathbf{U}\mathbf{V}^\mathbf{T})$ is proportional to the (negative) log-likelihood of the Itakara-Saito (IS), KL and Euclidean noise models up to a constant.

Recently, Bayesian inference has been introduced into NMF with a noise E between $\mathbf{X}$ and $\mathbf{U}\mathbf{V}^\mathbf{T}$.

$$\mathbf{X} = \mathbf{U}\mathbf{V}^\mathbf{T} + \mathbf{E} \tag{2}$$

Morten *et al.* [31] demonstrate how a Bayesian framework for model selection based on Automatic Relevance Determination (ARD) can be adapted to the Tucker and CandeComp/PARAFAC (CP) models. By assigning priors for the model parameters and learning the hyperparameters of these priors the method is able to turn off excess components and simplify the core structure at a computational cost of fitting the conventional Tucker/CP model. Morten *et al.* [32] also formulate a non-parametric Bayesian model for community detection consistent with an intuitive definition of communities, and present a Markov chain Monte Carlo procedure for inferring the community structure.

Automatic Relevance Determination is a hierarchical Bayesian approach that widely used for model selection. In ARD, hyperparameters explicitly represent the relevance of different features by defining the range of variation for these features, and are usually by modeling the width of a zero-mean prior imposed on the model parameters. If the width becomes zero, the corresponding feature cannot have any effect on the prediction. Hence, ARD optimizes these hyperparameters to discover which features are relevant. While ARD based on Gaussian or Poisson priors, we can prune excess components by admitting sparse representation and retain active components. Applying ARD in some real network community detection process, we can effectively find the relevant communities number without knowing in advance.

Jin *et al.* [17] extend the stochastic model method to detection of overlapping communities with the virtue of autonomous determination of the number of communities. Their approach hinges upon the idea of ranking node popularities within communities and using a Bayesian method to shrink communities to optimize an objective function based on the stochastic generative model. Wang *et al.* [46] propose a probabilistic model, Dynamic Bayesian Nonnegative Matrix Factorization, for automatic detection of overlapping communities in temporal networks. Their model can not only give the overlapping community structure based on the probabilistic memberships of nodes in each snapshot network but also automatically determines the number of communities in each snapshot network based on automatic relevance determination.

Schmidt *et al.* [42] present a Bayesian treatment of NMF based on a Gaussian likelihood and exponential priors, and approximate the posterior density of the NMF factors. This model equals to minimize the squares Euclidean distance $\mathbf{D_2}(\mathbf{X};\mathbf{UV^T})$ for NMF. Cemgil [5] proposes NMF models with a KL-divergence error measure in a statistical framework with a hierarchical generative model consisting of an observation and a prior component. We can see that this models of $\mathbf{D_1}(\mathbf{X};\mathbf{UV^T})$ is equals to NMF model with Poisson noise likelihood:

$$P(n;\lambda) = \frac{\lambda^n}{n!}exp(-\lambda) \tag{3}$$

$$P(X|U,V) = \prod_i \prod_j \frac{[UV^T]_{ij}^{X_{ij}} exp(-[UV^T]_{ij})}{X_{ij}!} \tag{4}$$

We further assume that all entries of X are independent of each other (the dependency structure is later induced by the matrix product), we can write:

$$ln(P(X|U,V)) = \sum_i \sum_j X_{ij} ln[UV^T]_{ij} - [UV^T]_{ij} - ln(X_{ij}!) \tag{5}$$

We use Stirling's formula $ln(n!) \approx nln(n) - n$ for $n >> 1$ to get approximated expression:

$$ln(P(X|U,V)) \approx \sum_i \sum_j X_{ij} ln\frac{[UV^T]_{ij}}{X_{ij}} - [UV^T]_{ij} + X_{ij} \tag{6}$$

3 Overlapping Community Detection with Bayesian NMF

3.1 Bayesian NMF Model

In this section, we introduce Bayesian inference process of our Adaptive Bayesian NMF (ABNMF) method. Given a network **G** consisting of n nodes $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$, we can represent the network as matrix **X** transformed from adjacency matrix. In our ABNMF processing, the diagonal elements are defined to be 1 rather than 0 as in usual clustering cases, and **X** is $n \times n$ square matrix and non-negative.

We consider there lies a relation between original network matrix **X** and combination of factorized matrix $\mathbf{UV^T}$. The distribution of this relation can be Gaussian [42] or Poisson [26] model. As Poisson noise model algorithm have much better performance than Gaussian noise models [19,21] to achieve better sparse estimation effect, we select Poisson likelihood in our ABNMF method. In maximum-likelihood solution to find **U** and **V**, $\mathbf{P(X|UV^T)}$ is maximized, or its energy function $-\mathbf{logP(X|UV^T)}$ is minimized.

To simplify likelihood with positive error [10,36], we chose the relation of $\mathbf{U}, \mathbf{V}$ and **X** as $X_{ij} \sim Poisson(\sum_k U_{ik} * V_{kj})$. In this Poisson model, the log-likelihood of **X** and $\mathbf{UV^T}$ is:

$$\begin{aligned} -\, ln(P(X|UV^T)) &= -\sum_i \sum_j \left\{ X_{ij} ln \tfrac{[UV^T]_{ij}}{X_{ij}} - [UV^T]_{ij} + X_{ij} \right\} \\ &= \quad -Xln(UV^T) + \mathbf{1}UV^T\mathbf{1}^T + const(X) \end{aligned} \tag{7}$$

where **1** is an n × n matrix with every elements equal to 1. We use independent half-normal prior over every column of **U** and **V**, where the mean is zero and precisian is β_j:

$$\begin{aligned} p(u_{ij}|\beta_j) &= \mathcal{HN}(x|0, \beta_j^{-1}) \\ p(v_{jk}|\beta_j) &= \mathcal{HN}(x|0, \beta_j^{-1}) \end{aligned} \tag{8}$$

when

$$\mathcal{HN}(x|0, \beta^{-1}) = \sqrt{\frac{2}{\pi}} \beta^{\frac{1}{2}} exp(-\frac{1}{2}\beta x^2) \tag{9}$$

We define the diagonal matrix **B** with $[\beta_1, ..., \beta_K]$ and zeros elsewhere, and the negative log priors of **U** and **V** are:

$$\begin{aligned} -\, ln(p(U|\beta)) &= \sum_i \sum_j \frac{1}{2}\beta_j u_{ij}^2 - \sum_j \frac{N}{2} log\beta_j + const \\ -ln(p(V|\beta)) &= \sum_j \sum_k \frac{1}{2}\beta_j v_{jk}^2 - \sum_j \frac{N}{2} log\beta_j + const \end{aligned} \tag{10}$$

At last, we set the independent prior distribution of β_j as a Gamma distribution with parameters a_j and b_j:

$$p(\beta_j|a_j, b_j) = \frac{b_j^{a_j}}{\Gamma(a_j)}\beta_j^{a_j-1}exp(-\beta_j b_j) \tag{11}$$

The negative log of β_j is:

$$-ln(p(\beta)) = \textstyle\sum_j [\beta_j b_j k - (a_j-1)ln\beta_j] + const \tag{12}$$

The MAP(Maximum a Posteriori) of ABNMF is:

$$\mathcal{U} = -lnP(X|UV^T)) - lnP(U|\beta)) - lnP(V|\beta)) - lnP(\beta)) \tag{13}$$

3.2 Iteration Rules of ABNMF

From Eq. (13), we can derive the multiplicative update rules of ABNMF with Poisson likelihood. Let ϕ_{ij}, ψ_{jk} be the Lagrange multiplier for constraint $u_{ij} \geq 0$ and $v_{jk} \geq 0$, respectively, and $\mathbf{\Phi} = [\phi_{ij}]$, $\mathbf{\Psi} = [\psi_{jk}]$. The Lagrange function $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{U} + \mathrm{tr}(\mathbf{\Phi U^T}) + \mathrm{tr}(\mathbf{\Psi V^T}) \tag{14}$$

Let the derivatives of $\mathcal{L}$ with respect to $\mathbf{U}$ or $\mathbf{V}$ vanish, we have:

$$\frac{\partial \mathcal{L}}{\partial U} = -2 * \frac{X}{UV^T}U + 2 * \mathbf{1}U + 2 * BU + \mathbf{\Phi} = 0 \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2 * \frac{X}{UV^T}V + 2 * \mathbf{1}V + 2 * BV + \mathbf{\Psi} = 0 \tag{16}$$

Using the KKT conditions $\phi_{ij}u_{ij} = 0$ and $\psi_{jk}v_{jk} = 0$, we get the following equations for u_{ij}, v_{jk}:

$$u_{ij} \longleftarrow u_{ij}\left(\frac{X}{UV^T}\right)_{ij}\left(\frac{U}{\mathbf{1}U + UB}\right)_{ij} \tag{17}$$

$$v_{jk} \longleftarrow v_{jk}\left(\frac{X}{UV^T}\right)_{jk}\left(\frac{V}{\mathbf{1}V + VB}\right)_{jk} \tag{18}$$

and the β_j will be updated below:

$$\beta_j \longleftarrow \frac{n + a_j - 1}{\frac{1}{2}(\sum_i u_{ij}^2 + \sum_k v_{jk}^2) + b_j} \tag{19}$$

We can get an approximate fixed value in convergence for iteration. Suppose the multiplicative updates stop after t iterations with parameters from Table 1, the overall computational complexity for ABNMF will be $O(tn^2c + n^2)$. A relatively small initial c will save running time of the algorithm.

Table 1. Parameters used in complexity analysis

Parameters	Description
n	Number of network nodes
c	Number of initial communities count
β_j	Paraments of communities number
a_j	Hyper-hyperparaments a
b_j	Hyper-hyperparaments b

3.3 Determination of Overlapping Community Number K and Threshold δ

In regular NMF methods for clustering, the object factorized dimension K should be given. But in community detection situation, we just know the relation of nodes without prior information of community number K, and it's hard to count out the suitable number. If K is too small, some communities will be very large and the model can not be fitted well. On contrary, If K is too large, we can not catch the group character effectively from an entire network and occur into overfitting. We need to find K with a appropriate solution between network fineness and overfitting.

To solve this problem, we propose a statistical *shrinkage* method in a Bayesian framework to find the number of communities and build a model selection method based on Automatic Relevance Determination [31,45]. In ABNMF, we principally iterate out v_{jk} with gradual change, and the prior will try to promote a *shrinkage* to zero of v_{jk} with a rate constant proportional to β_j. A large β_j represents a belief that the half-normal distribution over v_{jk} has small variance, and hence v_{jk} is expected to get close to zero. We can see the priors and the likelihood function (quantifying how well we explain the data) are combined with the effect that columns of V which have little effect in changing how well we explain the observed data will shrink close to zero. We can effectively estimate the communities number K by computer the non-zero column number from V with initial rank c.

In overlapping fuzzy detection, a sparse or dense network may have different overlapping degree. A dense network may contain more communities overlapped. Network density p describes the portion of the potential connections in a network that are actual connections. Every node potentially has a basic probability p to connect with rest nodes in a network, regardless of whether or not they actually connect:

$$p = \frac{2 * |E|}{n * (n-1)} \tag{20}$$

There is a fact that nodes shared multiple community memberships receive multiple chances to create a link in overlapping assumption. We may assume each overlapping sub-community is larger than a potential network, that refer every v_{ij} will large than one fixed threshold in each network. Yang *et al.* [52] suggest that the threshold value can be $\delta = \sqrt{-log(1-p)}$ to achieve good performance. Note that this process adaptively generates an increasing relationship between edge probability and the number of shared communities.

3.4 Performance Comparisons in Different Networks

We compare our algorithm with other four popular overlapping community detection methods. Five algorithms are listed below:

1. CFinder tries to find overlapping dense groups of nodes in networks, and is based on method Clique Percolation Method (CPM) [38].
2. COPRA (Community Overlap PRopagation Algorithm) is based on the label propagation technique for finding overlapping community structure in large networks [12].
3. OSLOM (Order Statistics Local Optimization Method) locally optimizes the statistical significance information of a cluster with respect to random fluctuation with Extreme and Order Statistics [24].
4. LCM (Link Communities Method) organizes community structures spanning inner-city to regional scales while maintaining pervasive overlap, and builds blocks that reveal overlap and hierarchical organization in networks [2].
5. ABNMF (Adaptive Bayesian Non-negative Matrix Factorization) with Poisson likelihood. Its overlapping threshold is related with network density.

We run OSLOM, LFM and ABNMF in different six network datasets without groundtruth to evaluate its communities number, overlap fraction and modularity. Then we generate a synthetic network with 5000 nodes in different overlap fraction [22]. The details of experiments are stated below:

(1). In ABNMF methods, we select 10 different initial communities count c and apply 10 independent experiments. Every experiment iterates for 500 times.
(2). We test the ABNMF method in Email network [14] to evaluate the performance with different initial dimension number of c in Table 1.
(3). We use six different size and different character networks to compare communities number, overlap fraction and Modularity [35]. **Football** (American College football), **Email** (Email network of University at Rovira i Virgili in Tarragona, Spain), and **PGP** (Pretty Good Privacy communication network) [9,13,14] are social networks. **Erdos** (Collaboration network with famous mathematician Erdos) and **Cmat** (Condensed matter collaborations 2003) [4,34] are collaborative networks. **Metabolic** (Metabolic Network) [16] is biological network.
(4). We use Omega Index [6,33] to evaluate overlapping communities detecting performance with benchmark in a synthetic network.

Modularity has widely used to measure the strength of non-overlapping or overlapping community structure found by community detection methods. In Eq. (21), A_{ij} is the adjacency matrix, and k_i, k_j are node degree of i, j. $\delta(c_i, c_j)$ is probability of having a link between i and j in the null model are weighted by the belonging of i and j to the same community, since $\delta(c_i, c_j)$ is equal to 1 only when i and j belong to the same community, and it is 0 otherwise.

$$Q_{ov} = \frac{1}{n} \sum_{i,j \in V} \left[A_{ij}\delta(c_i, c_j) - \frac{k_i k_j}{2n}\delta(c_i, c_j) \right] \tag{21}$$

The Omega Index can evaluate the extent of two different solutions for overlapping communities in which each pair of nodes is estimated to share same community:

$$Omega(C_1, C_2) = \frac{\sum_{j=0}^{min(J,K)} \frac{A_j}{N} - \sum_{j=0}^{min(J,K)} \frac{N_{j1}N_{j2}}{N^2}}{1 - \sum_{j=0}^{min(J,K)} \frac{N_{j1}N_{j2}}{N^2}} \tag{22}$$

where J and K represent the maximum number of communities in which any pair of nodes appears together in solution C_1 and C_2, respectively, A_j is the number of the pairs agreed by both solutions to be assigned to number of community j, and N is the number of pairs of nodes. N_{j1} is the total number of pairs assigned to number of communities j in solution C_1, and N_{j2} is the total number of pairs assigned to number of communities j in solution C_2.

Table 2. Overlapping community number K with different initial c in ABNMF

	c	K	O	Q_{ov}
1	23	22	0.3600	0.6876
2	30	26	0.3668	0.6887
3	52	30	0.3772	0.6975
4	76	34	0.3768	0.6960
5	114	35	0.3862	0.7058
6	227	36	0.3845	0.7063
7	378	35	0.3846	0.7064
8	567	37	0.3900	0.7133

We run ABNMF method to test the impact of different initial communities numbers c in Email network. Table 2 lists the different result K, relevant overlap fraction(O) and Modularity(Q_{ov}). ARD is effective well on features extraction with large initial c, and contractive communities count K is around 30 in **Email** network. We can find that different c has weak influence for O and Q_{ov} results when c is set from 567 to 52. In ABNMF, we choose the initial count c from 1/5 to 1/10 of total nodes n to keep the performance of algorithm and keep the operational efficiency.

3.5 Overlapping Community Detection in Different Network

On American Football Game real network with 115 nodes and 613 edges, we run ABNMF for case study and the visualization of our found overlapping community structure is shown in Fig. 1, where same color nodes are allocated into same overlapping community. From Fig. 1, we can see that our proposed method ABNMF automatically finds 10 strong sense communities which are gathered by crisp clustering, and most of the football teams are correctly assigned into their corresponding communities in our found overlapping community structure. Moreover, it is very interesting to note that our proposed method ABNMF detects 32 overlapping nodes in different communities in total, in which each overlapping node has two different colors indicating different communities the node belongs to. This is because, besides against other football teams in the same conference, these football teams corresponding to the overlapping nodes also frequently play many games against football teams in other conferences. Therefore, we can see that our proposed method ABNMF has a good performance in detecting overlapping community structures in this real world social network.

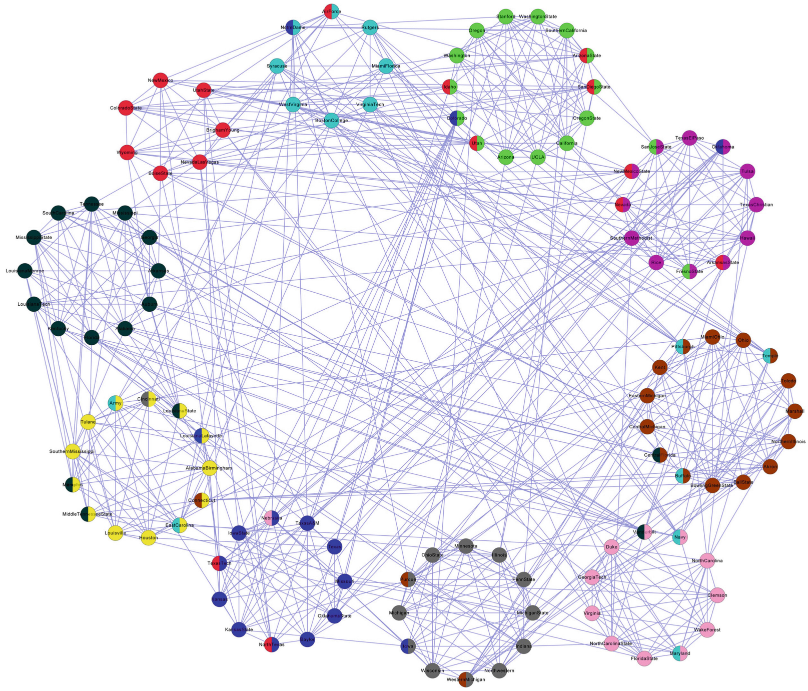

Fig. 1. Overlapping communities of Football network obtained by ABNMF.

We select 6 popular networks with different size, and compare community number(K), overlap fraction(O) and overlap modularity(Q_{ov}) in OSLOM, LCM and ABNMF. We can find from Table 3 that, our method ABNMF can effectively find overlapping community number and is highly close to results of OSLOM and LFM. ABNMF detects much dense communities in overlap fraction and achieves high overlapping modularity than other two methods. In ABNMF with

Multiplicative Update Rules [27], we achieve good performance in large **PGP** and **Cmat** networks both of which have more than ten thousands of nodes. We may also combine with Projected Gradient [28] method or Block Gradient Descent method [20] to solve Eq. (13) in larger datasets with millions of nodes.

Table 3. Overlapping community detection comparison on different networks

Network	Nodes	OSLOM			LCM			ABNMF		
		K	O	Q_{ov}	K	O	Q_{ov}	K	O	Q_{ov}
Football	115	9	0.1956	0.6032	11	0.2134	0.5992	10	0.2444	**0.6609**
Metabolic	453	32	0.4347	0.4212	31	0.4525	0.4678	27	0.5055	**0.6919**
Email	1133	28	0.2567	0.5796	27	0.2754	0.5821	26	0.3766	**0.6982**
Erdos	6927	77	0.2765	0.7187	81	0.2897	0.6837	73	0.3098	**0.8203**
PGP	10680	233	0.4688	0.8782	230	0.478	0.8843	227	0.4944	**1**
Cmat	27519	486	0.356	0.7216	483	0.4121	0.7255	475	0.534	**0.7340**

We evaluate the performance of our proposed algorithm on the LFR synthetic networks with benchmark, and compare with other four overlapping community detection algorithms. The LFR (Lancichinetti-Fortunato-Radicchi) benchmark [22] provides a class of artificial networks in which both the degrees of the nodes and the sizes of the communities follows power laws the same as many real-world networks. Here, we adopt a LFR benchmark with 5000 nodes respectively from the benchmark generator source code[1] in our experiment:

benchmark -N 5000 -k 10 -maxk 30 -mu 0.1 -minc 10 -maxc 50 -on 50 -om 2

In this LFR benchmark, we set the average degree of nodes $davg = 10$, the maximum degree of nodes $dmax = 30$, the minimum community size $minc = 10$, the maximum community size $maxc = 50$, the exponents of the power law of the community size distribution $t_1 = 1$, the exponents of the power law of the community size distribution $t_2 = 2$, the overlapping nodes in the entire network $on = 50$, and the number of communities that each overlapping node belongs to $om = 2$. Moreover, we define the mixing parameter μ as the average percentage of edges that connect a node to those in other communities which indicates that every node shares a fraction $(1-\mu)$ edges with other nodes in its community and a fraction μ edges with nodes outside its community. The network community structure will be weakened by increasing μ.

Five algorithms are executed on the LFR benchmark network, and the average Omega Index is used to measure the similarities between the known community structure and the obtained resultant community structure by these algorithms. The results of different algorithms in the LFR networks are shown in Table 4.

It can be seen that all these five algorithms perform well and our proposed ABNMF algorithm has slightly better performance comparing with the other

[1] https://sites.google.com/site/santofortunato/inthepress2.

Table 4. Omega index comparison on LFR 5000 network

Overlap fraction	CPM	COPRA	OSLOM	LPM	ABNMF
0.05	0.86	0.86	0.86	0.86	**0.89**
0.1	0.83	0.855	0.855	0.855	**0.88**
0.15	0.81	0.85	0.85	0.83	**0.86**
0.2	0.75	0.84	0.84	0.81	**0.86**
0.25	0.6	0.82	0.82	0.8	**0.85**
0.3	0.47	0.82	0.83	0.8	**0.84**
0.35	0.45	0.79	0.82	0.8	**0.83**
0.4	0.4	0.77	0.81	0.8	**0.83**
0.45	0.38	0.74	0.8	0.79	**0.82**
0.5	0.32	0.71	0.79	0.75	**0.81**
0.55	0.3	0.64	0.6	0.74	**0.8**
0.6	0.22	0.62	0.62	0.73	**0.77**
0.65	0.2	0.1	0.14	0.58	**0.7**

four algorithms when the value of overlap fraction μ is small on the LFR network. Moreover, as the value of μ increasing, the performance of our proposed algorithm does not degrade rapidly as shown in Table 4. Therefore, our proposed algorithm has a good ability to detect overlapping community structures in complex networks no matter whether they have dense or sparse overlapping structure.

4 Conclusions

In this paper, we solve an overlapping community detection problem using Adaptive Bayesian NMF. We propose a model that considerate Bayesian inference process with Poisson model into NMF, and derive the updating rules and conduct experiments to valid our model. We also apply Automatic Relevance Determination method with sparse constrain to learn the community count of a network, and compare the detection impact of different initial community rank. At last, we adaptively select a most proper value related to network density as overlapping threshold for mixture coefficient matrix. Our method can be applied in real network data without any given information, and achieves good performance than other overlapping community detection methods.

Acknowledgments. This work was supported by NSFC (No. 61272247, 61533012, 61472075), the Basic Research Project of "Innovation Action Plan" (16JC1402800), the Major Basic Research Program (15JC1400103) of Shanghai Science and Technology Committee, and the Arts and Science Cross Special Fund (13JCY14) of Shanghai JiaoTong University.

References

1. Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics **22**(8), 1021–1023 (2006)
2. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
3. Amelio, A., Pizzuti, C.: Overlapping community discovery methods: a survey. In: Gündüz-Öğüdücü, Ş., Etaner-Uyar, A.Ş. (eds.) Social Networks: Analysis and Case Studies. LNSN, pp. 105–125. Springer, Vienna (2014). doi:10.1007/978-3-7091-1797-2_6
4. Batagelj, V., Mrvar, A.: Some analyses of Erdos collaboration graph. Soc. Netw. **22**(2), 173–186 (2000)
5. Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. Comput. Intell. Neurosci. **2009**, 1–17 (2009)
6. Collins, L.M., Dent, C.W.: Omega: a general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. Multivar. Behav. Res. **23**(2), 231–242 (1988)
7. Fevotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the beta-divergence. Neural Comput. **23**(9), 2421–2456 (2011)
8. Gama, F., Segarra, S., Ribeiro, A.: Overlapping clustering of network data using cut metrics, pp. 6415–6419. IEEE (2016)
9. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
10. Gopalan, P., Ruiz, F.J., Ranganath, R., Blei, D.M.: Bayesian nonparametric Poisson factorization for recommendation systems. In: AISTATS, pp. 275–283 (2014)
11. Gopalan, P.K., Gerrish, S., Freedman, M., Blei, D.M., Mimno, D.M.: Scalable inference of overlapping communities. In: Advances in Neural Information Processing Systems, pp. 2249–2257 (2012)
12. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)
13. Guardiola, X., Guimera, R., Arenas, A., Diaz-Guilera, A., Streib, D., Amaral, L.: Macro-and micro-structure of trust networks. arXiv preprint arXiv:cond-mat/0206240 (2002)
14. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Phys. Rev. E **68**(6), 065103 (2003)
15. He, Y.C., Lu, H.T., Huang, L., Shi, X.H.: Non-negative matrix factorization with pairwise constraints and graph Laplacian. Neural Process. Lett. **42**(1), 167–185 (2015)
16. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
17. Jin, D., Wang, H., Dang, J., He, D., Zhang, W.: Detect overlapping communities via ranking node popularities. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
18. Jin, D., Yang, B., Baquero, C., Liu, D., He, D., Liu, J.: A Markov random walk under constraint for discovering overlapping communities in complex networks. J. Stat. Mech: Theor. Exp. **2011**(05), P05031 (2011)
19. Kaganovsky, Y., Han, S., Degirmenci, S., Politte, D.G., Brady, D.J., O'Sullivan, J.A., Carin, L.: Alternating minimization algorithm with automatic relevance determination for transmission tomography under poisson noise. SIAM J. Imaging Sci. **8**(3), 2087–2132 (2015)

20. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Global Optim. **58**(2), 285–319 (2014)
21. Kucukelbir, A., Ranganath, R., Gelman, A., Blei, D.: Automatic variational inference in stan. In: Advances in Neural Information Processing Systems, pp. 568–576 (2015)
22. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**(1), 016118 (2009)
23. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
24. Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding statistically significant communities in networks. PloS one **6**(4), e18961 (2011)
25. Lee, C., Reid, F., McDaid, A., Hurley, N.: Detecting highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827 (2010)
26. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 13 (2001)
27. Lee, D., Seung, H., et al.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
28. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
29. Macropol, K., Can, T., Singh, A.K.: Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. BMC Bioinf. **10**(1), 1 (2009)
30. Meena, J., Devi, V.S.: Overlapping community detection in social network using disjoint community detection. In: 2015 IEEE Symposium Series on Computational Intelligence, pp. 764–771. IEEE (2015)
31. Mørup, M., Hansen, L.K.: Automatic relevance determination for multi-way models. J. Chemometr. **23**(7–8), 352–363 (2009)
32. Mørup, M., Schmidt, M.N.: Bayesian community detection. Neural Comput. **24**(9), 2434–2456 (2012)
33. Murray, G., Carenini, G., Ng, R.: Using the omega index for evaluating abstractive community detection. In: Association for Computational Linguistics, pp. 10–18 (2012)
34. Newman, M.E.: Scientific collaboration networks. i. network construction and fundamental results. Phys. Rev. E **64**(1) (2001). 016131
35. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech: Theor. Exp. **2009**(03) (2009). P03024
36. Paisley, J., Blei, D., Jordan, M.I.: Bayesian nonnegative matrix factorization with stochastic variational inference. In: Handbook of Mixed Membership Models and Their Applications. Chapman and Hall/CRC, Boca Raton (2014)
37. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recogn. **37**(3), 487–501 (2004)
38. Palla, G., Barabási, A.L., Vicsek, T.: Quantifying social group evolution. Nature **446**(7136), 664–667 (2007)
39. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)

40. Psorakis, I., Roberts, S., Ebden, M., Sheldon, B.: Overlapping community detection using bayesian non-negative matrix factorization. Phys. Rev. E **83**(6). 066114 (2011)
41. Rabbany, R., Zaïane, O.R.: Generalization of clustering agreements and distances for overlapping clusters and network communities. Data Min. Knowl. Disc. **29**(5), 1458–1485 (2015)
42. Schmidt, M.N., Laurberg, H.: Nonnegative matrix factorization with Gaussian process priors. Comput. Intell. Neurosci. **2008**, 3 (2008)
43. Shankar, D.S., Bhavani, S.D.: Consensus clustering approach for discovering overlapping nodes in social networks. In: Proceedings of the 3rd IKDD Conference on Data Science, p. 21. ACM (2016)
44. Shi, X., Lu, H., He, Y., He, S.: 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, ASONAM 2015, pp. 541–546. ACM, New York (2015)
45. Tan, V.Y.F., Fevotte, C.: Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1592–1605 (2013)
46. Wang, W., Jiao, P., He, D., Jin, D., Pan, L., Gabrys, B.: Autonomous overlapping community detection in temporal networks: a dynamic bayesian nonnegative matrix factorization approach. Knowl.-Based Syst. **110**, 121–134 (2016)
47. Whang, J.J., Gleich, D.F., Dhillon, I.S.: Overlapping community detection using seed set expansion. In: Proceedings of the 22nd ACM International Conference on Conference on Information Knowledge Management - CIKM 2013. Association for Computing Machinery (ACM) (2013)
48. Wu, P., Fu, Q., Tang, F.: Social community detection from photo collections using Bayesian overlapping subspace clustering. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011. LNCS, vol. 6524, pp. 57–64. Springer, Heidelberg (2011). doi:10.1007/978-3-642-17829-0_6
49. Wu, Z.H., Lin, Y.F., Gregory, S., Wan, H.Y., Tian, S.F.: Balanced multi-label propagation for overlapping community detection in social networks. J. Comput. Sci. Technol. **27**(3), 468–479 (2012)
50. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput. Surv. **45**(4), 43:1–43:35 (2013)
51. Xie, J., Szymanski, B.K.: Towards linear time overlapping community detection in social networks. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012. LNCS (LNAI), vol. 7302, pp. 25–36. Springer, Heidelberg (2012). doi:10.1007/978-3-642-30220-6_3
52. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 587–596. ACM (2013)
53. Zhang, H., King, I., Lyu, M.R.: Incorporating implicit link preference into overlapping community detection. In: AAAI, pp. 396–402 (2015)
54. Zhang, Y., Yeung, D.Y.: Overlapping community detection via bounded nonnegative matrix tri-factorization. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2012. Association for Computing Machinery (ACM) (2012)
55. Zhubing, L., Jian, W., Yuzhou, L.: An overview on overlapping community detection. In: 2012 7th International Conference on Computer Science and Education (ICCSE), pp. 486–490. IEEE (2012)