Bayesian Analysis (2020) **15**, Number 3, pp. 711–736

Mixed Membership Stochastic Blockmodels for Heterogeneous Networks

Weihong Huang[*§], Yan Liu[†§], and Yuguo Chen[‡]

Abstract. Heterogeneous networks are useful for modeling complex systems that consist of different types of objects. However, there are limited statistical models to deal with heterogeneous networks. In this paper, we propose a statistical model for community detection in heterogeneous networks. We formulate a heterogeneous version of the mixed membership stochastic blockmodel to accommodate heterogeneity in the data and the content dependent property of the pairwise relationship. We also apply a variational algorithm for posterior inference. The proposed procedure is shown to be consistent for community detection under mixed membership stochastic blockmodels for heterogeneous networks. We demonstrate the advantage of the proposed method in modeling overlapping communities and multiple memberships through simulation studies and applications to a real data set.

Keywords: clustering, community detection, heterogeneous network, mixed membership model, stochastic blockmodel, variational algorithm.

1 Introduction

In recent years, network data have drawn a lot of attention from researchers in many areas, including statistics, computer science, biology, and economics. Network data appear in diverse applications such as social networks, protein-protein interaction (PPI) networks, the World Wide Web, and research publication networks. Modeling network data is an important topic, and Goldenberg et al. (2010) provided a review of statistical network models.

Many networks show the pattern of communities. That is, objects belonging to the same community tend to have similar behavior while objects belonging to different communities behave differently. One of the interesting problems in network analysis is clustering, or community detection, which is the process of uncovering the underlying community structure. The detected communities can also be meaningful in real applications. For example, the detected communities may correspond to functional groups (or proteins participating in the same cellular processes) associated with cancer and metastasis (Jonsson et al., 2006).

[*]Facebook, Menlo Park, CA 94025, whuang42@fb.com

[†]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, yanl5@illinois.edu

[‡]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, yuguo@illinois.edu

[§]Weihong Huang and Yan Liu made equal contributions to this paper and should be considered as joint first authors.

 https://doi.org/10.1214/19-BA1163

Some statistical methods for network clustering have been proposed in the literature. Hoff et al. (2002) proposed the latent space model, which was later extended by Handcock et al. (2007) for clustering, assuming that the latent positions come from a mixture of Gaussians. Nowicki and Snijders (2001) proposed the stochastic blockmodel, in which each object belongs to a cluster and the relationship between two objects depends on the pair of clusters these two objects belong to. One limitation of this model is that it assumes each object belongs to a single cluster and cannot handle overlapping communities. To model the situations that violate the single cluster assumption, Airoldi et al. (2008) proposed the mixed membership stochastic blockmodel (MMSB). In this model, the cluster of each object is content dependent, which means objects can show different functional contexts (or clusters) when interacting with different objects.

The methods mentioned above assume the network is homogeneous, i.e., all nodes in the network are objects of the same type, such as people in the social network and proteins in the PPI network. However in real world, objects of different types interact with each other to form a large heterogeneous network. For example, a university network consists of several types of objects (such as students, professors, courses and departments) and different types of links among them (such as the teaching relationship between professors and students, the registration relationship between students and courses, and the association relationship between students/professors and departments).

A heterogeneous network carries more information than its homogeneous sub-network. For example, a heterogeneous bibliographical network consists of authors, papers, and conferences as different types of nodes, and different types of relationships among them as edges. The homogeneous co-authorship network can be viewed as a projection of the heterogeneous network. Analyzing the co-authorship network only will result in an information loss, since the paper and conference nodes and the author-paper and author-conference links are ignored. Therefore, it is necessary to develop new methods to make use of the rich information in heterogeneous networks. Sun and Han (2012) provided an overview of the methods for mining heterogeneous networks in the computer science community. Sengupta and Chen (2015) proposed the spectral clustering method for the heterogeneous version of the stochastic blockmodel. However, similar to the homogeneous stochastic blockmodel, the heterogeneous version still assumes that each node belongs to a single cluster.

In this paper, we propose a heterogeneous version of the mixed membership stochastic blockmodel for community detection in heterogeneous networks. Similar to the homogeneous MMSB, each object is allowed to have multiple clusters and the clusters are content dependent. We present a variational EM algorithm for posterior inference so that it can scale up to large networks. The proposed procedure is shown to be consistent for community detection under mixed membership stochastic blockmodels for heterogeneous networks. We also apply our method to analyze a subset of the DBLP dataset to find out the community structure for authors.

The paper is organized as follows. Section 2 gives a review of the homogeneous MMSB and introduces the heterogeneous version of the MMSB. Section 3 describes the variational algorithm for posterior inference. Section 4 shows consistency of community detection under mixed membership stochastic blockmodels for heterogeneous networks.

Section 5 presents simulation studies comparing our method with the spectral clustering method of Sengupta and Chen (2015). Section 6 shows the results of our method applied to the DBLP dataset. Section 7 concludes with a discussion.

2 The mixed membership stochastic blockmodel

2.1 Homogeneous model

A homogeneous network or relational data can be represented as a graph $G(V, E)$, where V consists of all nodes (or vertices) and E consists of all links (or edges). Here we only consider unweighted graphs, but the edges can be directed or undirected. Suppose there are n nodes in the graph, denoted by $x_1, \ldots, x_n$. An adjacency matrix Y for this graph is an n-by-n binary matrix, where $Y(p, q) = 1$ if node x_p and node x_q are connected and $Y(p, q) = 0$ otherwise. For a homogeneous network, all of the nodes in V are of the same type, such as people in a friendship network, papers in a citation network, or proteins in a PPI network.

The original mixed membership stochastic blockmodel (Airoldi et al., 2008) considers a homogeneous network $G(V, E)$ and its adjacency matrix Y. Assume there are K groups. The MMSB models the adjacency matrix Y in a Bayesian hierarchical framework. For each pair of nodes (x_p, x_q), the presence or absence of a link between them is determined by a Bernoulli distribution with parameter depending on the latent group memberships of the two nodes. In other words, given the latent group membership $\boldsymbol{z}_{p,q,1}$, $\boldsymbol{z}_{p,q,2}$ and the Bernoulli probability matrix B,

$$Y(p, q)|\boldsymbol{z}_{p,q,1}, \boldsymbol{z}_{p,q,2}, B \sim \text{Bernoulli}(\boldsymbol{z}_{p,q,1}^T B \boldsymbol{z}_{p,q,2}).$$

The Bernoulli probability matrix B is K-by-K, where $B(g, h)$ represents the probability of having a link between a node in group g and a node in group h. Here $\boldsymbol{z}_{p,q,1}$ and $\boldsymbol{z}_{p,q,2}$ are K-dimensional membership indicator vectors, of which only one element equals to one and others equal to zero. The index of the non-zero element corresponds to the membership of the node. For undirected graphs, $\boldsymbol{z}_{p,q,1}$ denotes the latent group membership of node x_p when interacting with node x_q, and $\boldsymbol{z}_{p,q,2}$ denotes the latent group membership of node x_q when interacting with node x_p. For directed graphs, $\boldsymbol{z}_{p,q,1}$ and $\boldsymbol{z}_{p,q,2}$ denote the group membership of the initiator and receiver of the edge between x_p and x_q respectively. Note that the group membership of each node depends on the nodes it is interacting. That is, each node can have different membership when interacting or being interacted with different nodes. For example, a researcher may work as a biologist on a project about mass spectrometry analysis for proteins with other biologists. He/She may also work as a statistician on a project about network analysis with his/her students.

For the rest of the paper, we will focus on the undirected graph. For each node x_p, the latent group membership $\boldsymbol{z}_{p,\cdot,1} := \{\boldsymbol{z}_{p,q,1} : x_q \in V\}$ and $\boldsymbol{z}_{\cdot,p,2} := \{\boldsymbol{z}_{q,p,2} : x_q \in V\}$ have prior distribution with parameter $\boldsymbol{\pi}_p$,

$$\boldsymbol{z}_{p,\cdot,1}|\boldsymbol{\pi}_p \sim \text{Multinomial}_K(\boldsymbol{\pi}_p),$$

$$\boldsymbol{z}_{\cdot,p,2}|\boldsymbol{\pi}_p \sim \text{Multinomial}_K(\boldsymbol{\pi}_p),$$

and $\boldsymbol{\pi}_p$ has prior distribution

$$\boldsymbol{\pi}_p \sim \text{Dirichlet}_K(\boldsymbol{\alpha}),$$

where $\boldsymbol{\pi}_p$ and $\boldsymbol{\alpha}$ are K-dimensional vectors.

Let $Z_1 := \{\boldsymbol{z}_{p,q,1} : x_p, x_q \in V\}$, $Z_2 := \{\boldsymbol{z}_{p,q,2} : x_p, x_q \in V\}$ and $\boldsymbol{\pi} := \{\boldsymbol{\pi}_p : x_p \in V\}$. Then the joint distribution of data Y and the parameters $\{Z_1, Z_2, \boldsymbol{\pi}\}$ is

$$\begin{aligned} &p(Y, Z_1, Z_2, \boldsymbol{\pi}|\boldsymbol{\alpha}, B) \\ =&\prod_{p,q} p_1(Y(p,q)|\boldsymbol{z}_{p,q,1}, \boldsymbol{z}_{p,q,2}, B) p_2(\boldsymbol{z}_{p,q,1}|\boldsymbol{\pi}_p) p_2(\boldsymbol{z}_{p,q,2}|\boldsymbol{\pi}_q) \prod_p p_3(\boldsymbol{\pi}_p|\boldsymbol{\alpha}), \end{aligned}$$

where p_1, p_2, p_3 are the probability distributions of Bernoulli, multinomial and Dirichlet distributions, respectively.

2.2 Heterogeneous model

Different from the homogeneous network, the nodes in a heterogeneous network are of different types. Therefore the links in the heterogeneous network are also of different types. For example, in the Facebook network, other than people, we have object types such as posts, photos, movies, and events. Also, besides the friendship relation between people, there are relationships of other types, such as the person-photo tagging relationship, person-movie liking relationship, and person-post publishing relationship. To accommodate different types of nodes and links, we propose a mixed membership stochastic blockmodel for heterogeneous networks.

Given a heterogeneous network $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains all types of nodes and $\mathcal{E}$ contains all types of links. The graph is unweighted but can be directed or undirected. In this paper we focus on the undirected graph. Suppose there are N nodes of m different types, denoted by $\mathcal{X}_1 = \{x_{11}, \ldots, x_{1n_1}\}, \ldots, \mathcal{X}_m = \{x_{m1}, \ldots, x_{mn_m}\}$. Then $\mathcal{V} = \bigcup_{i=1}^m \mathcal{X}_i$ and $N = n_1 + \cdots + n_m$. Let G_{ij} be the subgraph between object types $\mathcal{X}_i$ and $\mathcal{X}_j$, and Y_{ij} be the adjacency matrix of G_{ij}, $1 \le i, j \le m$. Let Y be the following N-by-N matrix

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \ldots & Y_{1m} \\ Y_{21} & Y_{22} & \ldots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{m1} & Y_{m2} & \ldots & Y_{mm} \end{bmatrix}.$$

Suppose there are K groups. For any node x_{ip} from $\mathcal{X}_i$ and any node x_{jq} from $\mathcal{X}_j$, the probability that there is a link between the pair of nodes (x_{ip}, x_{jq}) is determined by a Bernoulli distribution with parameter depending on the group memberships of nodes x_{ip} and x_{jq}. Similar to the homogeneous MMSB, the group membership for each node depends on the nodes they interact with. The latent group membership of node x_{ip} when

interacting or being interacted with others is determined by a multinomial distribution with a node-specific parameter $\boldsymbol{\pi}_{ip}$. A Dirichlet prior is put on $\boldsymbol{\pi}_{ip}$, governed by a type-specific hyperparameter $\boldsymbol{\alpha}_i$. Therefore we have the following Bayesian hierarchical model:

$$\begin{aligned} Y_{ij}(p,q)|\boldsymbol{z}_{ip,jq,1},\boldsymbol{z}_{ip,jq,2},B \sim & \text{Bernoulli}(\boldsymbol{z}_{ip,jq,1}^T B_{ij}\boldsymbol{z}_{ip,jq,2}),\\ \boldsymbol{z}_{ip,jq,1}|\boldsymbol{\pi}_{ip} \sim & \text{Multinomial}_K(\boldsymbol{\pi}_{ip}),\\ \boldsymbol{z}_{ip,jq,2}|\boldsymbol{\pi}_{jq} \sim & \text{Multinomial}_K(\boldsymbol{\pi}_{jq}),\\ \boldsymbol{\pi}_{ip} \sim & \text{Dirichlet}_K(\boldsymbol{\alpha}_i), \end{aligned}$$

where $\boldsymbol{z}_{ip,jq,1}$ and $\boldsymbol{z}_{ip,jq,2}$ are K-dimensional membership indicator vectors. Here $\boldsymbol{z}_{ip,jq,1}$ denotes the latent group membership of node x_{ip} (or the initiator for directed graph) when interacting with node x_{jq}, and $\boldsymbol{z}_{ip,jq,2}$ denotes the latent group membership of node x_{jq} (or the receiver for directed graph) when interacting with node x_{ip}. The Bernoulli probability matrix B is mK-by-mK with the following structure:

$$B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1m} \\ B_{21} & B_{22} & \dots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \dots & B_{mm} \end{bmatrix},$$

where B_{st} is a K-by-K matrix whose (g,h) entry denotes the probability of having a link between a node of object type $\mathcal{X}_s$ from group g and a node of object type $\mathcal{X}_t$ from group h. The link probability depends on not only the group memberships (g and h), but also the types of node (s and t).

Let $Z_1 := \{\boldsymbol{z}_{ip,jq,1} : x_{ip}, x_{jq} \in \mathcal{V}\}$, $Z_2 := \{\boldsymbol{z}_{ip,jq,2} : x_{ip}, x_{jq} \in \mathcal{V}\}$, $\boldsymbol{\pi} := \{\boldsymbol{\pi}_{ip} : x_{ip} \in \mathcal{V}\}$, and $\boldsymbol{\alpha} := \{\boldsymbol{\alpha}_i : i = 1, 2, \dots, m\}$. Then the joint distribution of data Y and the parameters $\{Z_1, Z_2, \boldsymbol{\pi}\}$ is

$$\begin{aligned} & p(Y, Z_1, Z_2, \boldsymbol{\pi}|\boldsymbol{\alpha}, B) \\ = & \prod_{i,j}\prod_{p,q} p_1(Y_{ij}(p,q)|\boldsymbol{z}_{ip,jq,1},\boldsymbol{z}_{ip,jq,2},B) p_2(\boldsymbol{z}_{ip,jq,1}|\boldsymbol{\pi}_{ip}) p_2(\boldsymbol{z}_{ip,jq,2}|\boldsymbol{\pi}_{jq}) \prod_{i,p} p_3(\boldsymbol{\pi}_{ip}|\boldsymbol{\alpha}_i), \end{aligned}$$

where p_1, p_2, p_3 are the probability distributions of Bernoulli, multinomial and Dirichlet distributions, respectively.

3 Posterior inference and parameter estimation

We are interested in finding the posterior distribution of the latent variables, including the per-node mixed membership $\boldsymbol{\pi}$ and the membership indicators for per-pair interaction Z_1, Z_2, given the observed network. We also want to learn the Bernoulli probability matrix B. The hyperparameter $\boldsymbol{\alpha}$ is pre-specified. In this section, we present the variational method for posterior inference and parameter estimation.

3.1 Variational posterior inference

Let X be the collection of the latent variables $X = \{\boldsymbol{\pi}, Z_1, Z_2\}$, then the posterior distribution of X given data Y and hyperparameters $\Theta = \{\boldsymbol{\alpha}, B\}$ can be written as

$$p(X|Y,\Theta) = \frac{p(Y,X|\Theta)}{p(Y|\Theta)}.$$

One way to make inference on the posterior distribution is to use Markov chain Monte Carlo (MCMC) sampling. However, MCMC can be slow when the network size is large, so it is difficult to handle large networks. In addition, estimating the Bernoulli probability matrix B requires evaluating the normalizing constant

$$p(Y|\Theta) = \int_{\boldsymbol{\pi}} \sum_{Z_1,Z_2} \Bigg(\prod_{i,j,p,q} p_1(Y_{ij}(p,q)|\boldsymbol{z}_{ip,jq,1}, \boldsymbol{z}_{ip,jq,2}, B) \cdot p_2(\boldsymbol{z}_{ip,jq,1}|\boldsymbol{\pi}_{ip}) p_2(\boldsymbol{z}_{ip,jq,2}|\boldsymbol{\pi}_{jq}) \prod_{i,p} p_3(\boldsymbol{\pi}_{ip}|\boldsymbol{\alpha}_i) \Bigg) d\boldsymbol{\pi},$$

which requires integration over the latent variables and is not easy to compute.

Similar to the original algorithm for homogeneous MMSB, we use the variational method (Wainwright and Jordan, 2008) for posterior inference and parameter estimation. The main idea of variational method is to approximate the true posterior distribution by a variational distribution with free parameters (also called variational parameters). Then the free parameters are fitted to minimize the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution.

We introduce a fully factorized distribution q_Δ as variational distribution to approximate the true posterior distribution $p(X|Y,\Theta)$. The variational distribution is defined as

$$q_\Delta = q(X|\boldsymbol{\gamma}, \Phi_1, \Phi_2) = \prod_{i,p} q_1(\boldsymbol{\pi}_{ip}|\boldsymbol{\gamma}_{ip}) \prod_{i,j} \prod_{p,q} \left[q_2(\boldsymbol{z}_{ip,jq,1}|\boldsymbol{\phi}_{ip,jq,1}) q_2(\boldsymbol{z}_{ip,jq,2}|\boldsymbol{\phi}_{ip,jq,2}) \right],$$

where q_1 is the Dirichlet distribution, q_2 is the multinomial distribution, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_{ip} : x_{ip} \in \mathcal{V}\}$, $\Phi_1 = \{\boldsymbol{\phi}_{ip,jq,1} : x_{ip}, x_{jq} \in \mathcal{V}\}$ and $\Phi_2 = \{\boldsymbol{\phi}_{ip,jq,2} : x_{ip}, x_{jq} \in \mathcal{V}\}$. Let $\Delta = (\boldsymbol{\gamma}, \Phi_1, \Phi_2)$ be the variational parameters in q_Δ.

By Jensen's inequality, we have that minimizing the KL divergence between the variational distribution q_Δ and the true posterior distribution is equivalent to maximizing the lower bound

$$L(q_\Delta, \Theta) = E_{q_\Delta}[\log p(Y,X|\Theta) - \log q_\Delta(X)].$$

The details are given in the supplementary material (Huang et al., 2019).

3.2 Variational EM algorithm

To find the variational approximation of the posterior distribution of the latent variables and obtain the estimate of Bernoulli probability matrix B, we maximize $L(q_\Delta, \Theta)$

iteratively with respect to the variational parameters $\Delta = (\boldsymbol{\gamma}, \Phi_1, \Phi_2)$ and the hyperparameter B. The details of the derivation are in the supplementary material (Huang et al., 2019). By using the coordinate ascent algorithm, we can get the updating equation for variational multinomial parameters:

$$\begin{aligned}\hat{\phi}_{ip,jq,1,g} &\propto \exp\{\sum_h [\phi_{ip,jq,2,h} f(Y_{ij}(p,q), B_{ij}(g,h))] + \psi(\gamma_{ip,g}) - \psi(\sum_g \gamma_{ip,g})\} \qquad (3.1)\\ &= \exp\{\psi(\gamma_{ip,g}) - \psi(\sum_g \gamma_{ip,g})\}\\ &\quad \times \prod_h \left[B_{ij}(g,h)^{Y_{ij}(p,q)}(1 - B_{ij}(g,h))^{1-Y_{ij}(p,q)}\right]^{\phi_{ip,jq,2,h}},\end{aligned}$$

$$\begin{aligned}\hat{\phi}_{ip,jq,2,h} &\propto \exp\{\sum_g [\phi_{ip,jq,1,g} f(Y_{ij}(p,q), B_{ij}(g,h))] + \psi(\gamma_{jq,h}) - \psi(\sum_h \gamma_{jq,h})\} \qquad (3.2)\\ &= \exp\{\psi(\gamma_{jq,h}) - \psi(\sum_h \gamma_{jq,h})\}\\ &\quad \times \prod_g \left[B_{ij}(g,h)^{Y_{ij}(p,q)}(1 - B_{ij}(g,h))^{1-Y_{ij}(p,q)}\right]^{\phi_{ip,jq,1,g}}.\end{aligned}$$

Since $\boldsymbol{\phi}_{ip,jq,1}$ and $\boldsymbol{\phi}_{ip,jq,2}$ are probability vectors, they need to be normalized to make sure that $\sum_g \hat{\phi}_{ip,jq,1,g} = \sum_h \hat{\phi}_{ip,jq,2,h} = 1$. The updating equation for variational Dirichlet parameter $\hat{\gamma}_{ip,g}$ is

$$\hat{\gamma}_{ip,g} = \alpha_{i,g} + \sum_{j,q} \phi_{ip,jq,1,g} + \sum_{j,q} \phi_{jq,ip,2,g}. \qquad (3.3)$$

To get the estimate for the Bernoulli probability matrix B, we fix the variational parameters Δ to obtain the estimate of B that maximizes the lower bound $L(q_\Delta, \Theta)$. Therefore we get the update for $\hat{B}$:

$$\hat{B}_{ij}(g,h) = \frac{\sum_{p,q} \phi_{ip,jq,1,g}\phi_{ip,jq,2,h} Y_{ij}(p,q)}{\sum_{p,q} \phi_{ip,jq,1,g}\phi_{ip,jq,2,h}}, \qquad (3.4)$$

where $i, j = 1, \ldots, m$ and $g, h = 1, \ldots, K$.

In the proposed variational EM algorithm, we iteratively update the variational parameters $\Delta = (\boldsymbol{\gamma}, \Phi_1, \Phi_2)$ in E step, and the Bernoulli probability matrix B in M step until convergence. We use the value of the lower bound $L(q_\Delta, \Theta)$ to determine the convergence, i.e., the convergence is achieved when $L(q_{\Delta^{(t+1)}}, \Theta^{(t+1)}) - L(q_{\Delta^{(t)}}, \Theta^{(t)}) < \epsilon$, where $\epsilon > 0$ is the tolerance. The overall algorithm is summarized in Algorithm 1.

3.3 Initialization

As stated in Algorithm 1, the initial values of Φ_1 and Φ_2 need to be given first. Although the variational EM algorithm is fast and feasible to handle large networks, similar to the EM algorithm, it typically converges to a local maximum, not necessary the global

Algorithm 1 Variational EM Algorithm for MMSB for Heterogeneous Networks.

Initialize $\hat{\phi}^0_{ip,jq,1,g}$ and $\hat{\phi}^0_{ip,jq,2,h}$ for all pairs of nodes (x_{ip}, x_{jq}) and all pairs of groups g, h.
Initialize $\hat{\gamma}^0_{ip,g}$ for all nodes x_{ip} and all groups g by Equation (3.3).
Initialize $B^0, \boldsymbol{\alpha}$.
$t = 0$
repeat
  for $i = 1$ to m **do**
    for $p = 1$ to n_i **do**
      for $j = 1$ to m **do**
        for $q = 1$ to n_j **do**
          for $g = 1$ to K **do**
            update $\hat{\phi}^{t+1}_{ip,jq,1,g}$ by Equation (3.1)
          end for
          normalize $\{\hat{\phi}^{t+1}_{ip,jq,1,g}\}_{g=1}^K$ to sum to 1
          for $h = 1$ to K **do**
            update $\hat{\phi}^{t+1}_{ip,jq,2,h}$ by Equation (3.2)
          end for
          normalize $\{\hat{\phi}^{t+1}_{ip,jq,2,h}\}_{h=1}^K$ to sum to 1
        end for
      end for
    end for
  end for
  for $i = 1$ to m **do**
    for $p = 1$ to n_i **do**
      update $\hat{\gamma}^{t+1}_{ip,g}$ by Equation (3.3)
    end for
  end for
  for $i = 1$ to m **do**
    for $j = 1$ to m **do**
      update $B^{t+1}_{ij}(g, h)$ by Equation (3.4)
    end for
  end for
  $t = t + 1$
until convergence

maximum. Therefore, the results of the variational EM are sensitive to the initial values. It is helpful to use multiple initial values, and take the one with maximal value of $L(q_\Delta, \Theta)$ as the final output. However it is hard to determine the number of initial values needed and there is no guarantee that the global maximum can be found. In practice, using multiple initial values does not show much improvement and takes longer time to run the algorithm.

Another way to deal with this problem is to use the results from some pre-analysis as the initialization. Sengupta and Chen (2015) proposed a heterogeneous spectral cluster-

ing algorithm (Het-SC) for community detection in heterogeneous networks. Although the Het-SC algorithm assigns unique membership to each node, its results can still serve as a good guidance of the initial values. Since the result of Het-SC itself can be a local maximum or very close to a local maximum, the algorithm may get stuck in the local mode when starting from the results of Het-SC. A remedy is to use initial values close to the results of Het-SC to help the algorithm escape from the local mode.

4 Theoretical properties

In this section, we study the consistency of community detection under mixed membership stochastic blockmodels for heterogeneous networks. Although the consistency of community detection has been studies extensively in the literature, most of the existing work assumes that each node has a unique membership and the network is homogeneous. In the following, we first give a definition of community detection consistency in our setting.

Consider a heterogeneous network $G(\mathcal{V}, \mathcal{E})$ ($\mathcal{V} = \bigcup_{i=1}^{m} \mathcal{X}_i$) with community labels

$$\mathcal{Z} := \{z_{ip,jq,1}, z_{ip,jq,2} \in \{1, \ldots, K\} : \quad x_{ip}, x_{jq} \in \mathcal{V}\}.$$

Here $z_{ip,jq,1}$ denotes the community that node x_{ip} belongs to when it interacts with node x_{jq}, and $z_{ip,jq,2}$ denotes the community that node x_{jq} belongs to when it interacts with node x_{ip}.

A community detection criterion is generally a function of community labels $\mathcal{Z}$ and the observed network G. We define a community detection criterion $F(\mathcal{Z}, G)$ to be consistent if

$$\hat{\mathcal{Z}} := \arg\max_{\mathcal{Z}} F(\mathcal{Z}, G)$$

satisfies $\forall \epsilon > 0$,

$$P\Bigg[\frac{1}{n^2} \sum_{\substack{k \neq u \\ \text{or } l \neq v}} \Bigg(\sum_{i \neq j} \sum_{p,q} I(\hat{z}_{ip,jq,1} = u, \hat{z}_{ip,jq,2} = v) I(z_{ip,jq,1} = k, z_{ip,jq,2} = l) + \sum_{i} \sum_{p,q} I(\hat{z}_{ip,iq,1} = u, \hat{z}_{ip,iq,2} = v) I(z_{ip,iq,1} = k, z_{ip,iq,2} = l)\Bigg) > \epsilon\Bigg] \to 0 \tag{4.1}$$

as $n \to \infty$. This consistency definition is a generalization of the one proposed in Zhang and Chen (2019) for heterogeneous networks with unique membership for each node. In the unique membership case, consistency requires that the proportion of misclassified nodes goes to zero. For the mixed membership case, since each node is allowed to have different memberships when interacting with different nodes, we consider the proportion of misclassified pairs of nodes in the definition (4.1).

In order to study the consistency property, we first define the following key quantities. We use $\boldsymbol{z}$ to denote the ground truth of the community labels, and use $\boldsymbol{e}$ as a generic notation of a set of label assignment.

For any set of label assignment $\boldsymbol{e}$, define K-by-K matrices $O^{[i]}(\boldsymbol{e})$, $i = 1, \ldots, m$, and $O^{[ij]}(\boldsymbol{e})$, $1 \leq i \neq j \leq m$, as

$$\begin{aligned} O^{[i]}_{kl}(\boldsymbol{e}) &:= \sum_{p,q} Y_{ii}(p,q) I(e_{ip,iq,1} = k, e_{ip,iq,2} = l), \\ O^{[ij]}_{kl}(\boldsymbol{e}) &:= \sum_{p,q} Y_{ij}(p,q) I(e_{ip,jq,1} = k, e_{ip,jq,2} = l), \end{aligned}$$

where $I(\cdot)$ is the indicator function. Here $O^{[i]}_{kl}$ is the total number of edges between nodes in $\mathcal{X}_i$ with community label k and nodes in $\mathcal{X}_i$ with community label l, and $O^{[ij]}_{kl}$ is the total number of edges between nodes in $\mathcal{X}_i$ with community label k and nodes in $\mathcal{X}_j$ with community label l. Let

$$O(\boldsymbol{e}) = \sum_{i \neq j} O^{[ij]}(\boldsymbol{e}) + \sum_{i=1}^{m} O^{[i]}(\boldsymbol{e}).$$

Also, define K-by-K matrices $n^{[i]}(\boldsymbol{e})$, $i = 1, \ldots, m$, and $n^{[ij]}(\boldsymbol{e})$, $1 \leq i \neq j \leq m$, as

$$\begin{aligned} n^{[i]}_{kl}(\boldsymbol{e}) &:= \sum_{p,q} I(e_{ip,iq,1} = k, e_{ip,iq,2} = l), \\ n^{[ij]}_{kl}(\boldsymbol{e}) &:= \sum_{p,q} I(e_{ip,jq,1} = k, e_{ip,jq,2} = l). \end{aligned}$$

Define K-by-K matrices $f^{[i]}(\boldsymbol{e})$, $i = 1, \ldots, m$, $f^{[ij]}(\boldsymbol{e})$, $1 \leq i \neq j \leq m$, and $f(\boldsymbol{e})$ as

$$f^{[i]}_{kl}(\boldsymbol{e}) := \frac{n^{[i]}_{kl}(\boldsymbol{e})}{n^2}, \quad f^{[ij]}_{kl}(\boldsymbol{e}) := \frac{n^{[ij]}_{kl}(\boldsymbol{e})}{n^2},$$

$$f_{kl}(\boldsymbol{e}) := \frac{n_{kl}(\boldsymbol{e})}{n^2} := \frac{1}{n^2} \left(\sum_{i \neq j} n^{[ij]}_{kl}(\boldsymbol{e}) + \sum_{i=1}^{m} n^{[i]}_{kl}(\boldsymbol{e}) \right).$$

For $k, l = 1, \ldots, K$, $n_{kl}(\boldsymbol{e})$ is the number of node pairs that belong to community k and community l respectively when they interact with each other. Henceforth, we will refer to such node pairs as $[kl]$-node pairs. Then f_{kl} is the proportion of $[kl]$-node pairs in the observed network.

Furthermore, we define the following quantities that characterize the discrepancy between two community assignments $\boldsymbol{e}$ and $\boldsymbol{z}$ for the interactions between type i nodes and type j nodes ($1 \leq i \neq j \leq m$):

$$R^{[ij]}_{kluv}(\boldsymbol{e}, \boldsymbol{z}) := \frac{1}{n^2} \sum_{p,q} I(e_{ip,jq,1} = k, e_{ip,jq,2} = l) I(z_{ip,jq,1} = u, z_{ip,jq,2} = v),$$

and the quantities that characterize the discrepancy between $\boldsymbol{e}$ and $\boldsymbol{z}$ for the interactions within type i nodes ($1 \leq i \leq m$):

$$R^{[i]}_{kluv}(\boldsymbol{e}, \boldsymbol{z}) := \frac{1}{n^2} \sum_{p,q} I(e_{ip,iq,1} = k, e_{ip,iq,2} = l) I(z_{ip,iq,1} = u, z_{ip,iq,2} = v).$$

As discussed in (Zhao et al., 2012), a large class of community detection criteria can be expressed as

$$Q(\boldsymbol{e}) = F\left(\frac{O(\boldsymbol{e})}{\mu_n}, f(\boldsymbol{e})\right),$$

where $\mu_n = n^2\rho_n$, and $\rho_n = \sum_{i,j}\sum_{k,l}\pi_k\pi_l B_{ij}(k,l)$ is the probability of there being an edge between two nodes. The community membership estimator based on maximum likelihood (which will be discussed in the proof of Corollary 4.1) can be written in this form. For this type of community detection criteria, a natural condition for consistency result is that the "population version" of Q should be maximized by the correct community assignment (Zhao et al., 2012). In order to define the population version of Q, we define the population version of $O(\boldsymbol{e})$ and $f(\boldsymbol{e})$ as functions of the discrepancy characterization $\boldsymbol{R}$.

Let $S_{uv}^{[ij]} := \frac{1}{\rho_n}P(Y_{ij}(1,2)|z_{i1,j2,1} = u, z_{i1,j2,2} = v)$. For $1 \le i \ne j \le m$ and $k,l = 1,\ldots,K$, the population version of $O_{kl}^{[ij]}(\boldsymbol{e})$ is defined as its conditional expectation given the true label assignment $\boldsymbol{z}$:

$$\begin{aligned}
\hat{T}_{kl}^{[ij]} &:= \frac{1}{\mu_n}\mathbb{E}[O_{kl}^{[ij]}(\boldsymbol{e})|\boldsymbol{z}] \\
&= \frac{1}{\mu_n}\mathbb{E}\left[\sum_{p,q} Y_{ij}(p,q)I(e_{ip,jq,1} = k, e_{ip,jq,2} = l)|\boldsymbol{z}\right] \\
&= \frac{1}{\mu_n}\mathbb{E}\left[\sum_{p,q}\sum_{u,v} Y_{ij}(p,q)I(e_{ip,jq,1} = k, e_{ip,jq,2} = l)I(z_{ip,jq,1} = u, z_{ip,jq,2} = v)|\boldsymbol{z}\right] \\
&= \frac{1}{\mu_n}\sum_{u,v}\mathbb{E}\left[\sum_{p,q} Y_{ij}(p,q)I(e_{ip,jq,1} = k, z_{ip,jq,2} = u)I(e_{ip,jq,1} = l, z_{ip,jq,2} = v)|\boldsymbol{z}\right] \\
&= \sum_{u,v} S_{uv}^{[ij]}R_{kluv}^{[ij]} := H_{kl}(R^{[ij]}),
\end{aligned}$$

and the population version of $f_{kl}^{[ij]}(\boldsymbol{e})$ is defined by $\sum_{u,v} R_{kluv}^{[ij]}(\boldsymbol{e},\boldsymbol{z}) := h_{kl}(R^{[ij]})$.

Similarly, for $1 \le i \le m$, the population version of $O_{kl}^{[i]}(\boldsymbol{e})$ is defined by

$$\hat{T}_{kl}^{[i]} := \frac{1}{\mu_n}\mathbb{E}[O_{kl}^{[i]}(\boldsymbol{e})|\boldsymbol{z}] = \sum_{u,v} S_{uv}^{[i]}R_{kluv}^{[i]} := H_{kl}(R^{[i]}),$$

and the population version of $f_{kl}^{[i]}(\boldsymbol{e})$ is defined by $\sum_{u,v} R_{kluv}^{[i]}(\boldsymbol{e},\boldsymbol{z}) := h_{kl}(R^{[i]})$.

Now we state the assumptions we need for the consistency result.

Assumption 4.1. *Among the n nodes, only $\lfloor\sqrt{n}\rfloor$ of them have multiple memberships. Other nodes have unique membership.*

Assumption 4.2. *For each $i = 1,\ldots,m$, the multinomial parameter $\boldsymbol{\pi}_{ip}$ is the same for each $p = 1,\ldots,n_i$. (However, the multinomial parameter is allowed to take different values for nodes of different types.)*

Assumption 4.3. *The proportion of each type of nodes is stable. In other words, there exists* $\zeta_i \in (0,1)$, $i = 1, \ldots, m$, *such that* $\sum_{i=1}^m \zeta_i = 1$ *and for each* $i = 1, \ldots, m$,

$$\frac{n_i}{n} \to \zeta_i, \qquad \text{as } n \to \infty.$$

Assumption 4.4. *As* $n \to \infty$, *the edge density* ρ_n *is either fixed, or goes to 0 at a rate such that* $n^{-1/4} = o(\rho_n)$, *i.e.,* $\frac{n^{-1/4}}{\rho_n} \to 0$ *as* $n \to \infty$.

Assumption 4.5. *The function* F *is Lipschitz continuous in its arguments.*

Assumption 4.6. *The population version of the community detection criterion* Q *is uniquely maximized over the set*

$$\left\{ \boldsymbol{R} \in \mathbb{R}^{K\times K\times K\times K} : R_{kluv} \geq 0,\ \sum_{k,l} R^{[ij]}_{kluv} = \hat{\Pi}^{[ij]}_{uv},\ \sum_{k,l} R^{[i]}_{kluv} = \hat{\Pi}^{[i]}_{uv} \right\}$$

by $\boldsymbol{R} = \boldsymbol{D}$, *where* $D_{kluv} = \begin{cases} \hat{\Pi}_{uv}, & (k = u, l = v) \\ 0, & o.w. \end{cases}$, *and* $\hat{\Pi}_{uv}$ *is defined as*

$$\begin{aligned} \hat{\Pi}_{uv} :&= \sum_{i\neq j}\sum_{k,l} R^{[ij]}_{kluv}(\boldsymbol{e}, \boldsymbol{z}) + \sum_{i=1}^m \sum_{k,l} R^{[i]}_{kluv}(\boldsymbol{e}, \boldsymbol{z}) \\ &= \frac{1}{n^2}\sum_{i\neq j}\sum_{p,q} I(z_{ip,jq,1} = u, z_{ip,jq,2} = v) + \frac{1}{n^2}\sum_{i=1}^n \sum_{p,q} I(z_{ip,iq,1} = u, z_{ip,iq,2} = v). \end{aligned}$$

Assumption 4.7. *The function* F *can be expressed by the following summation*

$$F\left(\frac{O(\boldsymbol{e})}{\mu_n}, f(\boldsymbol{e})\right) = \sum_{i\neq j} F\left(\frac{O^{[ij]}(\boldsymbol{e})}{\mu_n}, f^{[ij]}(\boldsymbol{e})\right) + \sum_{i=1}^m F\left(\frac{O^{[i]}(\boldsymbol{e})}{\mu_n}, f^{[i]}(\boldsymbol{e})\right).$$

Assumption 4.8. *The Bernoulli probability matrix* B *does not have identical elements.*

Assumption 4.9. *For any* $1 \leq u, v \leq K$, *the* K*-by-*K *matrix* $R_{\cdot\cdot uv}$ *has at least one nonzero element.*

Now we are ready to state the consistency results of community detection under the proposed mixed membership stochastic blockmodel for heterogeneous networks.

Theorem 4.1. *Suppose Assumptions 4.1–4.7 hold. Then for any community detection criterion of the following form*

$$Q(\boldsymbol{e}) = F\left(\frac{O(\boldsymbol{e})}{\mu_n}, f(\boldsymbol{e})\right),$$

$Q(\boldsymbol{e})$ *is consistent under mixed membership stochastic blockmodels for heterogeneous networks.*

The following corollary gives the community detection consistency result for the maximum likelihood estimator.

Corollary 4.1. *Suppose that Assumptions 4.1–4.4, and 4.8–4.9 hold. Then the maximum likelihood estimator of the community assignment is consistent under the mixed membership stochastic blockmodel for heterogeneous networks.*

In this paper, we adopt a Bayesian approach and estimate the community labels by the posterior mode. The following corollary gives the community detection consistency result of this estimator.

Corollary 4.2. *Suppose that Assumptions 4.1–4.4, and 4.8–4.9 hold. Then the Bayesian estimator of the community assignment given by the posterior mode is consistent under the mixed membership stochastic blockmodel for heterogeneous networks.*

The proofs of Theorem 4.1, Corollary 4.1 and Corollary 4.2 are given in the supplementary material (Huang et al., 2019).

5 Simulation results

We conducted three simulation studies to compare the performance between our algorithm and the spectral clustering algorithms for heterogeneous networks (Sengupta and Chen, 2015). In all of the simulation studies, we consider bi-type heterogeneous networks simulated from the mixed membership stochastic blockmodel. The networks are simulated as follows:

- For each node x_{ip} from type i, $i = 1, 2$, $p = 1, \ldots, n_i$:
  - sample $\boldsymbol{\pi}_{ip} \sim \text{Dirichlet}_K(\boldsymbol{\alpha}_i)$.
- For each pair of nodes (x_{ip}, x_{jq}), $i, j = 1, 2$, $p = 1, \ldots, n_i$, $q = 1, \ldots, n_j$:
  - sample the membership of x_{ip}: $\boldsymbol{z}_{ip,jq,1} \sim \text{Multinomial}_K(\boldsymbol{\pi}_{ip})$.
  - sample the membership of x_{jq}: $\boldsymbol{z}_{ip,jq,2} \sim \text{Multinomial}_K(\boldsymbol{\pi}_{jq})$.
  - sample $Y_{ij}(p, q) \sim \text{Bernoulli}(\boldsymbol{z}_{ip,jq,1} B_{ij} \boldsymbol{z}_{ip,jq,2})$.

In all simulations, we studied networks with a total of $N = 200$ nodes, of which 100 were of type I and 100 of type II. The number of groups is fixed to be $K = 4$. We set $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \alpha \mathbf{1}_K$, where $\mathbf{1}_K$ is a K-vector of 1's. We tried three different values of α ($\alpha = 0.05$, 0.1 and 0.25) to simulate different settings of the membership. When $\alpha = 0.05$, each node has almost unique membership, and we assigned the group with the highest probability to each node so that all nodes take unique membership in this case. As α increases, the nodes will have more diffused membership. When $\alpha = 0.1$, each node belongs to about 1.5 groups on average. When $\alpha = 0.25$, each node belongs to about 2 groups on average.

The Bernoulli probability matrix B is given by

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where

$$\begin{aligned} B_{11} &= p_1 \mathbf{1}_K \mathbf{1}_K' + (r_1 - p_1) \boldsymbol{I}_K, \\ B_{12} = B_{21} &= p_2 \mathbf{1}_K \mathbf{1}_K' + (r_2 - p_2) \boldsymbol{I}_K, \\ B_{22} &= p_3 \mathbf{1}_K \mathbf{1}_K' + (r_3 - p_3) \boldsymbol{I}_K. \end{aligned}$$

Here $\boldsymbol{I}_K$ is a K-by-K identity matrix. Under this setting, the parameters r_1 and r_3 represent the intra-group link probability of the type I-type I and type II-type II homogeneous networks, respectively, and r_2 denotes the intra-group inter-type link probability of the type I-type II network. The strength of inter-group homophily for the type I-type I (type II-type II) homogeneous network is represented by p_1 (p_3), and p_2 denotes the strength of inter-group inter-type homophily for the type I-type II network.

We tried to compare our algorithm with the Het-SC algorithm proposed by Sengupta and Chen (2015). However, the spectral clustering algorithm is designed for the situation that each node belongs to one unique group. In order to better compare the performance of Het-SC and our algorithm under the mixed membership setting, we adopted a revised version of the spectral clustering which has been used before in Airoldi et al. (2005). Based on the cluster prediction of Het-SC, we calculated the relative distance between each node to the centroids of clusters. Then we normalized the inverse distance to obtain the mixed membership probability vector for each node. The closer the node is to the cluster centroid, the higher probability it is assigned to that cluster.

5.1 Performance evaluation

For unique membership case ($\alpha = 0.05$), we evaluated the performance of the algorithm by the error rate, which is the proportion of nodes that are assigned to the wrong group/cluster. Since for both Het-SC and our algorithm, there is an identifiability issue with the cluster labels, we searched through all permutations to find the one that maximizes the accuracy (or minimizes the error rate).

To measure the performance in the mixed membership case ($\alpha = 0.1$ or 0.25), we used another way to define the error rate. For each node, its error rate is computed by

$$1 - \frac{\mathrm{TP} + \mathrm{TN}}{K},$$

where K is the number of clusters, TP is the number of true positives and TN is the number of true negatives. True positives are the clusters containing the node and are correctly detected by the algorithm. True negatives are the clusters that the node does not belong to and are not classified by the algorithm. Then the error rate for the whole data set is the average of the error rate for all of the nodes. Similar to the unique

membership case, all possible permutations of the clusters are considered and the one with minimum error rate is taken as the final error rate.

To better measure the quality of the overlapping community detection under the mixed membership setting, we introduce the extended version of the normalized mutual information (NMI), which was proposed by Lancichinetti et al. (2009) and reviewed by Xie et al. (2013) as one of the most widely used measures for overlapping communities. Suppose the number of clusters is K and the number of nodes is N. For each node, its membership can be expressed as a binary vector of length K. Then we have an $N \times K$ assignment matrix for all the nodes. We use X and Y to denote the assignment matrices obtained by the algorithm and the truth, respectively. The kth entry of each row in X can be treated as a random variable X_k, with probability distribution defined as

$$P(X_k = 1) = N_k/N, \quad P(X_k = 0) = 1 - P(X_k = 1),$$

where N_k is the number of nodes belonging to cluster k. Similarly, we can obtain the probability distribution of Y_l (the lth column of Y) and the joint probability distribution of (X_k, Y_l). Define the entropies $H(X_k)$, $H(Y_l)$ and $H(X_k, Y_l)$ by

$$-\sum_i p_i \log p_i,$$

where p_i's are a discrete set of probabilities for the random variable (vector). Then we have

$$\begin{aligned} H(X_k|Y_l) &= H(X_k, Y_l) - H(Y_l), \\ H(X_k|Y) &= \min_l H(X_k|Y_l), \\ H(X|Y) &= \frac{1}{K}\sum_{k=1}^{K} \frac{H(X_k|Y)}{H(X_k)}. \end{aligned}$$

The extended NMI is defined by

$$NMI = 1 - [H(X|Y) + H(Y|X)]/2.$$

The extended NMI should be between 0 and 1, with 1 implying a perfect match between the true and assigned clusters, and 0 indicating random cluster assignment with respect to the true cluster labels. A larger extended NMI indicates a better match with the truth.

5.2 Simulation 1

In this simulation we consider the heterogeneous network with the following setting: $r_1 = 0.3$, $r_3 = 0.7$, $p_1 = p_2 = p_3 = 0.1$. Type II nodes within the same group have a larger link probability than type I nodes. We also let r_2 increase from 0.3 to 0.7 in increments of 0.1 to simulate networks with different strength of intra-group inter-type link probability. Three choices of α, 0.05, 0.1 and 0.25, were considered. We applied three

methods to the simulated networks: the Het-SC algorithm (SC), the revised version of Het-SC (R-SC) (for $\alpha = 0.1$ and 0.25 only) and our proposed algorithm (MMSB). The results of the three algorithms are shown in Figures 1 and 2. Panel (a) of Figure 1 shows the average error rate for Het-SC and MMSB under the unique membership setting ($\alpha = 0.05$). Other panels in Figures 1 and 2 show the error rate and also the extended NMI for Het-SC, revised Het-SC and MMSB algorithms under the mixed membership setting. The membership gets more diffused with $\alpha = 0.25$ than with $\alpha = 0.1$. All of the results shown in the figures are the average of 10 simulations.

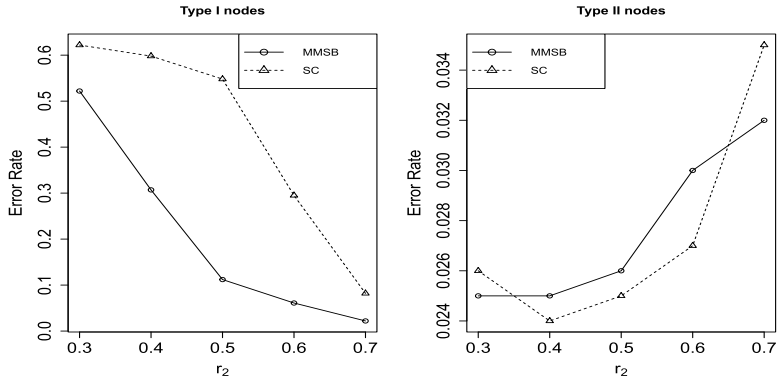
Based on the results in Figures 1 and 2, we can see the performance on type II nodes is always better than type I nodes due to the higher intra-group link probability. We can also see that as the intra-group inter-type link probability r_2 increases, the error rates usually decrease (and the NMIs increase) for almost all methods and all values of α. For the mixed membership case ($\alpha = 0.1$ or 0.25), the performance gets worse for all three methods when the membership becomes more diffused. When comparing MMSB with the other methods, in the unique membership case, although MMSB and the Het-SC algorithm have similar error rate for type II nodes, MMSB improves the accuracy of type I nodes a lot over the Het-SC for different r_2 values. When $\alpha = 0.1$ or 0.25, although we used the revised Het-SC to address the mixed membership case, it did not improve the performance of Het-SC and sometimes the performance is even worse. The revised Het-SC only outperforms Het-SC in terms of error rate when $\alpha = 0.25$ for type II nodes. On the other hand, MMSB has better performance (smaller error rate and larger NMI) than both Het-SC and the revised Het-SC for all values of r_2 and α.

We also compared the performance of variational method with MCMC for MMSB based on the setting in this simulation. To make the computation feasible for MCMC, we reduced the number of nodes to $n_1 = n_2 = 25$, and reduced the number of communities to $K = 3$. The hyperparameter α is set to be 0.1. All the results shown in Table 1 are the average of 10 simulations. For each simulation, variational method took about 30 seconds while MCMC took around 3.5 hours. The results in Table 1 show that variational method has better performance (smaller error rate and larger NMI) than MCMC for r_2 ranging from 0.3 to 0.7.

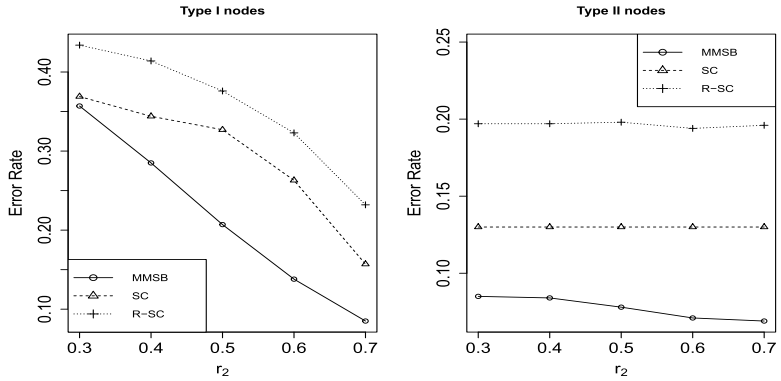
As the number of nodes increases, the number of latent variables grows at the rate of $O(n^2)$. We expect the computation time needed for MCMC will grow even faster than $O(n^2)$. The performance in this simulation indicates that MCMC will be too slow for large networks, such as the DBLP data in Section 6. The computational issue of MCMC is also discussed in Airoldi et al. (2008).

5.3 Simulation 2

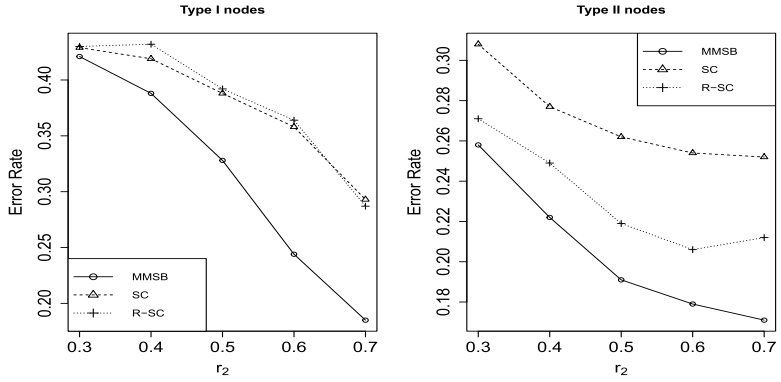
In this section, we consider the scenario that there is no homophilic community structure among type II-type II nodes. We set $r_1 = 0.3$, $r_3 = 0.1$ and $p_1 = p_2 = p_3 = 0.1$. We let r_2 increase from 0.3 to 0.7 in increments of 0.1. Under this setting, for type II nodes, the nodes within the same group do not have higher link probability than nodes in different groups. The results of the Het-SC algorithm, the revised Het-SC algorithm and the MMSB under different values of α are displayed in Figures 3 and 4. All of the results are the average of 10 simulations.



(a) $\alpha = 0.05$

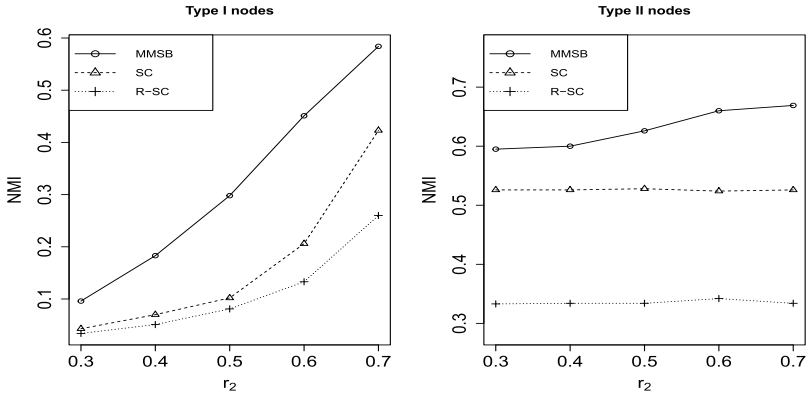


(b) $\alpha = 0.1$

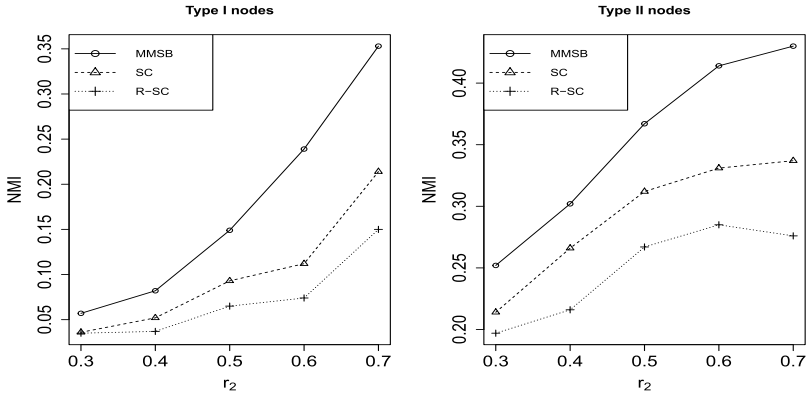


(c) $\alpha = 0.25$

Figure 1: Error rates of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 1.



(a) $\alpha = 0.1$

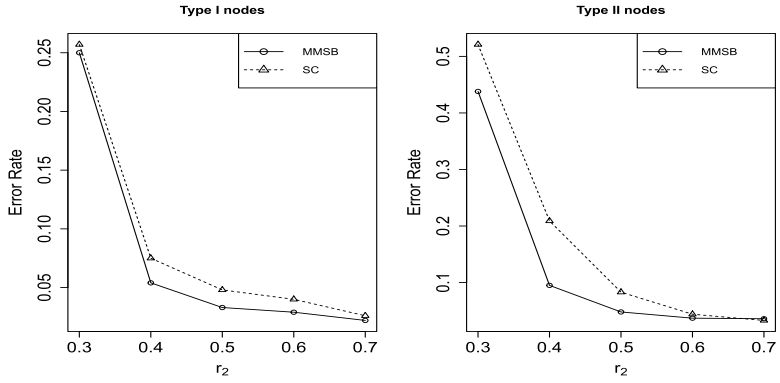


(b) $\alpha = 0.25$

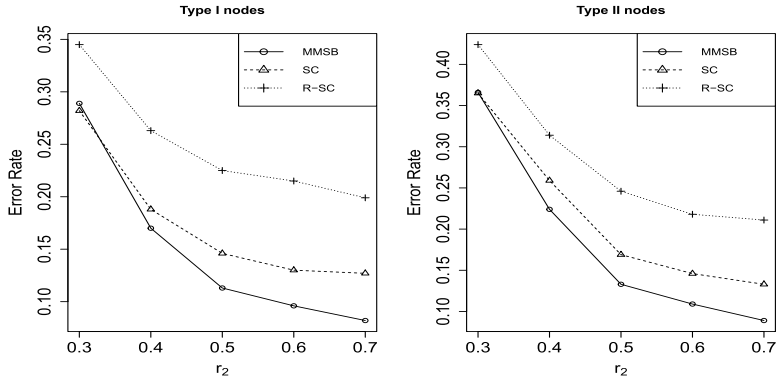
Figure 2: The extended NMI of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 1.

r_2	Error Rate				NMI			
	Type I nodes		Type II nodes		Type I nodes		Type II nodes	
	MCMC	VB	MCMC	VB	MCMC	VB	MCMC	VB
0.3	0.3400	0.1413	0.4040	0.1213	0.2847	0.4934	0.1914	0.5229
0.4	0.3333	0.1427	0.3907	0.1107	0.2903	0.4880	0.1991	0.5539
0.5	0.3240	0.1613	0.3760	0.1373	0.2653	0.4414	0.2260	0.5305
0.6	0.3627	0.1640	0.4053	0.1200	0.2478	0.4467	0.2006	0.5437
0.7	0.3427	0.1747	0.3960	0.1053	0.2587	0.4133	0.1937	0.6036

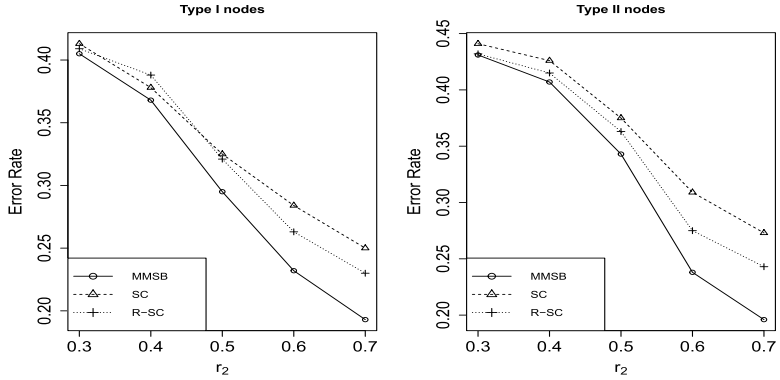
Table 1: Comparison between MCMC and variational method for MMSB.



(a) $\alpha = 0.05$

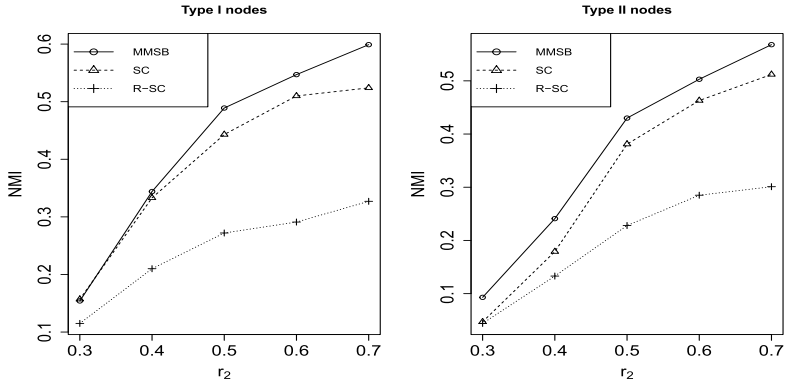


(b) $\alpha = 0.1$

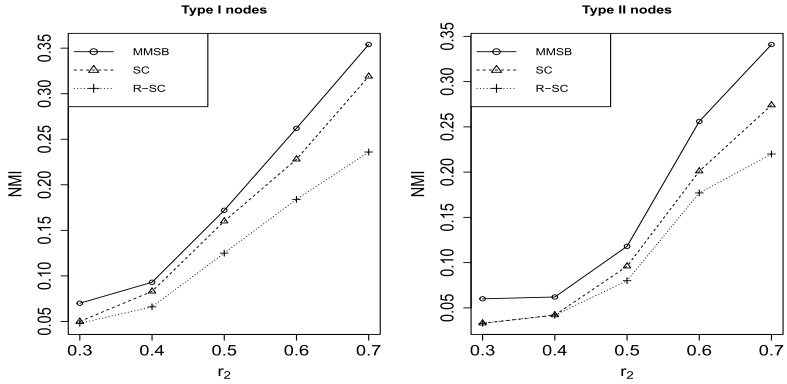


(c) $\alpha = 0.25$

Figure 3: Error rates of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 2.



(a) $\alpha = 0.1$



(b) $\alpha = 0.25$

Figure 4: The extended NMI of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 2.

Figures 3 and 4 show that the performance on type II nodes is always worse than type I nodes for all methods under different values of α and r_2. The lack of community structure in type II-type II network makes the clustering of type II nodes more difficult, since only the information in the type I-type II links can be used to assign groups. The revised Het-SC algorithm seems to have better performance than Het-SC in terms of error rate when the membership becomes more diffused. The revised Het-SC has smaller error rate than Het-SC when $\alpha = 0.25$, but its NMI is not better than Het-SC. Also, the revised Het-SC has worse performance than Het-SC when $\alpha = 0.1$. As for our method, MMSB improves the accuracy (in terms of both error rate and NMI) over both Het-SC and the revised Het-SC for both type of nodes and all values of α and almost all values of r_2.

5.4 Simulation 3

We consider the scenario that there are no type II-type II links in this section. We set $r_1 = 0.3$, $r_3 = 0$, $p_1 = p_2 = 0.1$ and $p_3 = 0$. Again r_2 increases from 0.3 to 0.7 in increments of 0.1. The results of the three algorithms under different values of α are displayed in Figures 5 and 6. All of the results are the average of 10 simulations.

Similar to simulation 2, the performance on type II nodes is always worse than type I nodes for all methods under different values of α and r_2, since type II-type II links are missing. The revised Het-SC algorithm still has better performance than Het-SC in terms of error rate when the membership is more diffused ($\alpha = 0.25$). Our MMSB method has the best performance (in terms of both error rate and NMI) in almost all cases compared with Het-SC and the revised Het-SC.

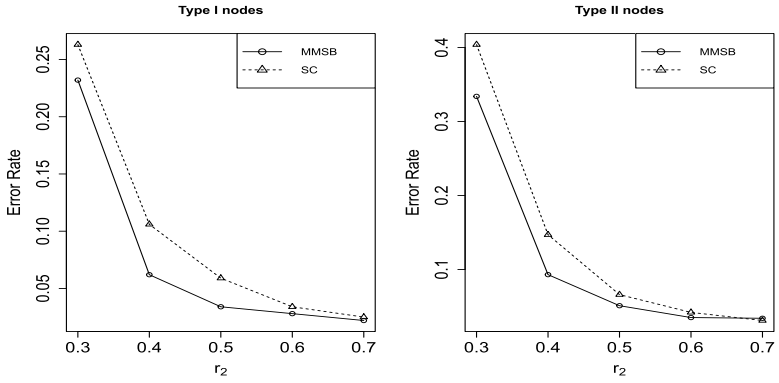
6 Analysis of the DBLP data

DBLP (Digital Bibliography & Library Project) is a computer science bibliography website, which contains over 3.3 million publications published by more than 1.7 million authors. Gao et al. (2009) and Ji et al. (2010) extracted a sub-network from the DBLP data set, which contains 14376 papers, 20 conferences, 14475 authors and 8920 terms. The sub-network focuses on four areas of computer science: database, data mining, artificial intelligence (AI) and information retrieval, which form four groups in the sub-network. Gao et al. (2009) and Ji et al. (2010) manually labeled the area of 4057 authors, 100 papers and all 20 conferences. In our study, we focus on the sub-network which contains the labeled 4057 authors and all 20 conferences. We consider only one type of links: the author-conference links (author attended conference or have papers presented at the conference). Therefore, we have a heterogeneous network with two types of nodes: 4057 authors and 20 conferences, and the author-conference links. The goal of this application is to identify the research areas (communities) of the authors.

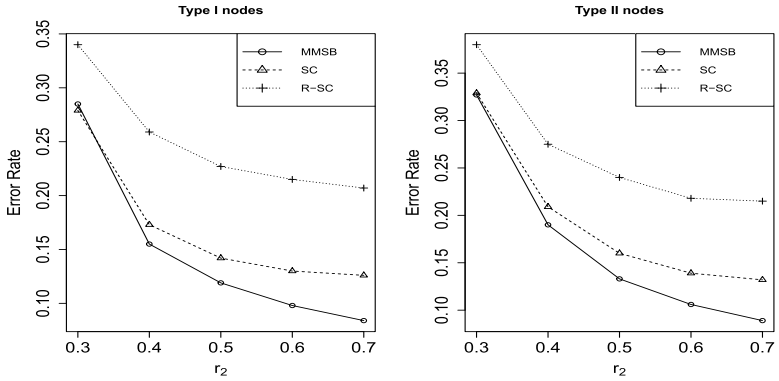
Gao et al. (2009) and Ji et al. (2010) only assigned a single group (research area) to each author when they manually labeled the data set. However in practice, it is possible that people have more than one research area. Since the author-conference links are very informative for clustering (Sengupta and Chen, 2015), here we create our own label for authors based on the areas of the conferences they attended. That is, we assigned the labels of all conferences the author attended to the author. After the assignment, there are more than 27% of the authors having more than one research area and the average number of areas for each author is 1.36.

We applied MMSB to the data with pre-specified number of groups $K = 4$, corresponding to the four research areas in computer science. Table 2 shows the percentage of authors in each research area. Overall about 20% of authors are active in more than one research area.

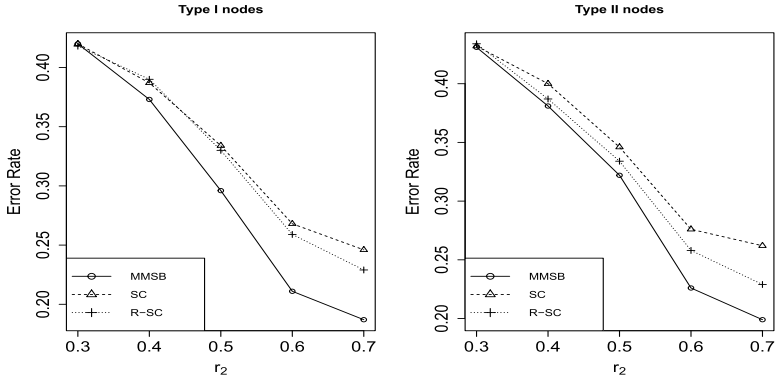
We also compared the performance of MMSB with the Het-SC and the revised Het-SC algorithms. Both the error rate and the extended NMI are used to evaluate the performance of each method. The results are shown in Table 3. The results in bold



(a) $\alpha = 0.05$

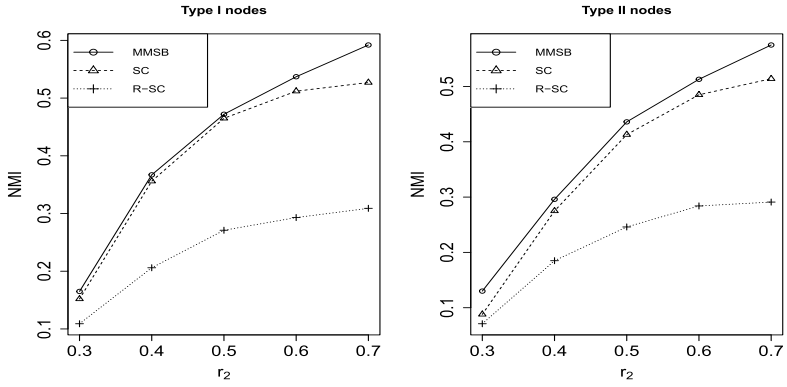


(b) $\alpha = 0.1$

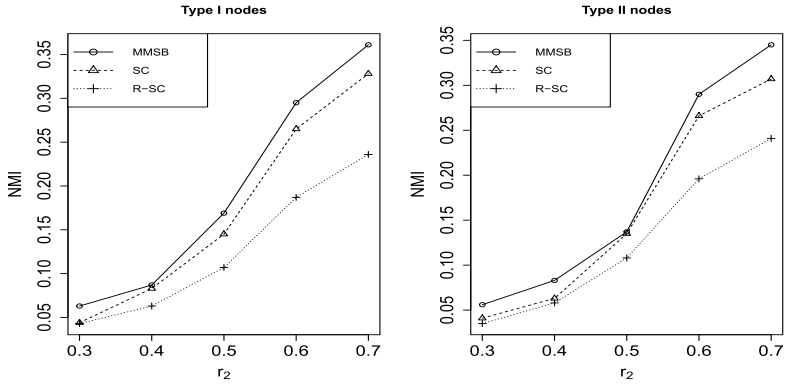


(c) $\alpha = 0.25$

Figure 5: Error rates of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 3.



(a) $\alpha = 0.1$



(b) $\alpha = 0.25$

Figure 6: The extended NMI of Het-SC (SC), revised Het-SC (R-SC) and MMSB algorithms for Simulation 3.

Community	Database	Data Mining	AI	Information Retrieval
% of authors	38%	26%	31%	27%

Table 2: The percentage of authors in each community.

denote the best performance under a certain criterion. The results suggest that MMSB outperforms Het-SC and the revised Het-SC in terms of both error rate and NMI. The revised version of Het-SC does improve the performance of Het-SC in terms of error rate and also the NMI.

To further assess the performance of the proposed method, we examined one specific case. In the original dataset, the author with id=78624 is manually labeled to the data mining area, while the proposed algorithm uncovers two research areas of this author:

Error rate			NMI		
SC	R-SC	MMSB	SC	R-SC	MMSB
0.0968	0.0792	**0.0458**	0.5850	0.6035	**0.7681**

Table 3: Performance of Het-SC (SC), revised Het-SC (R-SC) and MMSB for the DBLP dataset.

data mining and database. These two areas match with labels we assigned based on the conferences the author attended. To determine the true research areas of this author, we manually checked the author's publications on the DBLP webpage. We checked the journal/conference the paper is published in, the title, key words or abstract of the paper to determine whether a paper is database related. For the 60 publications shown on the DBLP webpage, 24 of them are related to database. It seems natural to consider database to be a research area of this author. The additional research area uncovered by the proposed method shows the advantage of MMSB.

7 Discussion

This article proposes a statistical framework for community detection in heterogeneous networks, which extends the original MMSB for homogeneous networks to heterogeneous networks. The proposed method relaxes the unique cluster limitation of the classical stochastic blockmodel and allows each node of different types to have multiple memberships. The use of the variational algorithm makes the method scalable to large networks. The proposed procedure is shown to be consistent for community detection under the MMSB for heterogeneous networks. Simulation studies show that the proposed method gave more accurate clustering results than the spectral clustering algorithm for heterogeneous version of the stochastic blockmodel, especially in diffused membership case. The analysis on the DBLP data also demonstrates the advantage of our method.

In the simulation studies, we tried different values of the hyperparameter α to simulate different settings of the membership. As α increases, the nodes will have more diffused membership. In the real data example, the results were not sensitive to the choice of α. In practice 0.05-0.25 seems to be a reasonable range for α. It is also possible to update α in the variational M step that maximizes the lower bound.

There is some connection between the MMSB and the latent space model (Hoff et al., 2002), as discussed in Goldenberg et al. (2010) and Airoldi et al. (2008). Both models try to study the latent structure of the network to explain the connectivity of the observed network. It is of interest to develop latent space models for heterogeneous networks.

Supplementary Material

Supplementary Material for "Mixed Membership Stochastic Blockmodels for Heterogeneous Networks" (DOI: 10.1214/19-BA1163SUPP; .pdf). The supplementary material contains the details of the variational posterior inference, variational EM algorithm, and the proofs of theoretical results in Section 4.

References

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2005). "A latent mixed membership model for relational data." In *Proceedings of the 3rd International Workshop on Link Discovery*, 82–89. ACM. 724

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research*, 9(Sep): 1981–2014. 712, 713, 726, 734

Gao, J., Liang, F., Fan, W., Sun, Y., and Han, J. (2009). "Graph-based consensus maximization among multiple supervised and unsupervised models." In *Advances in Neural Information Processing Systems*, 585–593. 731

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). "A survey of statistical network models." *Foundations and Trends in Machine Learning*, 2(2): 129–233. 711, 734

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). "Model-based clustering for social networks." *Journal of the Royal Statistical Society: Series A*, 170(2): 301–354. MR2364300. doi: https://doi.org/10.1111/j.1467-985X.2007.00471.x. 712

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). "Latent space approaches to social network analysis." *Journal of the American Statistical Association*, 97(460): 1090–1098. MR1951262. doi: https://doi.org/10.1198/016214502388618906. 712, 734

Huang, W., Liu, Y., and Chen, Y. (2019). "Supplementary Material for "Mixed Membership Stochastic Blockmodels for Heterogeneous Networks"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1163SUPP. 716, 717, 723

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). "Graph regularized transductive classification on heterogeneous information networks." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 570–586. Springer. 731

Jonsson, P. F., Cavanna, T., Zicha, D., and Bates, P. A. (2006). "Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis." *BMC Bioinformatics*, 7(1): 2. 711

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). "Detecting the overlapping and hierarchical community structure in complex networks." *New Journal of Physics*, 11(3): 033015. 725

Nowicki, K. and Snijders, T. A. B. (2001). "Estimation and prediction for stochastic blockstructures." *Journal of the American Statistical Association*, 96(455): 1077–1087. MR1947255. doi: https://doi.org/10.1198/016214501753208735. 712

Sengupta, S. and Chen, Y. (2015). "Spectral clustering in heterogeneous networks." *Statistica Sinica*, 25(3): 1081–1106. MR3410299. doi: https://doi.org/10.5705/ss.2013.231. 712, 713, 718, 723, 724, 731

Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers. 712

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical models, exponential families, and variational inference." *Foundations and Trends in Machine Learning*, 1(1–2): 1–305. 716

Xie, J., Kelley, S., and Szymanski, B. K. (2013). "Overlapping community detection in networks: The state-of-the-art and comparative study." *ACM Computing Surveys*, 45(4): 43:1–43:35. 725

Zhang, J. and Chen, Y. (2019). "Modularity based community detection in heterogeneous networks." *Statistica Sinica*, in press. doi: https://doi.org/10.5705/ss.202017.0399. 719

Zhao, Y., Levina, E., and Zhu, J. (2012). "Consistency of community detection in networks under degree-corrected stochastic block models." *The Annals of Statistics*, 40(4): 2266–2292. 721

Acknowledgments

The authors thank the editor, the associate editor, and the referees for valuable suggestions. This work was supported in part by National Science Foundation grant DMS-1406455.