Journal of Machine Learning Research 25 (2024) 1-42 Submitted 7/23; Revised 4/24; Published 5/24

Variational Estimators of the Degree-corrected Latent Block Model for Bipartite Networks

Yunpeng Zhao YUNPENG.ZHAO@COLOSTATE.EDU
Department of Statistics
Colorado State University
Fort Collins, CO 80523, USA

Ning Hao NHAO@MATH.ARIZONA.EDU
Department of Mathematics
University of Arizona
Tucson, AZ 85721, USA

Ji Zhu JIZHU@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Xiaotong Shen

Abstract

Bipartite graphs are ubiquitous across various scientific and engineering fields. Simultaneously grouping the two types of nodes in a bipartite graph via biclustering represents a fundamental challenge in network analysis for such graphs. The latent block model (LBM) is a commonly used model-based tool for biclustering. However, the effectiveness of the LBM is often limited by the influence of row and column sums in the data matrix. To address this limitation, we introduce the degree-corrected latent block model (DC-LBM), which accounts for the varying degrees in row and column clusters, significantly enhancing performance on real-world data sets and simulated data. We develop an efficient variational expectation-maximization algorithm by creating closed-form solutions for parameter estimates in the M steps. Furthermore, we prove the label consistency and the rate of convergence of the variational estimator under the DC-LBM, allowing the expected graph density to approach zero as long as the average expected degrees of rows and columns approach infinity when the size of the graph increases.

Keywords: biclustering, bipartite graph, identifiability, label consistency, variational expectation-maximization

1. Introduction

Biclustering or coclustering, first considered by Hartigan (1972), is an unsupervised learning task that simultaneously clusters the rows and columns of a rectangular data matrix. Biclustering is a machine learning technique with many applications, such as in genomics (Cheng and Church, 2000; Pontes et al., 2015), recommender systems (Alqadah et al., 2015), and text mining (de Castro et al., 2007; Orzechowski and Boryczko, 2016). Similar to standard cluster analysis, an exhaustive search for all possible partitions of rows and columns is intractable due to the exponential growth in the number of possible partitions with the

increase in row and column numbers. Many popular biclustering methods employ greedy algorithms to find the local optimal partition according to certain criteria. Examples include Minimum Sum-Squared Residue Coclustering (MSSRCC) (Cho et al., 2004) and Large Average Submatrices (LAS) (Shabalin et al., 2009). For a systematic review and comparison of typical biclustering algorithms, readers are referred to Padilha and Campello (2017).

Mixture models, such as Gaussian mixture models (Fraley and Raftery, 2002) for continuous data and Bernoulli mixture models for binary data (Celeux and Govaert, 1991), provide a natural probabilistic framework for standard cluster analysis, where each observation is associated with a latent cluster label that can be inferred by estimating posterior probabilities given the data. Similarly, model-based approaches have also been developed for biclustering. We focus on one of the most popular models for biclustering — the latent block model (LBM), first proposed by Govaert and Nadif (2003). The LBM is a natural generalization of mixture models to the "two-dimensional" case, where the probability distribution of each entry of the data matrix depends on both the row and column labels.

The expectation-maximization (EM) algorithm is the most widely-used algorithm for fitting a mixture model. However, the E step of the EM algorithm for the LBM becomes intractable due to the complex dependence structure among entries of the data matrix and cluster labels (Govaert and Nadif, 2008). To overcome this computational obstacle, Govaert and Nadif (2003) proposed the block classification EM (CEM) algorithm. It includes an additional C step that assigns hard cluster labels based on the estimated posterior probabilities in the current iteration. Once the hard row labels are obtained, the column labels can be updated using a standard EM algorithm, and vice versa. Govaert and Nadif (2006) introduced the fuzzy block criterion, which avoids the conversion of posterior probabilities into hard labels. The criterion function was later reinterpreted within the framework of variational EM algorithms and named block EM (Govaert and Nadif, 2008), a technique we will adapt in this paper. In this approach, the block EM algorithm maximizes a lower bound of the log-likelihood function, referred to as the variational approximation. It imposes a constraint that allows for factoring the posterior distribution of row and column labels.

We study the biclustering problem for data matrices in which an entry represents the relationship between the corresponding row and column objects. In other words, we consider biclustering on bipartite (two-mode) networks with single or multiple edges. This formulation is closely related to another widely-studied area—community detection for one-mode networks. Although developed almost independently from biclustering, the models and algorithms for community detection, and challenges that they face are similar to biclustering. The stochastic block model (SBM), first proposed by Holland et al. (1983), is the best studied model in the community detection literature. The SBM can be viewed as the analogue of the LBM for symmetric binary networks although there is little overlap between literatures of the two models until recently. Bickel and Chen (2009) established the first theoretical framework to study the consistency of estimated labels under the SBM, and in particular they proved the consistency of profile likelihood estimators. The theoretical framework was extended by Flynn and Perry (2020) to biclustering for a wide range of data modalities, including binary, count, and continuous observations. When fitting the SBM, the E step of the classical EM algorithm is intractable as in the LBM. Various computationally efficient approaches have been proposed for fitting the SBM, including, but not limited to, variational approximation (Daudin et al., 2008; Bickel et al., 2013), pseudo likelihood

(Amini et al., 2013), split likelihood (Wang et al., 2021a), and profile-pseudo likelihood (Wang et al., 2023). Readers are referred to Abbe (2017) and Zhao (2017) for surveys on the computational and theoretical advances for the SBM and related models.

The LBM has a notable limitation in practical applications to bipartite networks: it tends to cluster rows with similar row degrees (i.e. row sums) and columns with similar column degrees (i.e. column sums) together. This issue is also observed in the SBM used for symmetric networks. Karrer and Newman (2011) proposed the degree-corrected stochastic block model (DC-SBM) that includes an additional set of parameters controlling expected degrees. The degree parameters were usually estimated implicitly in the community detection literature, which is partially due to the model identifiability issue. For example, Amini et al. (2013) proposed the conditional pseudo likelihood which models the number of edges in a block as a multinomial variable conditional on the observed degrees. A similar approach was used in the split likelihood method (Wang et al., 2021a).

In this paper, we propose a degree-corrected latent block model (DC-LBM) to accommodate degree heterogeneity in biclustering. Instead of using any surrogate, we take a direct approach: we adapt the block EM algorithm to estimate all parameters in the original form of the DC-LBM, including degree parameters, without ad-hoc modification. We show that the observed row and column degrees, up to constants, are exactly the maximizers for the corresponding degree parameters in the M step given any probability assignment on the cluster labels, if we model the entries of the data matrix as independent Poisson variables conditional on the cluster labels. The estimates of the degree parameters therefore remain constant in the algorithm, which results an elegant and efficient estimating procedure under the variational EM framework.

The theoretical contribution of the present paper is the establishment of the label consistency and the rate of convergence of the variational estimator under the DC-LBM. Brault et al. (2020) proved the consistency of the maximum likelihood estimator and the variational estimator under the classical LBM by showing both the marginal likelihood and the variational approximation are asymptotically equivalent to the complete data likelihood. We take a new approach by directly proving the variational approximation uniformly converges to its population version and the true cluster labels are a well-separated maximizer of the population version, which implies label consistency. A key ingredient in the proof is a uniform concentration inequality over probability assignments of cluster labels. The proof can accommodate degree parameters and requires weaker conditions. In particular, we allow that the expected graph density goes to zero provided that both the average expected row and column degrees go to infinity as the size of the network increases, which is a typical condition for label consistency for symmetric networks (Bickel and Chen, 2009; Zhao et al., 2012).

This paper focuses on likelihood-based approaches to biclustering on bipartite networks. Spectral clustering, co-clustering, and their numerous variants, as computationally efficient non-likelihood-based approaches, have been widely applied to network data. This class of methods typically involves the construction of various types of graph Laplacians, followed by the application of eigendecomposition or singular value decomposition to the constructed matrices. Spectral clustering methods were first proposed for undirected networks (Rohe et al., 2011; Jin, 2015; Sarkar and Bickel, 2015; Lei and Rinaldo, 2015), and soon be generalized to directed networks (Rohe et al., 2016; Wang et al., 2021b; Zhang et al., 2022).

It is particularly worth mentioning that biclustering is closely related to co-clustering in directed networks when the row and column labels are presumed to be different; that is, they represent distinct communities for "sending" versus "receiving", as in the setup of Rohe et al. (2016). The only difference compared to the bipartite graph problem is the assumption of square matrices for directed graphs.

In terms of related theoretical results, Zhao et al. (2012) extended the theoretical framework of Bickel and Chen (2009) and proved a general theorem for label consistency under the DC-SBM, where the degree parameters can take a finite number of possible values. Amini et al. (2013); Wang et al. (2021a, 2023) also considered the DC-SBM but the theoretical analyses focused on the classical SBM. Flynn and Perry (2020) extended the theoretical framework of profile likelihood methods (Bickel and Chen, 2009) to biclustering. Mariadassou et al. (2015) proposed a unified framework for studying the convergence of the posterior distribution of cluster labels under both the SBM and LBM, assuming known parameter values.

The rest of this paper is organized as follows. In Section 2, we introduce the DC-LBM with the Poisson distribution. In Section 3, we propose a variational EM algorithm for DC-LBM, with a key property stating that the row and column degrees maximize the objective function in the M step, given any probability assignment on the cluster labels. Section 4 addresses asymptotic properties. We establish the consistency of label estimation and the rate of convergence under the DC-LBM with both the Poisson and the Bernoulli distributions on edges. In Section 5, we compare the performance of the proposed method with other popular biclustering algorithms in various setups. In Section 6, we apply the proposed method to a benchmark data set for biclustering—the MovieLens data set. All technical proofs are provided in the appendix. Additionally, we include the analysis of an SMS spam data set in the appendix.

2. Model

We introduce the degree-corrected latent block model (DC-LBM) for bipartite networks in this section. Consider an adjacency matrix $A = [A_{ij}]$ with m rows and n columns, where each A_{ij} is a non-negative integer that represents multiple edges or an edge with an integer weight from i to j.

We assume that the row (resp. column) indices are partitioned into K (resp. L) latent clusters. Denote the cluster labels on rows by $z = (z_1, ..., z_m)^T$ and the labels on columns by $w = (w_1, ..., w_n)^T$. Assume $z_1, ..., z_m$ are independently and identically distributed (i.i.d.) multinomial variables with Multi$(1, \pi = (\pi_1, ..., \pi_K)^T)$. Similarly, $w_1, ..., w_n$ are i.i.d. multinomial variables with Multi$(1, \rho = (\rho_1, ..., \rho_L)^T)$. In addition to the cluster structure, we use "degree parameters" $\theta = (\theta_1, ..., \theta_m)^T$ and $\lambda = (\lambda_1, ..., \lambda_n)^T$ to model the propensities of row and column objects to form links. Both cluster labels z, w and degree parameters θ, λ are unknown, and throughout the paper we treat z, w as latent variables and θ, λ as parameters.

Conditional on the cluster labels, $\{A_{ij}\}$ are independent Poisson variables with mean $\{\theta_i \lambda_j \mu_{z_i w_j}\}$ where $\mu = [\mu_{kl}]$ is a K-by-L matrix. Hence, the joint likelihood of z, w and A

is

$$P(z,w,A;\pi,\rho,\theta,\lambda,\mu)=\left(\prod_{i=1}^{m}\pi_{z_i}\right)\left(\prod_{j=1}^{n}\rho_{w_j}\right)\prod_{i=1}^{m}\prod_{j=1}^{n}e^{-\theta_i\lambda_j\mu_{z_iw_j}}\frac{(\theta_i\lambda_j\mu_{z_iw_j})^{A_{ij}}}{A_{ij}!}.$$

Certainly, the parameters θ, λ and μ must meet specific constraints for identifiability. We will discuss the resolution to this issue in Sections 3 and 4.

We assume the Poisson distribution in the model for the convenience of algorithm development. The DC-SBM for symmetric networks, first proposed by Karrer and Newman (2011), also assumes the Poisson distribution. As will be shown in simulation studies and data analysis, the model performs well for networks with binary edges. We prove label consistency and the rate of convergence under both the Bernoulli and Poisson models in Section 4. Furthermore, the model reduces to the classical LBM when $\theta_i\equiv 1,\lambda_j\equiv 1, i=1,...,m, j=1,...,n$.

Since the cluster labels are latent, we consider the marginal likelihood of A

$$P(A;\pi,\rho,\theta,\lambda,\mu)=\sum_{z\in\Omega_z}\sum_{w\in\Omega_w}P(z,w,A;\pi,\rho,\theta,\lambda,\mu), \tag{1}$$

where $\Omega_z=\{1,...,K\}^m$ and $\Omega_w=\{1,...,L\}^n$. Note that a brute-force calculation of the summation in (1) is intractable because the number of terms grows exponentially with m and n, and $P(z,w,A;\pi,\rho,\theta,\lambda,\mu)$ cannot factor under these sums. The classical EM algorithm is intractable for the same reason. Specifically, the E step involves a sum of K^mL^n terms as in (1), owing to the dependence structure among the variables A_{ij}.

3. Variational expectation-maximization algorithm

To address the computational challenges, we employ a strategy akin to the variational EM algorithm used for the classical LBM, as outlined by Govaert and Nadif (2006, 2008). This algorithm reframes the EM algorithm as a coordinate ascent method, where the E-step is treated as maximization across a range of probability measures. When the E-step is computationally daunting, a solution can be sought by maximizing over a *constrained* space instead.

3.1 General framework

We begin with the general framework of the algorithm. Due to Jensen's inequality,[1]

$$\log P(A)=\log\sum_{z\in\Omega_z}\sum_{w\in\Omega_w}P(z,w,A;\Phi)\geq\sum_{z\in\Omega_z}\sum_{w\in\Omega_w}q(z,w)\log\left\{\frac{P(z,w,A;\Phi)}{q(z,w)}\right\},$$

where $\Phi=(\pi,\rho,\theta,\lambda,\mu)$, and q is a probability measure over $\Omega_z\times\Omega_w$ satisfying

$$\sum_{z\in\Omega_z}\sum_{w\in\Omega_w}q(z,w)=1.$$

1. We use the convention $0\log 0=0$, which is consistent with $\lim_{x\to 0}x\log x=0$.

Equality holds if $q(z,w)$ is equal to the posterior probability $\mathbb{P}(z,w|A)$. The standard EM algorithm is to iteratively update Φ and q.

As mentioned in Section 2, a brute-force calculation of $\sum_{z\in\Omega_z}\sum_{w\in\Omega_w}$ is intractable. To resolve this issue, Govaert and Nadif (2008) proposed to impose the constraint $q(z,w) = q_1(z)q_2(w)$ with $\sum_{z\in\Omega_z} q_1(z) = 1$ and $\sum_{w\in\Omega_w} q_2(w) = 1$. Let

$$J(q_1, q_2, \Phi) = \sum_{z\in\Omega_z}\sum_{w\in\Omega_w} q_1(z)q_2(w)\log\left\{\frac{P(z,w,A;\Phi)}{q_1(z)q_2(w)}\right\}. \tag{2}$$

Conditional on A, the latent variables z and w are not independent. Therefore, $q_1(z)q_2(w)$ is not exactly the posterior probability $\mathbb{P}(z,w|A)$. But their dependence is weak for large m and n. The intuition behind the constraint $q(z,w) = q_1(z)q_2(w)$ is that if the detection of clusters is consistent (as we will prove in Section 4), the posterior distribution of (z,w) will eventually concentrate on a single realization—the true cluster labels. In other words, the posterior distribution of (z,w) can be approximated by a Dirac measure, allowing for factorization.

The variational EM algorithm iteratively updates the parameters Φ (M step) and the probability measures q_1 and q_2 (E step). We specify the two steps in the following subsections.

3.2 M step

The degree parameters in the likelihood were typically indirectly addressed in the literature. Instead of explicitly including the degree parameters, likelihood functions conditional on observed degrees were used, such as conditional pseudo-likelihood (Amini et al., 2013) and conditional split likelihood (Wang et al., 2021a). This approach is based on the observation that the conditional distribution of independent Poisson variables, given their sum, follows a multinomial distribution (Amini et al., 2013).

In this paper, we take a different and more direct approach. Instead of using any surrogate (Amini et al., 2013; Wang et al., 2021a), in the M step we rigorously maximize $J(q_1, q_2, \Phi)$ over all parameters including θ and λ given q_1 and q_2. A key observation here is that the row and column degrees are global optimizers for θ and λ for any given q_1 and q_2 (Proposition 1).

Let

$$\mathbb{P}_{q_1}(z_i = k) = \sum_{z\in\Omega_z} q_1(z_1, ..., z_{i-1}, k, z_{i+1}, ..., z_m),\ i = 1, ..., m, k = 1, ..., K,$$

and

$$\mathbb{P}_{q_2}(w_j = l) = \sum_{w\in\Omega_w} q_2(w_1, ..., w_{j-1}, l, w_{j+1}, ..., w_n),\ j = 1, ..., n, l = 1, ..., L.$$

Let $d_i^r = \sum_{j=1}^n A_{ij}$ $(i = 1, ..., m)$ and $d_j^c = \sum_{i=1}^m A_{ij}$ $(j = 1, ..., n)$ be the row and column degrees, respectively.

We have the following result:

Proposition 1 *For any fixed q_1 and q_2, $\hat{\Phi} = (\hat{\pi}, \hat{\rho}, \hat{\theta}, \hat{\lambda}, \hat{\mu})$ defined below is a global maximizer of $J(q_1, q_2, \Phi)$.*

$$\hat{\theta}_i = d_i^r,\ i = 1, ..., m,$$
$$\hat{\lambda}_j = d_j^c,\ j = 1, ..., n,$$
$$\hat{\mu}_{kl} = \frac{\sum_{i=1}^m \sum_{j=1}^n A_{ij} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)}{\sum_{i=1}^m \sum_{j=1}^n d_i^r d_j^c \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)},\ k = 1, ..., K, l = 1, ..., L,$$
$$\hat{\pi}_k = \frac{1}{m}\sum_{i=1}^m \mathbb{P}_{q_1}(z_i = k),\ k = 1, ..., K,$$
$$\hat{\rho}_l = \frac{1}{n}\sum_{j=1}^n \mathbb{P}_{q_2}(w_j = l),\ l = 1, ..., L.$$

Moreover, if $\mathbb{P}_{q_1}(z_i = k) \neq 0$ for all i and k, and $\mathbb{P}_{q_2}(w_j = l) \neq 0$ for all j and l, all maximizers are of the form $(\hat{\pi}, \hat{\rho}, e^{c_1}\hat{\theta}, e^{c_2}\hat{\lambda}, e^{-c_1-c_2}\hat{\mu})$, where c_1 and c_2 are two constants.

Proposition 1 indicates that for generic q_1 and q_2, the maximizer is unique up to two multiplicative scalars e^{c_1} and e^{c_2}. (Please refer to our proof on the uniqueness result for all possible q_1 and q_2.) For the convenience of the theoretical study in Section 4, we will choose

$$\hat{\theta}_i = \frac{d_i^r}{n\sqrt{D}},\ i = 1, ..., m,\ \hat{\lambda}_j = \frac{d_j^c}{m\sqrt{D}},\ j = 1, ..., n,\ \text{where } D = \frac{\sum_{ij} A_{ij}}{mn},$$

and adjust $\hat{\mu}$ accordingly.

3.3 E step

The E step concerns the computation of $J(q_1, q_2, \Phi)$, or equivalently, $\mathbb{P}_{q_1}(z_i = k)$ and $\mathbb{P}_{q_2}(w_j = l)$. At first glance, the number of terms in both quantities grow exponentially. The next proposition shows that both q_1 and q_2 can be factorized if the other one is fixed.

Proposition 2 *Define*

$$g_1(z_i) = -\sum_{j=1}^n \theta_i\lambda_j \left(\sum_{l=1}^L \mathbb{P}_{q_2}(w_j = l)\mu_{z_i l}\right) + \sum_{j=1}^n A_{ij}\left(\sum_{l=1}^L \mathbb{P}_{q_2}(w_j = l)\log\mu_{z_i l}\right) + \log\pi_{z_i},$$
$$i = 1, ..., m,$$
$$g_2(w_j) = -\sum_{i=1}^m \theta_i\lambda_j \left(\sum_{k=1}^K \mathbb{P}_{q_1}(z_i = k)\mu_{k w_j}\right) + \sum_{i=1}^m A_{ij}\left(\sum_{k=1}^K \mathbb{P}_{q_1}(z_i = k)\log\mu_{k w_j}\right) + \log\rho_{w_j},$$
$$j = 1, ..., n.$$

Given Φ and q_2,

$$\arg\max_{q_1} J(q_1, q_2, \Phi) = \prod_{i=1}^m \frac{e^{g_1(z_i)}}{\sum_{k=1}^K e^{g_1(k)}}. \tag{3}$$

Given Φ *and* q_1*,*

$$\arg\max_{q_2} J(q_1, q_2, \Phi) = \prod_{j=1}^{n} \frac{e^{g_2(w_j)}}{\sum_{l=1}^{L} e^{g_2(l)}}. \tag{4}$$

The factored form in (3) and (4) is the key reason that the variational EM is computationally feasible. It is worth emphasizing that unlike in the variational EM algorithm for symmetric networks (Daudin et al., 2008; Bickel et al., 2013), the factored form of q_1 and q_2 in the scenario of bipartite networks is not an assumption but a conclusion according to Proposition 2. The factored form for the classical LBM was proved by Govaert and Nadif (2008) and rediscovered by Wang et al. (2021a).

3.4 Initial values

As an iterative algorithm, the variational EM needs initial values to proceed. Our algorithm is designed to start from the M step. Therefore, we need to specify the initial values for q_1 and q_2. We use the widely-adopted spectral clustering method (Ng et al., 2002) on rows and columns respectively. Specifically, we carry out spectral clustering on AA^T and denote the estimated row cluster labels by $\hat{z}^{\text{init}}$. Let $q_1(z) \propto \prod_{i=1}^{m} 1(z_i = \hat{z}_i^{\text{init}})$. Similarly, carry out spectral clustering on A^TA and denote the estimated column labels by $\hat{w}^{\text{init}}$. Let $q_2(w) \propto \prod_{j=1}^{n} 1(w_j = \hat{w}_j^{\text{init}})$. Such initial values $\hat{z}^{\text{init}}$ and $\hat{w}^{\text{init}}$ have been used in the literature (Wang et al., 2021b, 2023). We faithfully implement the spectral clustering algorithm described in Ng et al. (2002). In particular, we normalize the embedded points (setting the norm equal to 1) before applying the k-means algorithm when conducting spectral clustering, as suggested in Ng et al. (2002) (Step 4 of the algorithm). This normalization typically aids in correcting for degree variation, in addition to the usage of the graph Laplacian. We summarize the variational EM in Algorithm 1.

4. Asymptotic properties

Brault et al. (2020) established the consistency of the estimators for the parameters in the classical LBM. They initially demonstrated the consistency of the maximum likelihood estimator (MLE) when cluster labels are observed. Furthermore, they showed that both the marginal likelihood and the variational approximation are asymptotically equivalent to the complete data likelihood. Consequently, this implies the consistency of the MLE and variational estimator of the parameters π, ρ, and μ.

We adopt a different approach to studying the DC-LBM. We focus on the consistency of clustering, that is, the consistency of q_1 and q_2. We prove the consistency and the convergence rate by showing that (2) converges uniformly to its population version and the population has a well-separated maximizer. The proof can handle degree parameters and necessitates weaker conditions. In particular, we allow the expected graph density goes to zero as long as both the average expected row and column degrees go to infinity, which is a typical condition for label consistency for symmetric networks (Bickel and Chen, 2009; Zhao et al., 2012).

Algorithm 1: Variational EM algorithm for the DC-LBM

Input: A, K, L;

Set $\hat{z}^{\text{init}}$ (resp. $\hat{w}^{\text{init}}$) be the outcome of spectral clustering on AA^T (resp. A^TA);

$q_1(z) \propto \prod_{i=1}^m 1(z_i = \hat{z}_i^{\text{init}})$; $q_2(w) \propto \prod_{j=1}^n 1(w_j = \hat{w}_j^{\text{init}})$;

$D = \sum_{ij} A_{ij}/(mn)$;

$\hat{\theta}_i = \sum_j A_{ij}/(n\sqrt{D}), i = 1, ..., m$; $\hat{\lambda}_j = \sum_i A_{ij}/(m\sqrt{D}), j = 1, ..., n$;

repeat

M step:

$$\hat{\mu}_{kl} = \frac{\sum_{i=1}^m \sum_{j=1}^n A_{ij} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)}{\sum_{i=1}^m \sum_{j=1}^n \hat{\theta}_i \hat{\lambda}_j \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)},\ k = 1, ..., K, l = 1, ..., L;$$

$$\hat{\pi}_k = \frac{1}{m}\sum_{i=1}^m \mathbb{P}_{q_1}(z_i = k),\ k = 1, ..., K; \quad \hat{\rho}_l = \frac{1}{n}\sum_{j=1}^n \mathbb{P}_{q_2}(w_j = l),\ l = 1, ..., L;$$

E step:

$$g_1(z_i) = -\sum_{j=1}^n \hat{\theta}_i \hat{\lambda}_j \left(\sum_{l=1}^L \mathbb{P}_{q_2}(w_j = l)\hat{\mu}_{z_i l}\right) + \sum_{j=1}^n A_{ij}\left(\sum_{l=1}^L \mathbb{P}_{q_2}(w_j = l)\log \hat{\mu}_{z_i l}\right) + \log \hat{\pi}_{z_i}, i = 1, ..., m;$$

$$q_1(z) = \prod_{i=1}^m \frac{e^{g_1(z_i)}}{\sum_{k=1}^K e^{g_1(k)}};$$

$$g_2(w_j) = -\sum_{i=1}^m \hat{\theta}_i \hat{\lambda}_j \left(\sum_{k=1}^K \mathbb{P}_{q_1}(z_i = k)\hat{\mu}_{k w_j}\right) + \sum_{i=1}^m A_{ij}\left(\sum_{k=1}^K \mathbb{P}_{q_1}(z_i = k)\log \hat{\mu}_{k w_j}\right) + \log \hat{\rho}_{w_j}, j = 1, ..., n;$$

$$q_2(w) = \prod_{j=1}^n \frac{e^{g_2(w_j)}}{\sum_{l=1}^L e^{g_2(l)}};$$

until *convergence*;

Output: q_1, q_2.

Let π^*, ρ^* be the true prior probabilities and z^*, w^* be the true row and column labels. According to the model assumption, the conditional expectation of A_{ij} has the form

$$E[A_{ij}|z_i^*, w_j^*] = \theta_i \lambda_j \mu_{z_i^* w_j^*}. \tag{5}$$

The parameters θ, λ and μ are clearly not identifiable. We therefore introduce the following canonical form for θ, λ and μ:

$$\begin{aligned}
\theta_i^* &= \frac{\frac{1}{n}\sum_{j=1}^n E[A_{ij}|z_i^*, w_j^*]}{(\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n E[A_{ij}|z_i^*, w_j^*])^{1/2}},\ i = 1, ..., m, \\
\lambda_j^* &= \frac{\frac{1}{m}\sum_{i=1}^m E[A_{ij}|z_i^*, w_j^*]}{(\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n E[A_{ij}|z_i^*, w_j^*])^{1/2}},\ j = 1, ..., n, \\
\mu_{kl}^* &= \frac{E[A_{ij}|z_i^*, w_j^*]}{\theta_i^* \lambda_j^*},\ \text{for some } i, j \text{ with } z_i^* = k, w_j^* = l,\ k = 1, ..., K, l = 1, ..., L.
\end{aligned} \tag{6}$$

We need to show that μ_{kl}^* are well defined; in other word, its value depends on row and column labels but not the row and column indices, which is given in the following proposition.

Proposition 3 *μ_{kl}^* defined in* (6) *depends on i and j only through their cluster labels $z_i^* = k$ and $w_j^* = l$.*

It is worth mentioning that, among the many possibilities for choosing canonical parameters, our selection is particularly convenient for the following reasons: The mean parameters $\mu_{kl}^* \asymp 1$ under this definition. Thus, the density of the network is fully characterized by θ_i^* and λ_j^* (Proposition 4). In the meantime, θ_i^* and λ_j^* can be directly and consistently estimated by (functions of) row and column degrees, without involving any unknown scale factor. Those properties facilitate the theoretical analysis (Theorem 6 and Theorem 9).

We make the following assumptions on the parameters throughout the theoretical analysis:

H_1 : $\pi_{\min} \le \pi_k^* \le \pi_{\max}$ $(k = 1, ..., K)$ and $\rho_{\min} \le \rho_l^* \le \rho_{\max}$ $(l = 1, ..., L)$, where $\pi_{\min}$, $\pi_{\max}$, $\rho_{\min}$ and $\rho_{\max}$ are positive constants. Furthermore, $\tilde{\pi}_{\min} \le (1/m)\sum_i 1(z_i^* = k) \le \tilde{\pi}_{\max}$ $(k = 1, ..., K)$ and $\tilde{\rho}_{\min} \le (1/n)\sum_j 1(w_j^* = l) \le \tilde{\rho}_{\max}$ $(l = 1, ..., L)$, where $\tilde{\pi}_{\min}$, $\tilde{\pi}_{\max}$, $\tilde{\rho}_{\min}$ and $\tilde{\rho}_{\max}$ are positive constants.

H_2 : $E[A_{ij}|z_i^*, w_j^*] = r_{mn}E_{ij}$ $(i = 1, ..., m,\ j = 1, ..., n)$ where $(mnr_{mn})/(m+n) \to \infty$ as $m, n \to \infty$, and $0 < E_{\min} \le E_{ij} \le E_{\max} < \infty$, where $E_{\min}$ and $E_{\max}$ are positive constants.

H_3 : Each row and each column of μ^* is unique. That is, there do not exist two rows k and k' such that $\mu_{kl}^* = \mu_{k'l}^*$ for all l, and there do not exist two columns l and l' such that $\mu_{kl}^* = \mu_{kl'}^*$ for all k.

All assumptions above are standard. Assumption H_1 ensures that no cluster size is too small. The second part of H_1 in fact automatically holds with high probability given the first part, which can be proved by applying Hoeffding's inequality (see Proposition 4.2 in Brault et al. (2020) for example). Here we directly assume the condition for simplicity. Assumption H_2 is an analogue of a typical assumption on graph density in many

works from the community detection literature (Bickel and Chen, 2009; Zhao et al., 2012). Specifically, the factor r_{mn} measures the rate of the expected graph density decay. Because $(mnr_{mn})/(m+n) \geq (mnr_{mn})/(2\max(m,n))$, $(mnr_{mn})/(m+n) \to \infty$ if and only if $nr_{mn} \to \infty$ and $mr_{mn} \to \infty$. That is, the average expected row and column degrees go to infinity. Assumption H_2 has the following implications on the canonical parameters.

Proposition 4 *Under Assumption H_2, for $i = 1, ..., m$, $\theta_{\min}\sqrt{r_{mn}} \leq \theta_i^* \leq \theta_{\max}\sqrt{r_{mn}}$, where $\theta_{\min}, \theta_{\max}$ are positive constants. Similarly, $\lambda_{\min}\sqrt{r_{mn}} \leq \lambda_j^* \leq \lambda_{\max}\sqrt{r_{mn}}$ for $j = 1, ..., n$, where $\lambda_{\min}, \lambda_{\max}$ are positive constants. Finally, $\mu_{\min} \leq \mu_{kl}^* \leq \mu_{\max}$, where $\mu_{\min}$ and $\mu_{\max}$ are positive constants.*

The proof is straightforward and hence is omitted.

Assumption H_3 has the same form of H_3 in Brault et al. (2020), which ensures μ^* is identifiable up to a permutation of the row and column labels. In the DC-LBM, this assumption is satisfied if no two rows (columns) of any μ that gives (5) are proportional to each other. The next proposition elaborates on the details.

Proposition 5 *H_3 holds if and only if for any parametrization (θ, λ, μ) satisfying (5), there do not exist two rows k and k' such that $\mu_{kl} = \mu_{k'l}a_{kk'}$ for all l and there do not exist two columns l and l' such that $\mu_{kl} = \mu_{kl'}b_{ll'}$ for all k.*

Our goal is to study the property of the variational approximation $J(q_1, q_2, \Phi)$. Proposition 2 shows that q_1 (resp. q_2) can be factorized if q_2 (resp. q_1) is given. We therefore study $J(q_1, q_2, \Phi)$ under the constraint that q_1 and q_2 are product measures, which can therefore be represented as matrices. Let $q^z = [q^z_{ik}]_{m\times K}$ and $q^w = [q^w_{jl}]_{n\times L}$ be $m \times K$ and $n \times L$ matrices with $q^z_{ik} = \mathbb{P}_{q_1}(z_i = k)$ and $q^w_{jl} = \mathbb{P}_{q_2}(w_j = l)$. We rewrite $J(q_1, q_2, \Phi)$ in terms of q^z, q^w:

$$\begin{aligned} J(q^z, q^w, \Phi) = & -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i\lambda_j\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q^z_{ik}q^w_{jl}\mu_{kl}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q^z_{ik}q^w_{jl}\log\mu_{kl}\right) \\ & + \sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}(\log\theta_i + \log\lambda_j) + \sum_{i=1}^{m}\left(\sum_{k=1}^{K}q^z_{ik}\log\pi_k\right) + \sum_{j=1}^{n}\left(\sum_{l=1}^{L}q^w_{jl}\log\rho_l\right) \\ & - \sum_{i=1}^{m}\sum_{k=1}^{K}q^z_{ik}\log q^z_{ik} - \sum_{j=1}^{n}\sum_{l=1}^{L}q^w_{jl}\log q^w_{jl}. \end{aligned}$$

Furthermore, we replace θ_i and λ_j in $J(q^z, q^w, \Phi)$ by the estimators derived in Proposition 1 since they remain unchanged throughout the algorithm:

$$\hat{\theta}_i = \frac{d^r_i}{\sqrt{D}},\ i = 1, ..., m,\ \hat{\lambda}_j = \frac{d^c_j}{\sqrt{D}},\ j = 1, ..., n,\ \text{where } D = \frac{\sum_{ij}A_{ij}}{mn}. \tag{7}$$

We exclude θ_i and λ_j from Φ thereafter, and omit the term $\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}(\log\theta_i + \log\lambda_j)$ in J because it does not affect the estimation of q^z and q^w when $\hat{\theta}$ and $\hat{\lambda}$ are fixed. Denote

the new criterion function by

$$\hat{J}(q^z, q^w, \Phi) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\theta}_i\hat{\lambda}_j\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q_{ik}^z q_{jl}^w \mu_{kl}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q_{ik}^z q_{jl}^w \log\mu_{kl}\right)$$
$$+\sum_{i=1}^{m}\left(\sum_{k=1}^{K}q_{ik}^z\log\pi_k\right) + \sum_{j=1}^{n}\left(\sum_{l=1}^{L}q_{jl}^w\log\rho_l\right) - \sum_{i=1}^{m}\sum_{k=1}^{K}q_{ik}^z\log q_{ik}^z - \sum_{j=1}^{n}\sum_{l=1}^{L}q_{jl}^w\log q_{jl}^w. \tag{8}$$

The next theorem shows $\hat{J}(q^z, q^w, \Phi)$ uniformly converges to its "population version" omitting lower order terms

$$\bar{J}(q^z, q^w, \mu) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i^*\lambda_j^*\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q_{ik}^z q_{jl}^w\mu_{kl}\right)$$
$$+\sum_{i=1}^{m}\sum_{j=1}^{n}E[A_{ij}|z^*, w^*]\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q_{ik}^z q_{jl}^w\log\mu_{kl}\right).$$

Theorem 6 *If $(mnr_{mn}\epsilon^2)/(m+n)\to\infty$ as $m, n\to\infty$, for ϵ being a positive constant or $o(1)$, then, under H_1 and H_2, we have*

$$\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \pi\in\mathcal{C}_\pi, \rho\in\mathcal{C}_\rho, \mu\in\mathcal{C}_\mu}\left|\hat{J}(q^z, q^w, \Phi) - \bar{J}(q^z, q^w, \mu)\right| \geq mnr_{mn}\epsilon \,\middle|\, z^*, w^*\right)\to 0,$$

where $\mathcal{C}_z$, $\mathcal{C}_w$, $\mathcal{C}_\pi$, $\mathcal{C}_\rho$, and $\mathcal{C}_\mu$ are the (compact) domains for the corresponding parameters. Specifically, $\mathcal{C}_z = \{q^z : q^z\in\mathbb{R}^{m\times K}, q_{ik}^z\in[0,1], \sum_{k=1}^K q_{ik}^z = 1\}$, $\mathcal{C}_w = \{q^w : q^w\in\mathbb{R}^{n\times L}, q_{jl}^w\in[0,1], \sum_{l=1}^L q_{jl}^w = 1\}$, $\mathcal{C}_\pi = \{\pi : \pi\in\mathbb{R}^K, \pi_k\in[\pi_{\min}, \pi_{\max}], \sum_{k=1}^K\pi_k = 1\}$, $\mathcal{C}_\rho = \{\rho : \rho\in\mathbb{R}^L, \rho_l\in[\rho_{\min}, \rho_{\max}], \sum_{l=1}^L\rho_l = 1\}$, and $\mathcal{C}_\mu = \{\mu : \mu\in\mathbb{R}^{K\times L}, \mu_{\min}\leq\mu_{kl}\leq\mu_{\max}\}$.

Next we state a result that the true labels (up to a permutation) are a well-separated maximizer of $\bar{J}(q^z, q^w, \mu)$. We give more definitions before proceeding.

Definition 7 (Soft confusion matrix) *For any row label assignment matrices q^z and $\tilde{q}^z$, let*

$$\mathbb{R}_{kk'}(q^z, \tilde{q}^z) = \frac{1}{m}\sum_{i=1}^{m}q_{ik}^z\tilde{q}_{ik'}^z.$$

In particular, the confusion matrix for the true row label z^* and q^z is

$$\mathbb{R}_{kk'}(1^{z^*}, q^z) = \frac{1}{m}\sum_{i=1}^{m}1_{ik}^{z^*}q_{ik'}^z,$$

where $1_{ik}^{z^*} = 1(z_i^* = k)$.

Similarly, for any column label assignments, let

$$\mathbb{R}_{ll'}(q^w, \tilde{q}^w) = \frac{1}{n}\sum_{j=1}^{n} q_{jl}^w \tilde{q}_{jl'}^w,\ \mathbb{R}_{ll'}(1^{w^*}, q^w) = \frac{1}{n}\sum_{j=1}^{n} 1_{jl}^{w^*} q_{jl'}^w,$$

where $1_{jl}^{w^*} = 1(w_j^* = l)$.

The soft confusion matrices defined above generalize the confusion matrix for comparing two hard label assignments to the case of comparing two probability matrices. Let S_K (S_L) be the set of permutations on $\{1, ..., K\}$ ($\{1, ..., L\}$). Taking permutations into account, the misclassification rates for row clusters and column clusters are defined as

$$M_{\text{row}}(q^z) = \min_{s\in S_K}\left(1 - \sum_{k'=1}^{K} \mathbb{R}_{s(k'),k'}(1^{z^*}, q^z)\right),$$
$$M_{\text{col}}(q^w) = \min_{t\in S_L}\left(1 - \sum_{l'=1}^{L} \mathbb{R}_{t(l'),l'}(1^{w^*}, q^w)\right).$$

The following theorem shows that $\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(q^z, q^w, \mu)$ is bounded below by $M_{\text{row}}(q^z)$ and $M_{\text{col}}(q^w)$.

Theorem 8 *For all $q^z \in \mathcal{C}_z$, $q^w \in \mathcal{C}_w$, and $\mu \in \mathcal{C}_\mu$, under H_1 and H_3, we have*

$$\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(q^z, q^w, \mu) \geq C_1 mnr_{mn} M_{\text{row}}(q^z),$$
$$\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(q^z, q^w, \mu) \geq C_2 mnr_{mn} M_{\text{col}}(q^w).$$

Finally, we give the rate of convergence of the misclassification rate, which implies label consistency. Let $\hat{\Phi} = (\hat{\mu}, \hat{\pi}, \hat{\rho})$ be a maximizer of $J(q^z, q^w, \Phi)$, that is,

$$(\hat{q}^z, \hat{q}^w, \hat{\Phi}) = \underset{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \mu\in\mathcal{C}_\mu, \pi\in\mathcal{C}_\pi, \rho\in\mathcal{C}_\rho}{\arg\max} \hat{J}(q^z, q^w, \Phi).$$

Theorem 9 *Assume H_1, H_2 and H_3. If $(mnr_{mn})/(m+n) \to \infty$ as $m, n \to \infty$, then for all positive constant δ, we have*

$$M_{\text{row}}(\hat{q}^z) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right),\ M_{\text{col}}(\hat{q}^w) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right).$$

Finally, we provide the rate of convergence of misclassification for networks with binary edges, parallel to the result in Theorem 9. Note that the closed-form solutions $\{\hat{\theta}_i\}$ and $\{\hat{\lambda}_i\}$, and thus the algorithm, rely on the form of the Poisson distribution. However, if we apply the same estimating procedure to a network with edges following Bernoulli distributions, we can establish the same convergence rate as in Theorem 9. Specifically, we assume that given labels, A_{ij}'s independently follow Bernoulli distributions with $E[A_{ij}|z_i^*, w_j^*] = \theta_i\lambda_j\mu_{z_i^* w_j^*}$. But we analyze the estimator from the Poisson model—that is, $\hat{\theta}_i$ and $\hat{\lambda}_j$ are computed by (7) and $\hat{\Phi} = (\hat{\mu}, \hat{\pi}, \hat{\rho})$ is still a maximizer of $J(q^z, q^w, \Phi)$, defined by (8). We obtain the following result.

Theorem 10 *Let A_{ij}'s independently follow Bernoulli distributions with $E[A_{ij}|z_i^*, w_j^*] = \theta_i\lambda_j\mu_{z_i^* w_j^*}$, given z^* and w^*. Assume the canonical parameters satisfy H_1, H_2 and H_3. If $(mnr_{mn})/(m+n) \to \infty$ as $m, n \to \infty$, then for all positive constant δ, we have*

$$M_{\text{row}}(\hat{q}^z) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right),\ M_{\text{col}}(\hat{q}^w) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right).$$

The proof is similar to that of Theorem 9. Note that all results, except for Theorem 6, only concern the mean parameters and therefore hold true under the Bernoulli model. The proof of Theorem 6 relies on a concentration inequality of Poisson variables (Canonne, 2019), which can be replaced by the Bernstein inequality of Bernoulli variables. We refer the readers to the appendix for details.

5. Simulation studies

In this section, we compare the proposed variational EM algorithm for the DC-LBM to two other methods, the profile likelihood based biclustering method (Flynn and Perry, 2020) and spectral clustering (Ng et al., 2002). The profile likelihood based biclustering extends the classical SBM to biclustering. The method assumes that $\{A_{ij}\}$ are sampled from distributions in an exponential family, which gives a flexible choice, such as Bernoulli, Poisson and Gaussian. However, the likelihood does not incorporate degree parameters. The model is identical to the classical LBM when assuming the Bernoulli distribution on $\{A_{ij}\}$. The profile likelihood based biclustering method treats cluster labels to be unknown fixed parameters and a local search technique based on the Kernighan-Lin heuristic (Kernighan and Lin, 1970) was applied to search the optimal row and column partitions. To reduce the possibility of the algorithm finding a local optimum, we use 30 random initial partitions in all simulation settings. We apply spectral clustering to biclustering in the same manner as described in Section 3.4—that is, carry out spectral clustering on AA^T to find the row labels and on A^TA to find the column labels.

We first evaluate the performance of the algorithms under the correctly-specified model for our proposed method—that is, the DC-LBM with the Poisson distribution. We simulate networks with the number of rows $m = 800$, the number of columns $n = 1000$, the number of row clusters $K = 3$, and the number of column clusters $L = 4$. Row cluster labels $\{z_i\}$ are independently generated from Multi$(1, \pi = (1/3, 1/3, 1/3)^T)$, and similarly, column cluster labels $\{w_j\}$ are independently generated from Multi$(1, \rho = (1/4, 1/4, 1/4, 1/4)^T)$. Degree parameters $\{\theta_i\}$ and $\{\lambda_j\}$ are independently generated from Uniform$(0.5, 1.5)$. For clarification, we will not use the prior information on $\{\theta_i\}$ and $\{\lambda_j\}$ in the algorithm. That is, we treat them as unknown fixed parameters. Furthermore, we set

$$\mu = r\begin{pmatrix} 0.15 & 0.05 & 0.05 & 0.06 \\ 0.05 & 0.15 & 0.05 & 0.08 \\ 0.05 & 0.05 & 0.15 & 0.10 \end{pmatrix}, \tag{9}$$

where r varies from 0.4, 0.6, 0.8 to 1, which controls the graph density.

We measure the accuracy of clustering by the adjusted Rand index (Vinh et al., 2010), which is a widely-used measure for comparing two partitions. The zero value of the index

corresponds to two independent partitions, and higher values indicate better agreement. All reported adjusted Rand index values in the figures below were based on 200 replicates. The error bars represent the range of plus or minus one standard deviation. To enhance readability, we jitter the error bars along the x-axis. However, all methods correspond to the same r values, namely $r = 0.4, 0.6, 0.8$, and 1.

(a) Row communities (b) Column communities

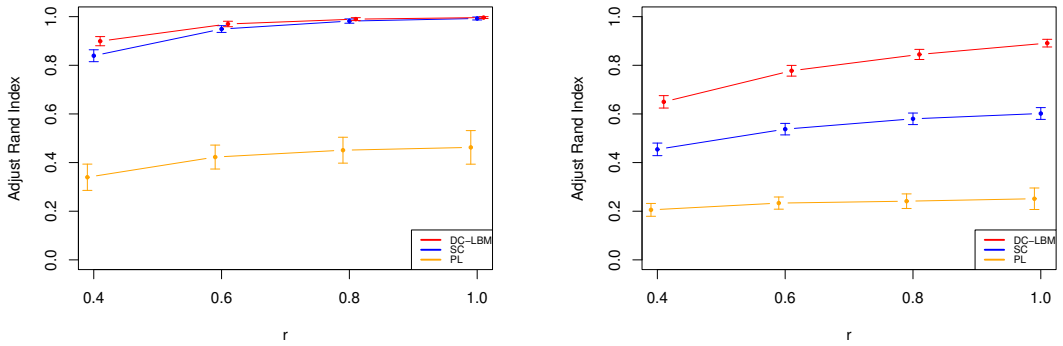


Figure 1: Performance of three algorithms under the DC-LBM with the Poisson distribution. r: the graph density factor in (9). Left panel: detection of row clusters. Right panel: detection of column clusters. SC: spectral clustering. PL: profile likelihood based biclustering.

Figure 1 shows the performance of three algorithms for detecting row and column clusters under the DC-LBM with the Poisson distribution. Our first observation is simply that the adjusted Rand index values for all methods improve as the graph density increases, which is in line with common sense in the community detection literature. Second, profile likelihood based biclustering gives the lowest adjusted Rand index values because the degree parameters are not considered by this method. Third, the performance of spectral clustering is between DC-LBM and profile likelihood based biclustering, because although not being a model-based approach, spectral clustering implicitly takes degree variation into account. The normalized Laplacian matrix includes the diagonal matrix whose elements are the row sums of AA^T (resp. A^TA). Moreover, the spectral clustering algorithm in Ng et al. (2002) renormalizes each row of the matrix whose columns are the K (resp. L) top eigenvectors of AA^T (resp. A^TA). This step alleviates the effect of degree variation. Lastly, the reason the improvement of the proposed method on row clustering is less substantial than the improvement on column clustering is that row clustering is easier in our setup, and therefore, has less room for improvement. It is noteworthy that all methods achieve better accuracy for row clustering than for column clustering in the simulations. The clustering problem on rows is easier according to the simulation setup for the following reasons: Firstly, the rows are grouped into a smaller number of clusters. Secondly, the estimated cluster label of row i (resp. column j) is mainly determined by the row vector $A_{i\cdot}$ (resp. the column vector

$A_{\cdot j}$), and a vector that contains more entries (in our case, the row vector) provides a clearer pattern. This is analogous to cluster analysis in a high-dimensional Euclidean space, where a large number of features aid in distinguish the clusters among the observations.

Our second simulation investigates how well the variational EM algorithm designed for the DC-LBM performs if the true model has equal degree parameters, that is, under the classical LBM with Poisson distribution. The model for this simulation is identical to the previous setup except that $\theta_i \equiv 1, \lambda_j \equiv 1, i = 1, ..., m, j = 1, ..., n$. From Figure 2, we can see that the adjusted Rand index values of the DC-LBM and of profile likelihood based biclustering are almost identical, which means that DC-LBM loses very little efficiency when introducing the extra parameters $\{\theta_i\}$ and $\{\lambda_j\}$, and can be safely used even if the true model is the classical LBM.

(a) Row communities (b) Column communities

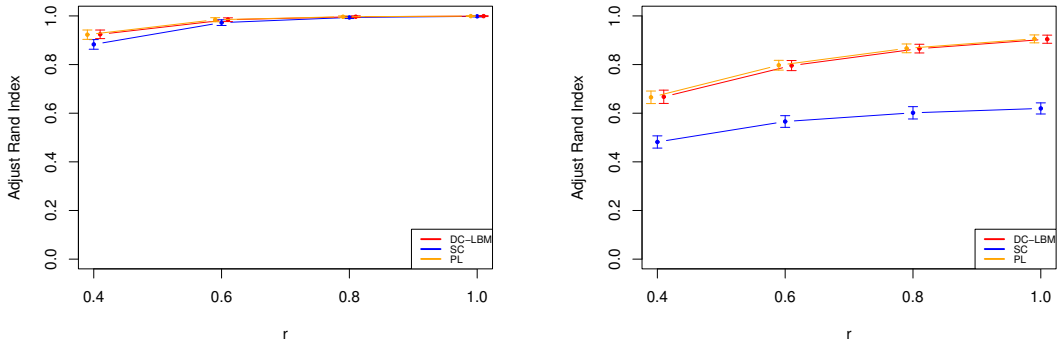


Figure 2: Performance of three algorithms under the classical LBM with the Poisson distribution. r: the graph density factor in (9). The red curve and green curve are almost identical. Left panel: detection of row clusters. Right panel: detection of column clusters. SC: spectral clustering. PL: profile likelihood based biclustering.

Additionally, we compare the CPU time for the three algorithms under both the classical LBM and the DC-LBM. The experiments are conducted on a high-performance computing cluster with fifth-generation Intel Core processors, and the results are reported in Figure 3. The recorded CPU time for the variational EM algorithm includes the time for spectral clustering, as it serves as the initial step in Algorithm 1. It can be seen that the variational EM algorithm costs less time than the profile likelihood based biclustering in the above simulations. Another notable pattern is that the running time for both the variational EM algorithm and the profile likelihood-based biclustering decreases as the graph density increases. This is because a network with a clearer community structure makes convergence easier for both algorithms.

Now, we evaluate the performance of the variational EM algorithm when the likelihood of A is misspecified. We assume a Poisson distribution on A_{ij} given the labels in our method, which is mainly due to the consideration of computation. In particular, $\{\hat{\theta}_i\}$ and $\{\hat{\lambda}_j\}$ in

(a) LBM, Poisson (b) DC-LBM, Poisson

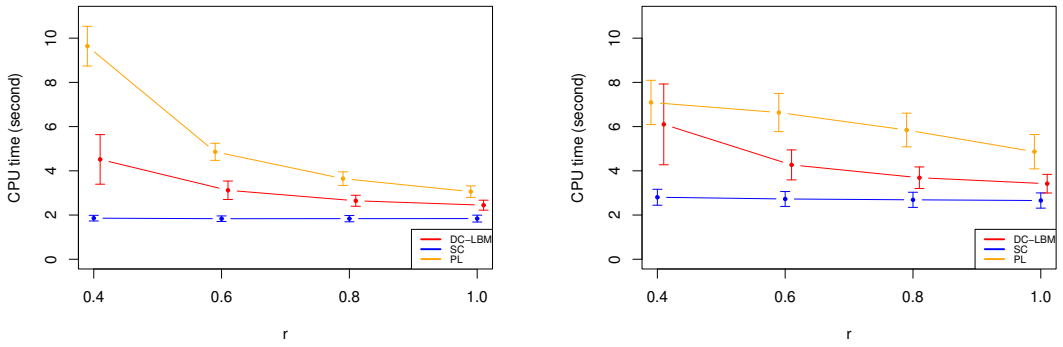


Figure 3: CPU time for three algorithms under Poisson models. r: the graph density factor in (9). Left panel: the true model is the classical LBM with the Poisson distribution. Right panel: the true model is the DC-LBM with the Poisson distribution.

(a) Row communities (b) Column communities

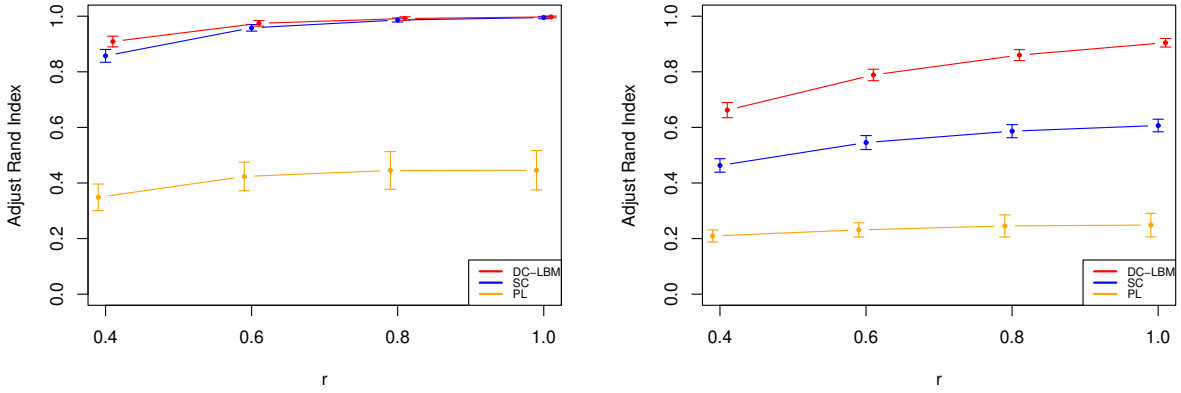


Figure 4: Performance of three algorithms under the DC-LBM with the Bernoulli distribution. r: the graph density factor in (9). Left panel: detection of row clusters. Right panel: detection of column clusters. SC: spectral clustering. PL: profile likelihood based biclustering.

the M step have a closed-form solution under the Poisson assumption, which is row and column degrees, respectively (Proposition 1). On the contrary, many real-world networks are unweighted graphs. Therefore, we investigate the behavior of the Poisson model when

(a) Row communities (b) Column communities

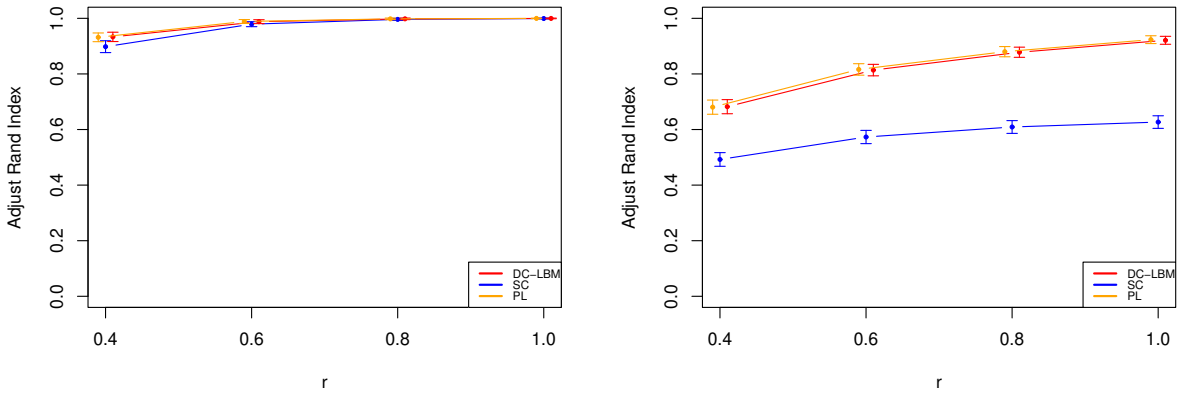


Figure 5: Performance of three algorithms under the classical LBM with the Bernoulli distribution. r: the graph density factor in (9). Left panel: detection of row clusters. Right panel: detection of column clusters. SC: spectral clustering. PL: profile likelihood based biclustering.

(a) LBM, Bernoulli (b) DC-LBM, Bernoulli

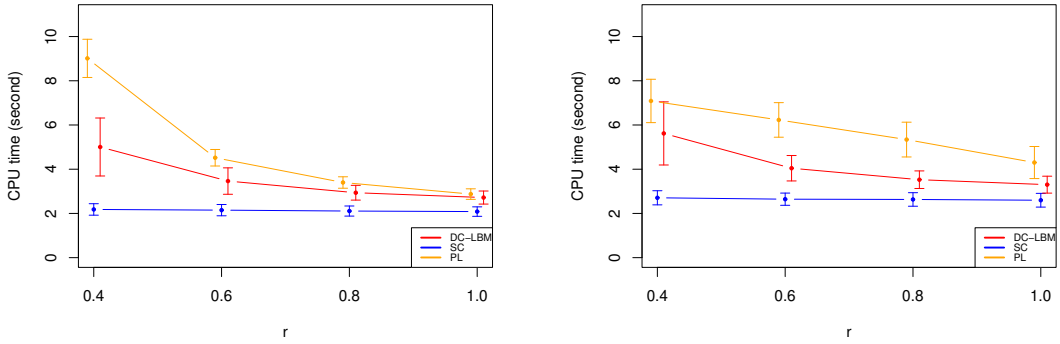


Figure 6: CPU time for three algorithms under Bernoulli models. r: the graph density factor in (9). Left panel: the true model is the classical LBM with the Bernoulli distribution. Right panel: the true model is the DC-LBM with the Bernoulli distribution.

$\{A_{ij}\}$ are actually Bernoulli variables. We carry out the two aforementioned simulations under the Bernoulli model. That is, the parameter settings are identical to the previous two simulations except that conditional on the cluster labels, $\{A_{ij}\}$ follow $\mathrm{Ber}(\theta_i\lambda_j\mu_{z_iw_j})$. The

algorithms for DC-LBM and spectral clustering are the same as before and profile likelihood based biclustering is implemented with the Bernoulli link function specified. Figure 4 and Figure 5 are identical to Figure 1 and Figure 2, which implies that the performance of the variational EM algorithm with the Poisson distribution is almost not affected by whether the true underlying distribution is Bernoulli or Poisson. Additionally, we report the CPU time for the three algorithms under both the classical LBM and the DC-LBM with the Bernoulli distribution in Figure 6. The pattern is again similar to that observed in the Poisson case (Figure 3).

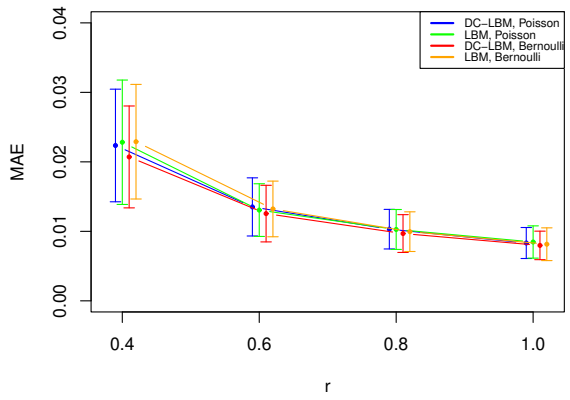


Figure 7: MAEs for estimates of μ by the variational EM algorithm. r: the graph density factor in (9).

Finally, we report the performance of the estimations of the block-wise mean parameters μ using the variational EM algorithm under the four scenarios, namely, DC-LBM and LBM with the Poisson distribution, and DC-LBM and LBM with the Bernoulli distribution. The estimates $\hat{\mu}$ in Algorithm 1 are for the canonical parameters μ^*. We therefore transform μ in our simulation setups to be the canonical parameters. Another benefit of this transformation is that the canonical parameters μ^* have the comparable scale under different density levels r. Moreover, note that the cluster labels are subject to permutations. Therefore, we choose the best match of the estimated row and column cluster labels with the true labels among all possible permutations and rearrange $\hat{\mu}_{kl}$ accordingly. We report the mean absolute deviations (MAEs) in Figure 7, where each point represents the MAE over 12 parameters and 200 replicates. The performance exhibits a consistent pattern across the four scenarios; that is, all MAEs decrease as the graph density grows. This demonstrates an improvement in the performance of the estimations of the canonical parameters with increasing graph density.

6. Application to MovieLens data

In this section we apply the proposed method to the well-known MovieLens data set (Harper and Konstan, 2015). The data was collected by a research team at the University of Minnesota through the MovieLens website (https://movielens.org/) during a seven-month period from September 19th, 1997 to April 22nd, 1998. The data set contains 100,000 ratings from 943 users on 1682 movies. Demographic information for the users were also available in the data set, which is not used in the present analysis. The goal of the analysis is to simultaneously identify the group structure of the users and of the movies with the expectation that it can reveal patterns of consumer behavior such as which group of users like to watch which types of movies.

We compare our method with PL that was applied to the same data set (Flynn and Perry, 2020). As in Flynn and Perry (2020), we constructed a 943-by-1682 binary matrix A where $A_{ij} = 1$ if user i has rated movie j and $A_{ij} = 0$ otherwise. We chose the number of user clusters $K = 3$ and the number of movie clusters $L = 4$ and ran PL with the Bernoulli link function and 250 random initial values, as described in Flynn and Perry (2020). The cluster numbers were chosen by Flynn and Perry (2020) based on the visualization of likelihoods in scree plots. We used the same cluster numbers, $K = 3$ and $L = 4$, when fitting the DC-LBM to the data for a fair comparison.

(a) PL (b) DL-LBM

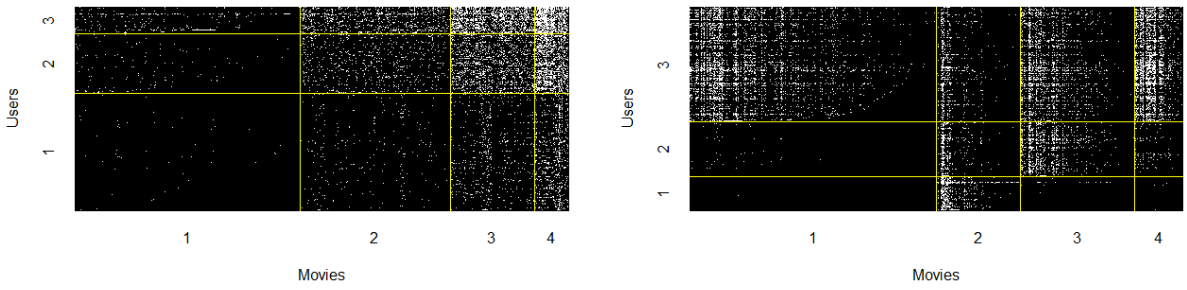
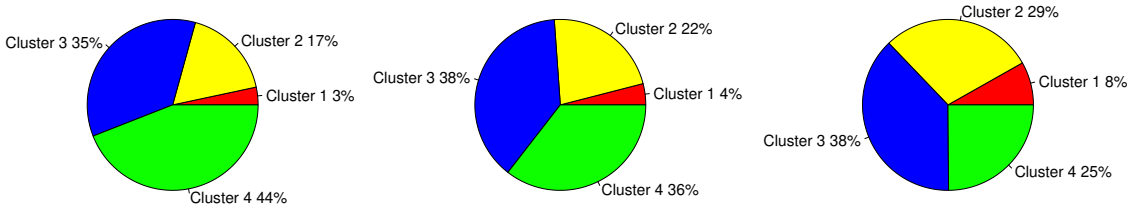


Figure 8: Heatmaps on the MovieLens matrix A with rows and columns rearranged by biclustering results. White pixels: $A_{ij} = 1$. Black pixels: $A_{ij} = 0$. Yellow lines: boundaries of estimated clusters.

Figure 8 presents the heatmaps of the data matrix with rows and columns rearranged based on the biclustering results from PL and DC-LBM, respectively. From the left panel, we can see that the label assignment by PL is largely dominated by marginal information on rows and columns, i.e., row and column degrees. The DC-LBM, by contrast, allows a higher of level of degree heterogeneity within cluster. The biclustering result by DC-LBM reveals certain patterns of consumer behavior—for example, users in cluster 1 almost only reviewed movies in cluster 2, and movies in clusters 1 were primarily reviewed by users in cluster 3.

The difference in user habits from different groups is better illustrated by the following pie charts in Figure 9. The pie charts show the percentages of movies in each movie cluster that are rated by user clusters identified by PL and DC-LBM, respectively. For example, panel (a) indicates that movies in cluster 4 by PL occupy 44% of total movies rated by user cluster 1. Pie charts for DC-LBM show a much more heterogeneous pattern on percentages across the three user clusters than the pie charts for PL. Specifically, user cluster 1 mainly rated movie cluster 2; user cluster 2 mainly rated movie cluster 3; user cluster 3 mainly rated movie clusters 1 and 4.

(a) User cluster 1 by PL (b) User cluster 2 by PL (c) User cluster 3 by PL



(d) User cluster 1 by DC-LBM (e) User cluster 2 by DC-LBM (f) User cluster 3 by DC-LBM

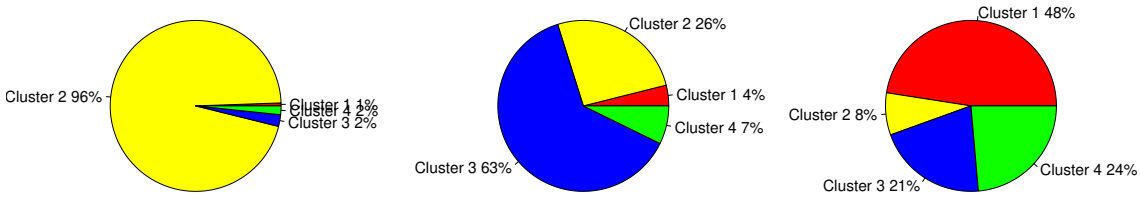


Figure 9: Pie charts showing the percentages of movies in each movie cluster that are rated by user clusters identified by PL and DC-LBM, respectively.

Furthermore, we compare the frequencies of degrees in user and movie clusters identified by PL and DC-LBM in Figure 10. As in the histograms, the degrees are much more homogeneous within a user or a movie cluster found by PL than in the clusters found by DC-LBM, which is in line with the observation from Figure 8.

Finally, we study whether the estimated movie clusters are associated with the true movie categories provided in the MovieLens data set. The 1682 movies in the data set were labeled with 19 categories such as "Action" or "Romance" and many of the movies belong to multiple categories. A direct comparison between the estimated clusters and the

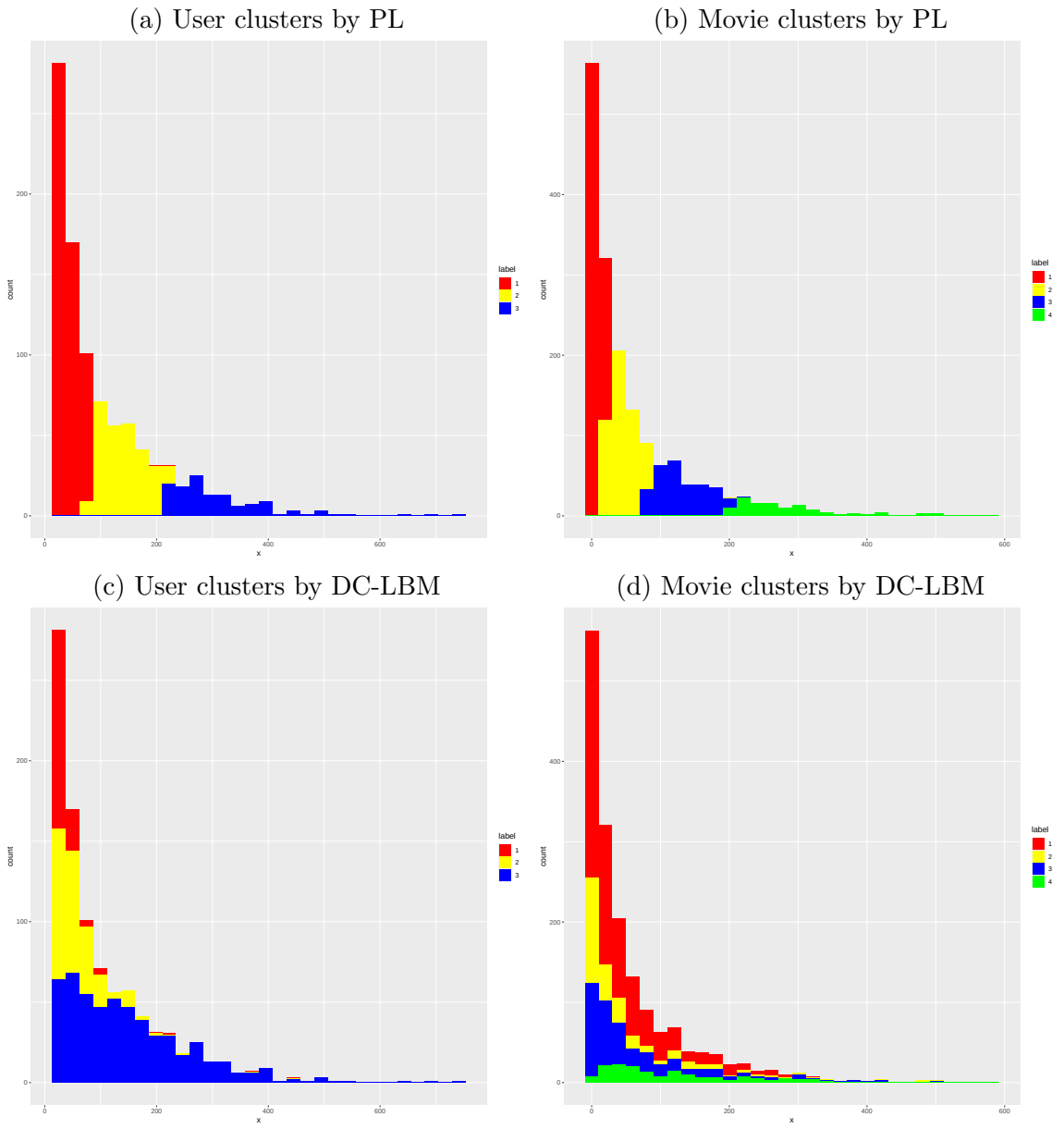


Figure 10: Histograms of degrees in user and movie clusters identified by PL and DC-LBM, respectively.

true categories is difficult due to the relatively large number of movie categories and the overlaps. We instead construct contingency tables with row categories being the estimated movie clusters and column categories being the ground truth and evaluate how much the table deviates from an independence model. Specifically, we filtered the data to only include movies belonging to a single category, resulting 833 movies, and constructed contingency tables for PL and DC-LBM, respectively. We then ran the chi-squared test of independence

on the two tables. The p-values for the tables constructed from clusters estimated by PL and DC-LBM are 0.0415 and 2.66×10^{-7}, respectively, which suggests that the movie clusters estimated by DC-LBM have a stronger association with the true movie categories.

7. Conclusion

In this paper, we proposed a degree-corrected latent block model (DC-LBM) for biclustering in bipartite networks. By introducing additional parameters to characterize row and column degrees, we achieved significant improvements in biclustering results compared to the classical LBM on both simulated and real-world data sets. We demonstrated that under the Poisson assumption, the maximizer of the variational approximation corresponds exactly to the row and column degrees. Furthermore, we established the label consistency and the convergence rate of the variational estimator under the DC-LBM with the Bernoulli and Poisson distributions, allowing for the expected graph density to approach zero as the average expected degrees of rows and columns go to infinity. For networks with weighted edges of more general types, label consistency is expected if the weights are nonnegative and bounded, although the assumption on graph densities may change depending on the variance of the weights. In the more general case, new concentration inequalities need to be developed (to replace (11) and the Bernstein inequality) to establish label consistency.

The proposed method can also be applied to clustering in directed networks, where row and column cluster labels are assumed to be distinct, capturing different behaviors of nodes as link senders and receivers. If the row and column cluster labels need to be identical, an additional step after the E step can be incorporated to enforce label probability assignment identity between rows and columns, similar to the split likelihood method (Wang et al., 2021a).

There are several directions for future work. One interesting area is the selection of the number of clusters in bipartite networks. In recent years, significant progress has been made on the selection of communities in networks (Saldana et al., 2017; Wang and Bickel, 2017; Hu et al., 2019; Ma et al., 2021; Le and Levina, 2022; Watanabe and Suzuki, 2021). It is interesting to explore how to adapt these methods to DC-LBM. Additionally, we would like to explore the generalization of the DC-LBM to other clustering problems within the context of bipartite networks, such as estimating mixed memberships (Airoldi et al., 2008; Jin et al., 2017; Zhang et al., 2020) and incorporating node features (Zhang et al., 2016; Zhao et al., 2019). Furthermore, the theory in the paper guarantees that the global optimizer of (8) is consistent with the true cluster labels. However, the consistency of the solution produced by the proposed variational EM algorithm remains an open problem. Recent rigorous studies have delved into the consistency of EM algorithms (Balakrishnan et al., 2017) and the K-means algorithm (Lu and Zhou, 2016) within the context of classical cluster analysis. In the domain of network community detection, Amini et al. (2013) established the consistency of the EM algorithm for a pseudo-likelihood in the Stochastic Block Model (SBM) with two communities. This method of proof has subsequently been adapted in Zhao et al. (2019); Wang et al. (2021a, 2023). We plan to explore how to prove the consistency of EM algorithms to networks with more than two communities and with degree parameters in future work.

Acknowledgments

This work was supported by the National Science Foundation Grant, DMS-2245380 (YZ), DMS-2245381 (NH), DMS-1722691 (NH), DMS-1821243 (JZ) and DMS-2210439 (JZ), and Simons Foundation Grant 524432 (NH).

Appendix A. Proofs

We give technical proofs in this section.

A.1 Proof of Proposition 1

Up to a constant independent of Φ,

$$
\begin{aligned}
& J(q_1, q_2, \Phi) \\
= & \sum_{z \in \Omega_z} \sum_{w \in \Omega_w} q_1(z) q_2(w) \left(- \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_i \lambda_j \left(\sum_{k=1}^{K} \sum_{l=1}^{L} 1(z_i = k) 1(w_j = l) \mu_{kl} \right) \right. \\
& + \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \left(\sum_{k=1}^{K} \sum_{l=1}^{L} 1(z_i = k) 1(w_j = l) \log \mu_{kl} \right) + \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} (\log \theta_i + \log \lambda_j) \\
& \left. + \sum_{i=1}^{m} \sum_{k=1}^{K} 1(z_i = k) \log \pi_k + \sum_{j=1}^{n} \sum_{l=1}^{L} 1(w_j = l) \log \rho_l \right) \\
= & - \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_i \lambda_j \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k) \mathbb{P}_{q_2}(w_j = l) \mu_{kl} \right) + \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} (\log \theta_i + \log \lambda_j) \\
& + \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k) \mathbb{P}_{q_2}(w_j = l) \log \mu_{kl} \right) \\
& + \sum_{i=1}^{m} \left(\sum_{k=1}^{K} \mathbb{P}_{q_1}(z_i = k) \log \pi_k \right) + \sum_{j=1}^{n} \left(\sum_{l=1}^{L} \mathbb{P}_{q_2}(w_j = l) \log \rho_l \right).
\end{aligned}
$$

The maximization of π and ρ is trivial.

We first prove $(\hat{\theta}, \hat{\lambda}, \hat{\mu})$ defined in the proposition is a stationary point for any q_1, q_2.

$$
\begin{aligned}
&\frac{\partial J}{\partial \theta_i}\bigg|_{\hat{\theta},\hat{\lambda},\hat{\mu}} \\
&= -\sum_{j=1}^{n} \hat{\lambda}_j \left(\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)\hat{\mu}_{kl} \right) + \frac{\sum_{j=1}^{n} A_{ij}}{\hat{\theta}_i} \\
&= -\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k) \left(\sum_{j=1}^{n} d_j^c \mathbb{P}_{q_2}(w_j = l) \right) \frac{\sum_{i'=1}^{m}\sum_{j'=1}^{n} A_{i'j'}\mathbb{P}_{q_1}(z_{i'} = k)\mathbb{P}_{q_2}(w_{j'} = l)}{\sum_{i'=1}^{m}\sum_{j'=1}^{n} d_{i'}^r d_{j'}^c \mathbb{P}_{q_1}(z_{i'} = k)\mathbb{P}_{q_2}(w_{j'} = l)} + 1 \\
&= -\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k) \frac{\sum_{i'=1}^{m}\sum_{j'=1}^{n} A_{i'j'}\mathbb{P}_{q_1}(z_{i'} = k)\mathbb{P}_{q_2}(w_{j'} = l)}{\sum_{i'=1}^{m} d_{i'}^r \mathbb{P}_{q_1}(z_{i'} = k)} + 1 \\
&= -\sum_{k=1}^{K} \mathbb{P}_{q_1}(z_i = k) \frac{\sum_{i'=1}^{m}\sum_{j'=1}^{n} A_{i'j'}\mathbb{P}_{q_1}(z_{i'} = k)\sum_{l=1}^{L}\mathbb{P}_{q_2}(w_{j'} = l)}{\sum_{i'=1}^{m} d_{i'}^r \mathbb{P}_{q_1}(z_{i'} = k)} + 1 \\
&= -\sum_{k=1}^{K} \mathbb{P}_{q_1}(z_i = k) \frac{\sum_{i'=1}^{m}(\sum_{j'=1}^{n} A_{i'j'})\mathbb{P}_{q_1}(z_{i'} = k)}{\sum_{i'=1}^{m} d_{i'}^r \mathbb{P}_{q_1}(z_{i'} = k)} + 1 \\
&= -\sum_{k=1}^{K} \mathbb{P}_{q_1}(z_i = k) + 1 = 0,\ i = 1, ..., m.
\end{aligned}
$$

Similarly,

$$
\frac{\partial J}{\partial \lambda_j}\bigg|_{\hat{\theta},\hat{\lambda},\hat{\mu}} = 0,\ j = 1, ..., n.
$$

And it is easy to check that

$$
\frac{\partial J}{\partial \mu_{kl}}\bigg|_{\hat{\theta},\hat{\lambda},\hat{\mu}} = 0,\ k = 1, ..., K,\ l = 1, ..., L.
$$

Let $\alpha_i = \log \theta_i, \beta_j = \log \lambda_j$ and $\gamma_{kl} = \log \mu_{kl}$. Then as a function of α, β and γ, $J(q_1, q_2, \Phi)$ has the form (omitting the last two terms that depend on only π and ρ)

$$
\begin{aligned}
J(q_1, q_2, \Phi) = &-\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)\exp(\alpha_i + \beta_j + \gamma_{kl}) \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}(\alpha_i + \beta_j) + \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij} \left(\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)\gamma_{kl} \right).
\end{aligned} \tag{10}
$$

It is easy to see $J(q_1, q_2, \Phi)$ is concave by noticing that $\exp(\alpha_i + \beta_j + \gamma_{kl})$ is convex by definition and the last two terms in $J(q_1, q_2, \Phi)$ are linear.

By the chain rule, $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ with $\hat{\alpha}_i = \log \hat{\theta}_i, \hat{\beta}_j = \log \hat{\lambda}_j$ and $\hat{\gamma}_{kl} = \log \hat{\mu}_{kl}$ is also a stationary point of $J(q_1, q_2, \Phi)$, and therefore a global maximizer by concavity. This implies $(\hat{\theta}, \hat{\lambda}, \hat{\mu})$ is a global maximizer.

Now we work on the uniqueness of the maximizer. For a maximizer of (10), say, $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, it is straightforward to check $(\hat{\alpha} + c_1 1_m, \hat{\beta} + c_2 1_n, \hat{\gamma} + 1_K 1_L^T)$ is also a maximizer, for any constant c_1 and c_2. Here 1_m is a m-dimensional vector with all entries equal to 1. The argument below implies all maximizers are of the form $(\hat{\alpha} + c_1 1_m, \hat{\beta} + c_2 1_n, \hat{\gamma} + 1_K 1_L^T)$ if $\mathbb{P}_{q_1}(z_i = k) \neq 0$ for all i and k, and $\mathbb{P}_{q_2}(w_j = l) \neq 0$ for all j and l. In terms of original parametrization, all maximizers are of the form $(e^{c_1}\hat{\theta}, e^{c_2}\hat{\lambda}, e^{-c_1-c_2}\hat{\mu})$.

Consider two maximizers $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ of (10). For i and i', if there is a k such that $\mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_1}(z_{i'} = k) \neq 0$, we show $\hat{\alpha}_i - \tilde{\alpha}_i = \hat{\alpha}_{i'} - \tilde{\alpha}_{i'}$. First find a j and an l with $\mathbb{P}_{q_2}(w_j = l) \neq 0$. By Lemma 11 below, we have $\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{kl} = \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\gamma}_{kl}$ and $\hat{\alpha}_{i'} + \hat{\beta}_j + \hat{\gamma}_{kl} = \tilde{\alpha}_{i'} + \tilde{\beta}_j + \tilde{\gamma}_{kl}$, which implies $\hat{\alpha}_i - \tilde{\alpha}_i = \hat{\alpha}_{i'} - \tilde{\alpha}_{i'}$. We can make a similar conclusion for $\hat{\beta}$ and $\tilde{\beta}$. As a consequence, if $\mathbb{P}_{q_1}(z_i = k) \neq 0$ for all i and k, $\hat{\alpha}_i - \tilde{\alpha}_i$ is constant for all i; if $\mathbb{P}_{q_2}(w_j = l) \neq 0$ for all j and l, $\hat{\beta}_j - \tilde{\beta}_j$ is constant for all j. If there is a partition of the index set $\{1, ..., m\} = \coprod_{s=1}^{S} \mathcal{C}_s$ such that (i) $\mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_1}(z_i' = k) = 0$ for all k whenever $i \in \mathcal{C}_s$ and $i' \in \mathcal{C}_{s'}$ with $s \neq s'$; (ii) for $i, i' \in \mathcal{C}_s$, there is a k such that $\mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_1}(z_i' = k) \neq 0$; then we have $\hat{\alpha}_i - \tilde{\alpha}_i = d_s$ when $i \in \mathcal{C}_s$. That is, $\hat{\alpha}_i - \tilde{\alpha}_i$ can take S different values based on the membership of i with respect to the partition. We can make a similar conclusion for $\hat{\beta}$ and $\tilde{\beta}$.

A.2 Proof of Proposition 2

We first prove a lemma.

Lemma 11 *For any two maximizers $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ of (10), we have $\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{kl} = \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\gamma}_{kl}$ when $\mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l) \neq 0$.*

Proof By concavity of J, any vector in the segment connecting two maximizers $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ and $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ is also a maximizer. That is, $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) + t(\tilde{\alpha} - \hat{\alpha}, \tilde{\beta} - \hat{\beta}, \tilde{\gamma} - \hat{\gamma})$ is a maximizer for $0 \leq t \leq 1$. This implies J is a constant on the segment. Define $\check{J}(t) = J(q_1, q_2, \hat{\Phi} + t(\tilde{\Phi} - \hat{\Phi}))$. Therefore, $\check{J}$ is constant over $[0, 1]$ and $\check{J}''|_{t=0} = 0$. In fact, we have

$$\check{J}''|_{t=0} = \\ -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{kl} - \tilde{\alpha}_i - \tilde{\beta}_j - \tilde{\gamma}_{kl})^2 \exp(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{kl})$$

$\check{J}''|_{t=0} = 0$ implies $\mathbb{P}_{q_1}(z_i = k)\mathbb{P}_{q_2}(w_j = l)(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{kl} - \tilde{\alpha}_i - \tilde{\beta}_j - \tilde{\gamma}_{kl})^2 = 0$ for all i, j, k, l, which leads to the conclusion. ■

We now prove Proposition 2. We only prove the result for q_1 and the other part is similar. Recall that

$$g_1(z_i) = -\sum_{j=1}^{n} \theta_i \lambda_j \left(\sum_{l=1}^{L} \mathbb{P}_{q_2}(w_j = l)\mu_{z_i l} \right) + \sum_{j=1}^{n} A_{ij} \left(\sum_{l=1}^{L} \mathbb{P}_{q_2}(w_j = l) \log \mu_{z_i l} \right) + \log \pi_{z_i}, \\ i = 1, ..., m.$$

Up to a constant independent of q_1,

$$\begin{aligned}
J(q_1, q_2, \Phi) &= \sum_{z\in\Omega_z}\sum_{w\in\Omega_w} q_1(z)q_2(w)\log P(A, z, w) - \sum_{z\in\Omega_z} q_1(z)\log q_1(z) \\
&= \sum_{z\in\Omega_z}\sum_{w\in\Omega_w} q_1(z)q_2(w)\left(-\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i\lambda_j\left(\sum_{l=1}^{L}1(w_j=l)\mu_{z_i l}\right)\right. \\
&\quad \left. +\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}\left(\sum_{l=1}^{L}1(w_j=l)\log\mu_{z_i l}\right)+\sum_{i=1}^{m}\log\pi_{z_i}\right) - \sum_{z\in\Omega_z} q_1(z)\log q_1(z) \\
&= \sum_{z\in\Omega_z} q_1(z)\left(-\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i\lambda_j\left(\sum_{l=1}^{L}\mathbb{P}_{q_2}(w_j=l)\mu_{z_i l}\right)\right. \\
&\quad \left. +\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}\left(\sum_{l=1}^{L}\mathbb{P}_{q_2}(w_j=l)\log\mu_{z_i l}\right)+\sum_{i=1}^{m}\log\pi_{z_i}\right) - \sum_{z\in\Omega_z} q_1(z)\log q_1(z) \\
&= \sum_{z\in\Omega_z} q_1(z)\sum_{i=1}^{m} g_1(z_i) - \sum_{z\in\Omega_z} q_1(z)\log q_1(z) \\
&= \sum_{z\in\Omega_z} q_1(z)\log\left\{\frac{\prod_{i=1}^{m} e^{g_1(z_i)}}{q_1(z)}\right\} \\
&\leq \log\prod_{i=1}^{m}\left(\sum_{z_i=1}^{K} e^{g_1(z_i)}\right),
\end{aligned}$$

where the last inequality is Jensen's inequality and equality holds if and only if

$$q_1(z) = \prod_{i=1}^{m}\frac{e^{g_1(z_i)}}{\sum_{k=1}^{K} e^{g_1(k)}}.$$

A.3 Proof of Proposition 3

Let θ, λ and μ be a set of arbitrarily chosen parameters that gives (5). Then

$$\begin{aligned}
\theta_i^* &= \theta_i \frac{\frac{1}{n}\sum_{j=1}^{n}\lambda_j\mu_{kw_j^*}}{\sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}E[A_{ij}|z_i^*, w_j^*]}}, \text{ for } z_i^* = k, \\
\lambda_j^* &= \lambda_j \frac{\frac{1}{m}\sum_{i=1}^{m}\theta_i\mu_{z_i^* l}}{\sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}E[A_{ij}|z_i^*, w_j^*]}}, \text{ for } w_j^* = l, \\
\mu_{kl}^* &= \mu_{kl}\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}E[A_{ij}|z_i^*, w_j^*]}{\left(\sum_{j=1}^{n}\lambda_j\mu_{kw_j^*}\right)\left(\sum_{i=1}^{m}\theta_i\mu_{z_i^* l}\right)}.
\end{aligned}$$

The proposition immediately follows by noticing $\sum_{j=1}^{n}\lambda_j\mu_{kw_j^*}$ depends only on k and $\sum_{i=1}^{m}\theta_i\mu_{z_i^* l}$ depends only on l.

A.4 Proof of Proposition 5

Let θ, λ and μ be a set of arbitrarily chosen parameters that gives (5). According to Proposition 3,

$$\begin{aligned}\theta_i^* &= \theta_i \alpha_k, \text{ for } z_i^* = k,\\ \lambda_j^* &= \lambda_j \beta_l, \text{ for } w_j^* = l,\\ \mu_{kl}^* &= \frac{\mu_{kl}}{\alpha_k \beta_l},\end{aligned}$$

where

$$\begin{aligned}\alpha_k &= \frac{\frac{1}{n}\sum_{j=1}^n \lambda_j \mu_{k w_j^*}}{\sqrt{\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n E[A_{ij}|z_i^*, w_j^*]}},\\ \beta_l &= \frac{\frac{1}{m}\sum_{i=1}^m \theta_i \mu_{z_i^* l}}{\sqrt{\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n E[A_{ij}|z_i^*, w_j^*]}}.\end{aligned}$$

Now we show that $\mu_{kl}^* = \mu_{k'l}^*$ for all l if and only if $\mu_{kl}/\mu_{k'l}$ is constant for all l, which leads to the conclusion of this proposition. On one hand,

$$\mu_{kl}^* = \mu_{k'l}^* \Rightarrow \frac{\mu_{kl}}{\alpha_k \beta_l} = \frac{\mu_{k'l}}{\alpha_k' \beta_l} \Rightarrow \frac{\mu_{kl}}{\mu_{k'l}} = \frac{\alpha_k}{\alpha_{k'}}, \text{ for all } l.$$

On the other hand, if $\mu_{kl}/\mu_{k'l} = a_{kk'}$ for all l, it is straightforward to check $\alpha_k/\alpha_{k'} = a_{kk'}$, which implies $\mu_{kl}^* = \mu_{k'l}^*$.

A.5 Proof of Theorem 6

We first define a few quantities. Let

$$\begin{aligned}J_1(q^z, q^w, \mu) &= \sum_{i=1}^m \sum_{j=1}^n A_{ij} \left(\sum_{k=1}^K \sum_{l=1}^L q_{ik}^z q_{jl}^w \log \mu_{kl} \right),\\ \bar{J}_1(q^z, q^w, \mu) &= \sum_{i=1}^m \sum_{j=1}^n E[A_{ij}|z^*, w^*] \left(\sum_{k=1}^K \sum_{l=1}^L q_{ik}^z q_{jl}^w \log \mu_{kl} \right),\\ J_2(q^z, q^w, \mu) &= -\sum_{i=1}^m \sum_{j=1}^n \hat{\theta}_i \hat{\lambda}_j \left(\sum_{k=1}^K \sum_{l=1}^L q_{ik}^z q_{jl}^w \mu_{kl} \right),\\ \bar{J}_2(q^z, q^w, \mu) &= -\sum_{i=1}^m \sum_{j=1}^n \theta_i^* \lambda_j^* \left(\sum_{k=1}^K \sum_{l=1}^L q_{ik}^z q_{jl}^w \mu_{kl} \right),\\ J_3(q^z, q^w, \pi, \rho) &= \sum_{i=1}^m \left(\sum_{k=1}^K q_{ik}^z \log \pi_k \right) + \sum_{j=1}^n \left(\sum_{l=1}^L q_{jl}^w \log \rho_l \right)\\ &\quad - \sum_{i=1}^m \sum_{k=1}^K q_{ik}^z \log q_{ik}^z - \sum_{j=1}^n \sum_{l=1}^L q_{jl}^w \log q_{jl}^w.\end{aligned}$$

It is easy to check that

$$\hat{J}(q^z, q^w, \Phi) = J_1(q^z, q^w, \mu) + J_2(q^z, q^w, \mu) + J_3(q^z, q^w, \pi, \rho),$$
$$\bar{J}(q^z, q^w, \mu) = \bar{J}_1(q^z, q^w, \mu) + \bar{J}_2(q^z, q^w, \mu).$$

Lemma 12 *Let $\{X_{ij}\}$ be independent Poisson variables with mean[2] $E[X_{ij}] \leq r_{mn}C$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left(\max_{0\leq u_i\leq 1, i=1,\ldots,m, 0\leq v_j\leq 1, j=1,\ldots,n} \left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right| \geq mnr_{mn}\epsilon\right) \leq 2^{m+n+1}\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C,\epsilon)}\right).$$

Proof First note that

$$\max_{0\leq u_i\leq 1, i=1,\ldots,m, 0\leq v_j\leq 1, j=1,\ldots,n} \left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right|$$
$$= \max_{u_i\in\{0,1\}, i=1,\ldots,m, v_j\in\{0,1\}, j=1,\ldots,n} \left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right|.$$

To prove this, let $f(u,v) = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j$. If the maximum occurs at (u,v) with $u_1 \in (0,1)$, then consider u' and u'' with $u'_1 = 0$, $u''_1 = 1$, and rest of entries identical to that of u. It is easy to check $f(u,v) = (1-u_1)f(u',v) + u_1 f(u'',v)$. So the value of $f(u,v)$ must be between $f(u',v)$ and $f(u'',v)$. This implies that both (u',v) and (u'',v) are also maximizers. We can consider (u',v) instead of (u,v). If there are other entries of (u',v) strictly between 0 and 1, we can continue this argument until we find a maximizer with all entries equal to 0 or 1.

Then according to Canonne (2019), for $u_i \in \{0,1\}, i = 1,...,m, v_j \in \{0,1\}, j = 1,...,n$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right| \geq mnr_{mn}\epsilon\right)$$
$$\leq 2\exp\left(-\frac{m^2n^2r_{mn}^2\epsilon^2}{2(\sum_{ij}E[X_{ij}] + mnr_{mn}\epsilon)}\right) \leq 2\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C,\epsilon)}\right). \quad (11)$$

2. All constants, such as C_1 and C_2, are defined locally. This means that the constants in different lemmas can vary.

It follows that

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\le u_i\le 1, i=1,\ldots,m, 0\le v_j\le 1, j=1,\ldots,n}\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-E[X_{ij}])u_iv_j\right|\ge mnr_{mn}\epsilon\right)\\
&=\mathbb{P}\left(\max_{u_i\in\{0,1\}, i=1,\ldots,m, v_j\in\{0,1\}, j=1,\ldots,n}\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-E[X_{ij}])u_iv_j\right|\ge mnr_{mn}\epsilon\right)\\
&\le 2^{m+n}2\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C,\epsilon)}\right).
\end{aligned}$$

■

Lemma 13 *For sufficiently small positive ϵ,*

$$\begin{aligned}
&\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \mu\in\mathcal{C}_\mu}\left|J_1(q^z,q^w,\mu)-\bar{J}_1(q^z,q^w,\mu)\right|\ge mnr_{mn}\epsilon\,\middle|\, z^*,w^*\right)\\
&\qquad\le C_1 2^{m+n}\exp\left(-C_2 mnr_{mn}\epsilon^2\right).
\end{aligned}$$

Proof Note that

$$\begin{aligned}
&\left|J_1(q^z,q^w,\mu)-\bar{J}_1(q^z,q^w,\mu)\right|\\
=&\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-E[A_{ij}|z^*,w^*])\left(\sum_{k=1}^{K}\sum_{l=1}^{L}q_{ik}^z q_{jl}^w\log\mu_{kl}\right)\right|\\
=&\left|\sum_{k=1}^{K}\sum_{l=1}^{L}\log\mu_{kl}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z q_{jl}^w\right)\right|\\
\le&\sum_{k=1}^{K}\sum_{l=1}^{L}|\log\mu_{kl}|\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z q_{jl}^w\right|\\
\le&\max\{|\log\mu_{\min}|,|\log\mu_{\max}|\}\sum_{k=1}^{K}\sum_{l=1}^{L}\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z q_{jl}^w\right|.
\end{aligned}$$

The lemma follows immediately from Lemma 12. ■

Lemma 14 *For sufficiently small positive $\epsilon>0$,*

$$\begin{aligned}
&\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \mu\in\mathcal{C}_\mu}\left|J_2(q^z,q^w,\mu)-\bar{J}_2(q^z,q^w,\mu)\right|\ge mnr_{mn}\epsilon\,\middle|\, z^*,w^*\right)\\
&\le C_1 2^m\exp\left(-C_2 mnr_{mn}\epsilon^2\right)+C_3 2^n\exp\left(-C_4 mnr_{mn}\epsilon^2\right).
\end{aligned}$$

Proof By a similar argument as in Lemma 12,

$$\begin{aligned}
&\mathbb{P}\left(\max_{q_{ik}^z\in[0,1],i=1,\ldots,m}\left|\sum_{i=1}^m\sum_{j=1}^n(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z\right|\geq mnr_{mn}\epsilon\middle|z^*,w^*\right)\\
&=\mathbb{P}\left(\max_{q_{ik}^z\in\{0,1\},i=1,\ldots,m}\left|\sum_{i=1}^m\sum_{j=1}^n(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z\right|\geq mnr_{mn}\epsilon\middle|z^*,w^*\right)\\
&\leq 2^{m+1}\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C,\epsilon)}\right). \qquad (12)
\end{aligned}$$

Let $E_i^r=E[d_i^r|z^*,w^*]=\sum_{j=1}^n E[A_{ij}|z^*,w^*]$ and $E_j^c=E[d_j^c|z^*,w^*]=\sum_{i=1}^m E[A_{ij}|z^*,w^*]$. Note that

$$\begin{aligned}
&\max_{q^z\in\mathcal{C}_z,q^w\in\mathcal{C}_w}\left|\sum_{i=1}^m\sum_{j=1}^n\left(\frac{d_i^r}{n}\frac{E_j^c}{m}-\frac{E_i^r}{n}\frac{E_j^c}{m}\right)q_{ik}^zq_{jl}^w\right|\\
&\leq\max_{q^w\in\mathcal{C}_w}\left|\sum_{j=1}^n\frac{E_j^c}{m}q_{jl}^w\right|\max_{q^z\in\mathcal{C}_z}\left|\sum_{i=1}^m\left(\frac{d_i^r}{n}-\frac{E_i^r}{n}\right)q_{ik}^z\right|\\
&\leq nE_{\max}r_{mn}\max_{q^z\in\mathcal{C}_z}\left|\sum_{i=1}^m\left(\frac{d_i^r}{n}-\frac{E_i^r}{n}\right)q_{ik}^z\right|\\
&\leq E_{\max}r_{mn}\max_{q^z\in\mathcal{C}_z}\left|\sum_{i=1}^m\sum_{j=1}^n(A_{ij}-E[A_{ij}|z^*,w^*])q_{ik}^z\right|. \qquad (13)
\end{aligned}$$

From (12) and (13)

$$\begin{aligned}
&\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z,q^w\in\mathcal{C}_w}\left|\sum_{i=1}^m\sum_{j=1}^n\left(\frac{d_i^r}{n}\frac{E_j^c}{m}-\frac{E_i^r}{n}\frac{E_j^c}{m}\right)q_{ik}^zq_{jl}^w\right|\geq mnr_{mn}^2\epsilon\middle|z^*,w^*\right)\\
&\qquad\leq C_5 2^m\exp\left(-C_6mnr_{mn}\epsilon^2\right). \qquad (14)
\end{aligned}$$

Next, we analyze the term $\max_{q^z\in\mathcal{C}_z,q^w\in\mathcal{C}_w}\left|\sum_{i=1}^m\sum_{j=1}^n\left(\frac{d_i^r}{n}\frac{d_j^c}{m}-\frac{d_i^r}{n}\frac{E_j^c}{m}\right)q_{ik}^zq_{jl}^w\right|$.

If $\max_{q^z \in \mathcal{C}_z} \left| \sum_{i=1}^m \left(\frac{d_i^r}{n} - \frac{E_i^r}{n} \right) q_{ik}^z \right| \le m r_{mn} \epsilon$,

$$
\begin{aligned}
& \max_{q^z \in \mathcal{C}_z, q^w \in \mathcal{C}_w} \left| \sum_{i=1}^m \sum_{j=1}^n \left(\frac{d_i^r}{n} \frac{d_j^c}{m} - \frac{d_i^r}{n} \frac{E_j^c}{m} \right) q_{ik}^z q_{jl}^w \right| \\
\le & \max_{q^z \in \mathcal{C}_z} \left| \sum_{i=1}^m \frac{d_i^r}{n} q_{ik}^z \right| \max_{q^w \in \mathcal{C}_w} \left| \sum_{j=1}^n \left(\frac{d_j^c}{m} - \frac{E_j^c}{m} \right) q_{jl}^w \right| \\
= & \max_{q^z \in \mathcal{C}_z} \left| \sum_{i=1}^m \left(\frac{d_i^r}{n} - \frac{E_i^r}{n} \right) q_{ik}^z + \sum_{i=1}^m \frac{E_i^r}{n} q_{ik}^z \right| \max_{q^w \in \mathcal{C}_w} \left| \sum_{j=1}^n \left(\frac{d_j^c}{m} - \frac{E_j^c}{m} \right) q_{jl}^w \right| \\
\le & \left(m r_{mn} \epsilon + m E_{\max} r_{mn} \right) \max_{q^w \in \mathcal{C}_w} \left| \sum_{j=1}^n \left(\frac{d_j^c}{m} - \frac{E_j^c}{m} \right) q_{jl}^w \right| \\
= & (\epsilon + E_{\max}) r_{mn} \max_{q^w \in \mathcal{C}_w} \left| \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - E[A_{ij} | z^*, w^*]) q_{jl}^w \right| . \qquad (15)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \mathbb{P} \left(\max_{q^z \in \mathcal{C}_z, q^w \in \mathcal{C}_w} \left| \sum_{i=1}^m \sum_{j=1}^n \left(\frac{d_i^r}{n} \frac{d_j^c}{m} - \frac{d_i^r}{n} \frac{E_j^c}{m} \right) q_{ik}^z q_{jl}^w \right| \ge m n r_{mn}^2 \epsilon \,\middle|\, z^*, w^* \right) \\
\le & \mathbb{P} \left(\max_{q^z \in \mathcal{C}_z} \left| \sum_{i=1}^m \left(\frac{d_i^r}{n} - \frac{E_i^r}{n} \right) q_{ik}^z \right| \ge m r_{mn} \epsilon \,\middle|\, z^*, w^* \right) \\
& + \mathbb{P} \left((\epsilon + E_{\max}) r_{mn} \max_{q^w \in \mathcal{C}_w} \left| \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - E[A_{ij} | z^*, w^*]) q_{jl}^w \right| \ge m n r_{mn}^2 \epsilon \,\middle|\, z^*, w^* \right) \\
\le & C_7 2^m \exp\left(-C_8 m n r_{mn} \epsilon^2 \right) + C_9 2^n \exp\left(-C_{10} m n r_{mn} \epsilon^2 \right) . \qquad (16)
\end{aligned}
$$

To sum up, from (14) and (16), for sufficiently small positive ϵ,

$$
\begin{aligned}
& \mathbb{P} \left(\max_{q^z \in \mathcal{C}_z, q^w \in \mathcal{C}_w} \left| \sum_{i=1}^m \sum_{j=1}^n \left(\frac{d_i^r}{n} \frac{d_j^c}{m} - \frac{E_i^r}{n} \frac{E_j^c}{m} \right) q_{ik}^z q_{jl}^w \right| \ge m n r_{mn}^2 \epsilon \,\middle|\, z^*, w^* \right) \\
\le & C_5 2^m \exp\left(-C_6 m n r_{mn} \epsilon^2 \right) + C_7 2^m \exp\left(-C_8 m n r_{mn} \epsilon^2 \right) + C_9 2^n \exp\left(-C_{10} m n r_{mn} \epsilon^2 \right) .
\end{aligned}
$$

Next,

$$
\begin{aligned}
& \mathbb{P} \left(|D - E[D | z^*, w^*]| \ge r_{mn} \epsilon \,|\, z^*, w^* \right) \\
= & \mathbb{P} \left(\left| \frac{1}{mn} \sum_{ij} (A_{ij} - E[A_{ij} | z^*, w^*]) \right| \ge r_{mn} \epsilon \,\middle|\, z^*, w^* \right) \\
\le & C_{11} \exp\left(-C_{12} m n r_{mn} \epsilon^2 \right) .
\end{aligned}
$$

Recall that

$$\hat{\theta}_i\hat{\lambda}_j = \frac{d_i^r d_j^c}{mnD} \quad \text{and} \quad \theta_i^*\lambda_j^* = \frac{E_i^r E_j^c}{mnE[D|z^*, w^*]}.$$

Then by a similar argument for convergence of ratio of two random variables,

$$\begin{aligned}
&\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w}\left|\sum_{i=1}^m\sum_{j=1}^n(\hat{\theta}_i\hat{\lambda}_j - \theta_i^*\lambda_j^*)q_{ik}^z q_{jl}^w\right| \geq mnr_{mn}\epsilon \,\middle|\, z^*, w^*\right) \\
&\leq C_{13}2^m \exp\left(-C_{14}mnr_{mn}\epsilon^2\right) + C_{15}2^n \exp\left(-C_{16}mnr_{mn}\epsilon^2\right)
\end{aligned}$$

Finally, the lemma holds since

$$\begin{aligned}
&|J_2(q^z, q^w, \mu) - E[J_2(q^z, q^w, \mu)|z^*, w^*]| \\
=& \left|\sum_{i=1}^m\sum_{j=1}^n(\hat{\theta}_i\hat{\lambda}_j - \theta_i^*\lambda_j^*)\left(\sum_{k=1}^K\sum_{l=1}^L q_{ik}^z q_{jl}^w\mu_{kl}\right)\right| \\
\leq& \sum_{k=1}^K\sum_{l=1}^L\mu_{kl}\left|\sum_{i=1}^m\sum_{j=1}^n(\hat{\theta}_i\hat{\lambda}_j - \theta_i^*\lambda_j^*)q_{ik}^z q_{jl}^w\right| \\
\leq& \mu_{\max}\sum_{k=1}^K\sum_{l=1}^L\left|\sum_{i=1}^m\sum_{j=1}^n(\hat{\theta}_i\hat{\lambda}_j - \theta_i^*\lambda_j^*)q_{ik}^z q_{jl}^w\right|.
\end{aligned}$$

■

Note that

$$\begin{aligned}
&\left|\hat{J}(q^z, q^w, \Phi) - \bar{J}(q^z, q^w, \mu)\right| \\
\leq& \left|J_1(q^z, q^w, \mu) - \bar{J}_1(q^z, q^w, \mu)\right| + \left|J_2(q^z, q^w, \mu) - \bar{J}_2(q^z, q^w, \mu)\right| + \left|J_3(q^z, q^w, \pi, \rho)\right|,
\end{aligned}$$

and

$$\max_{q^z, q^w, \pi\in\mathcal{C}_\pi, \rho\in\mathcal{C}_\rho}|J_3(q^z, q^w, \pi, \rho)| = O(m+n).$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \pi\in\mathcal{C}_\pi, \rho\in\mathcal{C}_\rho, \mu\in\mathcal{C}_\mu}\left|\hat{J}(q^z, q^w, \Phi) - \bar{J}(q^z, q^w, \mu)\right| \geq mnr_{mn}\epsilon \,\middle|\, z^*, w^*\right) \\
&\leq \mathbb{P}\left(\max_{q^z\in\mathcal{C}_z, q^w\in\mathcal{C}_w, \pi\in\mathcal{C}_\pi, \rho\in\mathcal{C}_\rho, \mu\in\mathcal{C}_\mu}\left|(\hat{J}_1 + \hat{J}_2)(q^z, q^w, \mu) - (\bar{J}_1 + \bar{J}_2)(q^z, q^w, \mu)\right| + C(m+n)\right. \\
&\qquad \left.\geq mnr_{mn}\epsilon \,\middle|\, z^*, w^*\right). \qquad (17)
\end{aligned}$$

For (17) converges to 0, it is sufficient to assume $(mnr_{mn}\epsilon)/(m+n) \to \infty$ and $(mnr_{mn}\epsilon^2)/(m+n) \to \infty$ by Lemmas 13 and 14. But the first condition is implied by the second one. Therefore, Theorem 6 is proven.

A.6 Proof of Theorem 8

Let

$$G(q^z, q^w, \mu) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L} \mathbb{R}_{kk'}(1^{z^*}, q^z)\mathbb{R}_{ll'}(1^{w^*}, q^w)\mathrm{KL}(\mu^*_{kl}, \mu_{k'l'}),$$

where $\mathrm{KL}(\mu^*_{kl}, \mu_{k'l'}) = \mu^*_{kl}\log(\mu^*_{kl}/\mu_{k'l'}) - (\mu^*_{kl} - \mu_{k'l'})$.

The next proposition shows that $\bar{J}(q^z, q^w, \mu)$ is maximized at the true label assignments and true parameters, and gives a lower bound of the difference between $\bar{J}(q^z, q^w, \mu)$ and the maximum in the form of the confusion matrices.

Proposition 15 *For any q^z, q^w, and μ,*

$$\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(q^z, q^w, \mu) \geq mnr_{mn}\theta_{\min}\lambda_{\min}G(q^z, q^w, \mu) \geq 0.$$

Proof A straightforward calculation shows that

$$\begin{aligned}
\bar{J}_1(q^z, q^w, \mu) &= \sum_{i=1}^{m}\sum_{j=1}^{n} E[A_{ij}|z^*, w^*]\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\log\mu_{k'l'}\right)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \theta^*_i\lambda^*_j\mu^*_{z^*_i w^*_j}\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\log\mu_{k'l'}\right)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \theta^*_i\lambda^*_j\left(\sum_{k=1}^{K}\sum_{l=1}^{L} 1^{z^*}_{ik}1^{w^*}_{jl}\mu^*_{kl}\right)\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\log\mu_{k'l'}\right)\\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}\theta^*_i 1^{z^*}_{ik}q^z_{ik'}\right)\left(\sum_{j=1}^{n}\lambda^*_j 1^{w^*}_{jl}q^w_{jl'}\right)\mu^*_{kl}\log\mu_{k'l'}.
\end{aligned}$$

$$\begin{aligned}
\bar{J}_2(q^z, q^w, \mu) &= -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta^*_i\lambda^*_j\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\mu_{k'l'}\right)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta^*_i\lambda^*_j\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\mu_{k'l'}\right)\left(\sum_{k=1}^{K}\sum_{l=1}^{L} 1^{z^*}_{ik}1^{w^*}_{jl}\right)\\
&= -\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}\theta^*_i 1^{z^*}_{ik}q^z_{ik'}\right)\left(\sum_{j=1}^{n}\lambda^*_j 1^{w^*}_{jl}q^w_{jl'}\right)\mu_{k'l'}.
\end{aligned}$$

$$\begin{aligned}
\bar{J}_1(1^{z^*}, 1^{w^*}, \mu^*) &= \sum_{i=1}^{m}\sum_{j=1}^{n}\theta^*_i\lambda^*_j\left(\sum_{k=1}^{K}\sum_{l=1}^{L} 1^{z^*}_{ik}1^{w^*}_{jl}\mu^*_{kl}\log\mu^*_{kl}\right)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\theta^*_i\lambda^*_j\left(\sum_{k=1}^{K}\sum_{l=1}^{L} 1^{z^*}_{ik}1^{w^*}_{jl}\mu^*_{kl}\log\mu^*_{kl}\right)\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} q^z_{ik'}q^w_{jl'}\right)\\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}\theta^*_i 1^{z^*}_{ik}q^z_{ik'}\right)\left(\sum_{j=1}^{n}\lambda^*_j 1^{w^*}_{jl}q^w_{jl'}\right)\mu^*_{kl}\log\mu^*_{kl}.
\end{aligned}$$

$$\begin{aligned}\bar{J}_2(1^{z^*}, 1^{w^*}, \mu^*) &= -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i^*\lambda_j^*\left(\sum_{k=1}^{K}\sum_{l=1}^{L}1_{ik}^{z^*}1_{jl}^{w^*}\mu_{kl}^*\right)\\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n}\theta_i^*\lambda_j^*\left(\sum_{k=1}^{K}\sum_{l=1}^{L}1_{ik}^{z^*}1_{jl}^{w^*}\mu_{kl}^*\right)\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L}q_{ik'}^{z}q_{jl'}^{w}\right)\\ &= -\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}\theta_i^*1_{ik}^{z^*}q_{ik'}^{z}\right)\left(\sum_{j=1}^{n}\lambda_j^*1_{jl}^{w^*}q_{jl'}^{w}\right)\mu_{kl}^*.\end{aligned}$$

Therefore,

$$\begin{aligned}&\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(q^z, q^w, \mu)\\ &= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}\theta_i^*1_{ik}^{z^*}q_{ik'}^{z}\right)\left(\sum_{j=1}^{n}\lambda_j^*1_{jl}^{w^*}q_{jl'}^{w}\right)\mathrm{KL}(\mu_{kl}^*, \mu_{k'l'})\\ &\geq r_{mn}\theta_{\min}\lambda_{\min}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\left(\sum_{i=1}^{m}1_{ik}^{z^*}q_{ik'}^{z}\right)\left(\sum_{j=1}^{n}1_{jl}^{w^*}q_{jl'}^{w}\right)\mathrm{KL}(\mu_{kl}^*, \mu_{k'l'})\\ &= mnr_{mn}\theta_{\min}\lambda_{\min}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\mathbb{R}_{kk'}(1^{z^*}, q^z)\mathbb{R}_{ll'}^{w}(1^{w^*}, q^w)\mathrm{KL}(\mu_{kl}^*, \mu_{k'l'}) \geq 0.\end{aligned}$$

■

We now present a lemma on KL divergence:

Lemma 16 $\mathrm{KL}(a, b) = a\log\frac{a}{b} - (a-b) \geq \min\{(a-b)^2/(6b), |a-b|\}$, *where* a, $b > 0$.

Proof Define $x = \log\frac{a}{b}$. We show below that $a\log\frac{a}{b} - (a-b) \geq (a-b)^2/(6b)$ when $x \leq 2$, and $a\log\frac{a}{b} - (a-b) \geq |a-b|$ when $x > 2$.

Note that

$$\begin{aligned}& a\log\frac{a}{b} - (a-b) \geq (a-b)^2/(6b)\\ \iff\quad & \frac{a}{b}\log\frac{a}{b} - (\frac{a}{b}-1) \geq (a-b)^2/(6b^2)\\ \iff\quad & e^x x - (e^x - 1) \geq \frac{1}{6}(e^x-1)^2\\ \iff\quad & 6xe^x - e^{2x} - 4e^x + 5 \geq 0,\end{aligned}$$

where the last line holds when $x \leq 2$. Moreover, under the condition $x > 2$ that implies $a > b$, we have

$$\begin{aligned}& a\log\frac{a}{b} - (a-b) \geq |a-b|\\ \iff\quad & \frac{a}{b}\log\frac{a}{b} - (\frac{a}{b}-1) \geq \frac{a}{b} - 1\\ \iff\quad & xe^x \geq 2(e^x - 1),\end{aligned}$$

where the last line trivially holds as $x > 2$ and $e^x > e^x - 1$.

■

We now prove Theorem 8. By Lemma 16, we have

$$\mathrm{KL}(\mu^*_{kl}, \mu_{k'l'}) \geq \min\{(\mu^*_{kl} - \mu_{k'l'})^2/(6\mu_{\max}), |\mu^*_{kl} - \mu_{k'l'}|\} \geq \min\{(\mu^*_{kl} - \mu_{k'l'})^2/(6\mu_{\max}), 6\mu_{\max}\}.$$

We define a new metric in $\mathbb{R}^1$ by $|a-b|_{\text{new}} = \min\{|a-b|, 6\mu_{\max}\}$. Then we have $\mathrm{KL}(a,b) \geq |a-b|^2_{\text{new}}/(6\mu_{\max})$. For vectors a and b, define $\|a-b\|_{\text{new}} = \sqrt{\sum_i |a_i - b_i|^2_{\text{new}}}$.

We derive the lower bound for $G(q^z, q^w, \mu)$. Because

$$\sum_{l'=1}^{L} \frac{1}{n} \sum_{j=1}^{n} 1^{w^*}_{jl} \, q^w_{jl'} = \frac{1}{n} \sum_{j=1}^{n} 1^{w^*}_{jl},$$

for all l, there exists l', denoted by $h(l)$, such that

$$\frac{1}{n} \sum_{j=1}^{n} 1^{w^*}_{jl} \, q^w_{jl'} \geq \frac{1}{L} \frac{1}{n} \sum_{j=1}^{n} 1^{w^*}_{jl} \geq \frac{1}{L} \tilde{\rho}_{\min}.$$

Denote $\tilde{\mu}_{k',l} = \mu_{k',h(l)}$. Then

$$\begin{aligned} G(q^z, q^w, \mu) &\geq \frac{1}{L} \hat{\rho}_{\min} \sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\frac{1}{m} \sum_{i=1}^{m} 1^{z^*}_{ik} q^z_{ik'} \right) \sum_{l=1}^{L} \mathrm{KL}(\mu^*_{kl}, \tilde{\mu}_{k',l}) \\ &\geq \frac{1}{L} \hat{\rho}_{\min} \sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\frac{1}{m} \sum_{i=1}^{m} 1^{z^*}_{ik} q^z_{ik'} \right) 6\mu_{\max} \|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}}. \end{aligned}$$

Let $d_{\min} = \min_{k \neq k'} \|\mu^*_{k\cdot} - \mu^*_{k'\cdot}\|_{\text{new}}$. Note that based on H_3, $d_{\min} > 0$. Furthermore, for any $\tilde{\mu}_{k'\cdot}$ there exists at most one $\mu^*_{k\cdot}$ such that $\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|_{\text{new}} < d_{\min}/2$.

There are two possible cases:

Case 1: For each $\mu^*_{k\cdot}$, there exists one and only one $\tilde{\mu}_{k'\cdot}$ such that $\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|_{\text{new}} < d_{\min}/2$.

Case 2: There exists some $\mu^*_{k\cdot}$ such that no $\tilde{\mu}_{k'\cdot}$ is within its $d_{\min}/2$-radius.

The one-to-one correspondence in Case 1 induces a permutation s on $\{1, ..., K\}$. Case 1 implies $\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \geq d^2_{\min}/4$ for $k \neq s(k')$.

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\frac{1}{m} \sum_{i=1}^{m} 1^{z^*}_{ik} q^z_{ik'} \right) \|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \\ \geq &\sum_{k'=1}^{K} \sum_{k \neq s(k')} \left(\frac{1}{m} \sum_{i=1}^{m} 1^{z^*}_{ik} q^z_{ik'} \right) \|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \\ \geq &\frac{d^2_{\min}}{4} \sum_{k'=1}^{K} \sum_{k \neq s(k')} \left(\frac{1}{m} \sum_{i=1}^{m} 1^{z^*}_{ik} q^z_{ik'} \right) \geq \frac{d^2_{\min}}{4} M_{\text{row}}(q^z). \end{aligned}$$

In Case 2, let k be the class label such that $\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \geq d^2_{\min}/4$ for all k'.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{k'=1}^{K}\left(\frac{1}{m}\sum_{i=1}^{m}1^{z^*}_{ik}q^z_{ik'}\right)\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \\
\geq\ &\sum_{k'=1}^{K}\left(\frac{1}{m}\sum_{i=1}^{m}1^{z^*}_{ik}q^z_{ik'}\right)\|\mu^*_{k\cdot} - \tilde{\mu}_{k'\cdot}\|^2_{\text{new}} \\
\geq\ &\frac{d^2_{\min}}{4}\frac{1}{m}\sum_{i=1}^{m}1^{z^*}_{ik} \geq \frac{d^2_{\min}}{4}\tilde{\rho}_{\min}M_{\text{row}}(q^z).
\end{aligned}$$

In summary, $G(q^z, q^w, \mu) \geq C_1 M_{\text{row}}(q^z)$. Based on the same argument, $G(q^z, q^w, \mu) \geq C_2 M_{\text{col}}(q^w)$.

A.7 Proof of Theorem 9

Let $\Phi^* = (\pi^*, \rho^*, \mu^*)$. By Theorems 6 and 8, if $(mnr_{mn}\epsilon^2)/(m+n) \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(M_{\text{row}}(\hat{q}^z) \geq \epsilon \middle| z^*, w^*\right) \\
\leq\ &\mathbb{P}\left(\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \bar{J}(\hat{q}^z, \hat{q}^w, \hat{\mu}) \geq C_1 mnr_{mn}\epsilon \middle| z^*, w^*\right) \\
=\ &\mathbb{P}\Big(\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \hat{J}(1^{z^*}, 1^{w^*}, \Phi^*) + \hat{J}(1^{z^*}, 1^{w^*}, \Phi^*) - \hat{J}(\hat{q}^z, \hat{q}^w, \hat{\Phi}) \\
&+\hat{J}(\hat{q}^z, \hat{q}^w, \hat{\Phi}) - \bar{J}(\hat{q}^z, \hat{q}^w, \hat{\mu}) \geq C_1 mnr_{mn}\epsilon \Big| z^*, w^*\Big) \\
\leq\ &\mathbb{P}\left(\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \hat{J}(1^{z^*}, 1^{w^*}, \Phi^*) + \hat{J}(\hat{q}^z, \hat{q}^w, \hat{\Phi}) - \bar{J}(\hat{q}^z, \hat{q}^w, \hat{\mu}) \geq C_1 mnr_{mn}\epsilon \middle| z^*, w^*\right) \\
\leq\ &\mathbb{P}\left(|\bar{J}(1^{z^*}, 1^{w^*}, \mu^*) - \hat{J}(1^{z^*}, 1^{w^*}, \Phi^*)| \geq (C_1/2) mnr_{mn}\epsilon \middle| z^*, w^*\right) \\
&+\mathbb{P}\left(|\hat{J}(\hat{q}^z, \hat{q}^w, \hat{\Phi}) - \bar{J}(\hat{q}^z, \hat{q}^w, \hat{\mu})| \geq (C_1/2) mnr_{mn}\epsilon \middle| z^*, w^*\right) \to 0.
\end{aligned}$$

It implies, for $\epsilon = \left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\epsilon_1$, where δ and ϵ_1 are positive constants,

$$\mathbb{P}\left(M_{\text{row}}(\hat{q}^z) \geq \epsilon \middle| z^*, w^*\right) \to 0;$$

in other words,

$$M_{\text{row}}(\hat{q}^z) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right).$$

Similarly,

$$M_{\text{col}}(\hat{q}^w) = o_p\left(\left(\frac{mnr_{mn}}{m+n}\right)^{-1/2+\delta}\right).$$

A.8 Proof of Theorem 10

We only need to prove the uniform convergence of $\hat{J}(q^z, q^w, \Phi)$ to $\bar{J}(q^z, q^w, \mu)$ under the Bernoulli model, which relies on the following lemma:

Lemma 17 *Let $\{X_{ij}\}$ be independent Bernoulli variables with mean $E[X_{ij}] \le r_{mn}C$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left(\max_{0\le u_i\le 1, i=1,\dots,m, 0\le v_j\le 1, j=1,\dots,n} \left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right| \ge mnr_{mn}\epsilon\right) \le 2^{m+n+1}\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C, \epsilon/3)}\right).$$

Proof According to the Bernstein inequality, for $u_i \in \{0,1\}, i = 1, ..., m, v_j \in \{0,1\}, j = 1, ..., n$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])u_i v_j\right| \ge mnr_{mn}\epsilon\right) \le 2\exp\left(-\frac{m^2n^2r_{mn}^2\epsilon^2}{2(\sum_{ij}\mathrm{Var}[X_{ij}] + mnr_{mn}\epsilon/3)}\right) \le 2\exp\left(-\frac{mnr_{mn}\epsilon^2}{4\max(C, \epsilon/3)}\right).$$

The rest of the proof is identical to that of Lemma 12.

■

Appendix B. Application to SMS spam data set

We apply the proposed method to an SMS spam data set collected by Almeida et al. (2011). The data set has a total of 4827 SMS legitimate messages (labeled as "ham") and a total of a total of 747 spam messages (labeled as "spam"). We followed a standard procotol for data preprocessing (https://kharshit.github.io/blog/2017/08/25/email-spam-filtering-text-analysis-in-r) using the R package `tm`: we removed punctuation and stopwords (such as "that") from the messages, and only kept words appearing in at least 1% of the messages. This results in a total of 4938 messages, with 4211 labeled as "ham" and 727 labeled as "spam", and a vocabulary size of 139. The messages were then encoded in a 4938×139 matrix $A = [A_{ij}]$ where each entry A_{ij} represents the count of the j-th word in the i-th message. The data matrix, therefore, represents a bipartite network with integer weights.

The data set typically serves as a benchmark for supervised learning in spam filtering. We took an unsupervised learning approach; that is, we applied biclustering methods to this data set and compared the estimated labels on the messages with the true labels. We applied the three aforementioned methods—SC, PL, and DC-LBM—to the data set with $K = 2$ (since the messages are classified as "ham" or "spam") and $L = 2, 3, 4, 5, 6$. We reported two metrics—ARI and accuracy—for evaluating the performances. The latter is the fraction of estimated labels that match the true labels, allowing for a permutation of

the two classes. The results for SC, PL, and DC-LBM are reported in Table 1. Note that the performance of SC on row clustering does not depend on L. The best performance of PL is achieved at $L = 3$, and the best performance of DC-LBM is achieve at $L = 5$. For all values of L, DC-LBM consistently outperforms PL and SC in terms of both ARI and accuracy. This pattern aligns with the findings in the simulation studies.

L	2	3	4	5	6
ARI (SC)	0.419				
ARI (PL)	0.085	0.165	0.112	0.125	0.101
ARI (DC-LBM)	0.634	0.617	0.719	0.729	0.644
Accuracy (SC)	0.851				
Accuracy (PL)	0.654	0.716	0.672	0.683	0.661
Accuracy (DC-LBM)	0.923	0.916	0.944	0.946	0.924

Table 1: ARI and accuracy of SC, PL. and DC-LBM on the SMS spam data set.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pages 259–262, 2011.

Faris Alqadah, Chandan K Reddy, Junling Hu, and Hatim F Alqadah. Biclustering neighborhood-based collaborative filtering method for top-n recommender systems. *Knowledge and Information Systems*, 44(2):475–491, 2015.

Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435. URL `https://doi.org/10.1214/16-AOS1435`.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106: 21068–21073, 2009.

Peter J Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

Vincent Brault, Christine Keribin, and Mahendra Mariadassou. Consistency and asymptotic normality of latent block model estimators. *Electronic Journal of Statistics*, 14(1):1234–1268, 2020.

Clément Canonne. A short note on poisson tail bounds. *manuscript*, 2019. http://www.cs.columbia.edu/ ccanonne/files/misc/2017-poissonconcentration.pdf.

Gilles Celeux and Gérard Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176, 1991.

Yizong Cheng and George M Church. Biclustering of expression data. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 93–103, 2000.

Hyuk Cho, Inderjit S Dhillon, Yuqiang Guan, and Suvrit Sra. Minimum sum-squared residue co-clustering of gene expression data. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 114–125. SIAM, 2004.

J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, 2008.

Pablo AD de Castro, Fabrício O de França, Hamilton M Ferreira, and Fernando J Von Zuben. Applying biclustering to text mining: an immune-inspired approach. In *International Conference on Artificial Immune Systems*, pages 83–94. Springer, 2007.

Cheryl Flynn and Patrick Perry. Profile likelihood biclustering. *Electronic Journal of Statistics*, 14(1):731–768, 2020.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Asscociation*, 97(458):611–631, 2002.

Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003.

Gérard Govaert and Mohamed Nadif. Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5):415–422, 2006.

Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6): 3233–3245, 2008.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

John A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Asscociation*, 67(337):123–129, 1972.

Paul W Holland, Kathryn B Laskey, and Samuel Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

Jianwei Hu, Hong Qin, Ting Yan, and Yunpeng Zhao. Corrected bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, pages 1–13, 2019.

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

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

Brian Karrer and Mark E J Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011.

Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307, 1970.

Can M Le and Elizaveta Levina. Estimating the number of communities by spectral methods. *Electronic Journal of Statistics*, 16(1):3315–3342, 2022.

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

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Shujie Ma, Liangjun Su, and Yichong Zhang. Determining the number of communities in degree-corrected stochastic block models. *Journal of Machine Learning Research*, 22(69): 1–63, 2021.

Mahendra Mariadassou, Catherine Matias, et al. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

Patryk Orzechowski and Krzysztof Boryczko. Text mining with hybrid biclustering algorithms. In *International Conference on Artificial Intelligence and Soft Computing*, pages 102–113. Springer, 2016.

Victor A Padilha and Ricardo JGB Campello. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18(1):1–25, 2017.

Beatriz Pontes, Raúl Giráldez, and Jesús S Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, 2015.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic block model. *The Annals of Statistics*, 39(4):1878—-1915, 2011.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45): 12679–12684, 2016.

D Franco Saldana, Yi Yu, and Yang Feng. How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017.

Purnamrita Sarkar and Peter J. Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962 – 990, 2015. doi: 10.1214/14-AOS1285. URL https://doi.org/10.1214/14-AOS1285.

Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Jiangzhou Wang, Binghui Liu, and Jianhua Guo. Efficient split likelihood-based method for community detection of large-scale networks. *Stat*, 10(1):e349, 2021a.

Jiangzhou Wang, Jingfei Zhang, Binghui Liu, Ji Zhu, and Jianhua Guo. Fast network community detection with profile-pseudo likelihood methods. *Journal of the American Statistical Association*, 118(542):1359–1372, 2023.

YX Rachel Wang and Peter J Bickel. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528, 2017.

Zhe Wang, Yingbin Liang, and Pengsheng Ji. Spectral algorithms for community detection in directed networks. *Journal of Machine Learning Research*, 21:1–45, 2021b.

Chihiro Watanabe and Taiji Suzuki. Goodness-of-fit test for latent block models. *Computational Statistics & Data Analysis*, 154:107090, 2021.

Jingnan Zhang, Xin He, and Junhui Wang. Directed community detection with network embedding. *Journal of the American Statistical Association*, 117(540):1809–1819, 2022.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283, 2020.

Yunpeng Zhao. A survey on theoretical advances of community detection in networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5), 2017.

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

Yunpeng Zhao, Qing Pan, and Chengan Du. Logistic regression augmented community detection for network data with application in identifying autism-related gene pathways. *Biometrics*, 75(1):222–234, 2019.