Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2015, Article ID 657965, 14 pages
http://dx.doi.org/10.1155/2015/657965

Research Article

Estimating Parameters of a Probabilistic Heterogeneous Block Model via the EM Algorithm

Marianna Bolla and Ahmed Elbanna

Institute of Mathematics, Budapest University of Technology and Economics, Műegyetem Rakpart 3, Budapest 1111, Hungary

Correspondence should be addressed to Marianna Bolla; marib@math.bme.hu

Received 16 June 2015; Revised 6 November 2015; Accepted 10 November 2015

Academic Editor: Hyungjun Cho

We introduce a semiparametric block model for graphs, where the within- and between-cluster edge probabilities are not constants within the blocks but are described by logistic type models, reminiscent of the 50-year-old Rasch model and the newly introduced α-β models. Our purpose is to give a partition of the vertices of an observed graph so that the induced subgraphs and bipartite graphs obey these models, where their strongly interlaced parameters give multiscale evaluation of the vertices at the same time. In this way, a profoundly heterogeneous version of the stochastic block model is built via mixtures of the above submodels, while the parameters are estimated with a special EM iteration.

1. Introduction

So far many parametric and nonparametric methods have been proposed for so-called community detection in networks. In the nonparametric scenario, hierarchical or spectral methods were applied to maximize the two- or multiway Newman-Girvan modularity [1–4]; more generally, spectral clustering tools (SC), based on Laplacian or modularity spectra, proved to be feasible to find community, anticommunity, or regular structures in networks [5]. In the parametric setup, certain model parameters are estimated, usually via maximizing the *likelihood function* of the graph, that is, the joint probability of our observations under the model equations. This so-called ML estimation is a promising method of statistical inference, has solid theoretical foundations [6, 7], and also supports the common-sense goal of accepting parameter values based on which our sample is the most likely.

As for the parametric scenario, in the 2010s, α and β models [8, 9] were developed as the unique graph models where the *degree sequence is a sufficient statistic*: given the degree sequence, the distribution of the random graph does not depend on the parameters any more (microcanonical distribution over the model graphs). This fact makes it possible to derive the ML estimate of the parameters in a standard way [10]. Indeed, in the context of network data, a lot of information is contained in the degree sequence, though, perhaps, in a more sophisticated way. The vertices may have *clusters* (groups or modules) and their memberships may affect their affinity to make ties. We will find groups of the vertices such that the within- and between-cluster edge probabilities admit certain parametric graph models, the parameters of which are highly interlaced. Here the degree sequence is not a sufficient statistic any more, only if it is restricted to the subgraphs. When making inference, we are partly inspired by the stochastic block model, partly by the Rasch model, and by the rectangular analogue of the α-β models.

The generalized random graph model, sometimes called stochastic block model (SBM), was first introduced in [11] and discussed later in [12–18]. This model is the generalization of the classical Erdős-Renyi random graph $G_n(p)$, the first random graph of the history introduced in [19] which corresponds to the one-cluster case: between any pair of the n vertices edges come into existence independently, with the same probability p. The graph $G_n(\mathbf{P}, \mathscr{P}_k)$ on n vertices is a *generalized random graph* with $k \times k$ symmetric probability matrix $\mathbf{P} = (p_{uv})$ and proper k-partition $\mathscr{P}_k = (C_1, \ldots, C_k)$ of the vertices if vertices of C_u and C_v are connected independently, with probability p_{uv}, $1 \le u < v \le$

k; further, any pair of the vertices within C_u is connected with probability p_{uu} $(u = 1, \ldots, k)$. Therefore, the subgraph of $G_n(\mathbf{P}, \mathscr{P}_k)$ confined to the vertex set C_u is an Erdős-Renyi type random graph, while the bipartite subgraphs connecting vertices of C_u and C_v $(u \neq v)$ are random bipartite graphs of edge probability p_{uv}. Sometimes we refer to $\mathscr{P}_k$ as *clustering*, where $C_1, \ldots, C_k$ are the clusters. In fact, seminal Szemeredi's regularity lemma [20] guarantees the existence of such a structure in any huge graph, albeit with an enormously large number of clusters; therefore, it is not applicable for practical purposes.

Though in [18] the SBM is called heterogeneous, it is, in fact, a homogeneous one: the probability to make ties is the same within the clusters or between the cluster pairs. Nonetheless, this probability depends on the actual cluster memberships; given the memberships of the vertices, the probability that they are connected is a given constant, for the estimation of which an algorithm (maximizing the likelihood modularity) is proposed in [12] and the the EM algorithm is described in [5]. Here we want to model more complicated within- and between-cluster relations.

As for the nonparametric scenario, the role of the degree sequence is the best enhanced in the normalized Laplacian or normalized modularity based SC. In [4] we extended the notion of the *modularity matrix* to weighted graphs as follows. Let $G = (V, \mathbf{W})$ be an *edge-weighted graph* on the n-element vertex-set V with the $n \times n$ symmetric weight matrix $\mathbf{W}$; the entries satisfy $w_{ij} = w_{ji} \geq 0$, $w_{ii} = 0$, and they are similarities between the vertex pairs. The *modularity matrix* of G is defined as $\mathbf{M} = \mathbf{W} - \mathbf{d}\mathbf{d}^T$, where the entries of $\mathbf{d}$ are the *generalized vertex degrees* $d_i = \sum_{j=1}^{n} w_{ij}$ $(i = 1, \ldots, n)$. Here $\mathbf{W}$ is normalized in such a way that $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$, an assumption that does not hurt the generality, since the forthcoming *normalized modularity matrix*, to be mostly used, is not affected by the scaling of the entries of $\mathbf{W}$:

$$\mathbf{M}_\mathbf{D} = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2}, \tag{1}$$

where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ is the diagonal *degree matrix*.

In [4], we also introduced the following spectral relaxation technique to approximate the k-partition of the vertices minimizing the within- and between-cluster discrepancies in the spirit of Szemeredi's regularity lemma. (This discrepancy measures the homogeneity of the clusters; we will not use this notion here; see [21] for more information.) Let the eigenvalues of $\mathbf{M}_\mathbf{D}$, enumerated in decreasing absolute values, be $1 > |\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_n| = 0$. Assume that $|\mu_{k-1}| > |\mu_k|$, and denote by $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ the corresponding unit-norm, pairwise orthogonal eigenvectors. Let $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^{k-1}$ be the row vectors of the $n \times (k-1)$ matrix of column vectors $\mathbf{D}^{-1/2}\mathbf{u}_1, \ldots, \mathbf{D}^{-1/2}\mathbf{u}_{k-1}$; they are called $(k-1)$-dimensional representatives of the vertices. The *weighted k-variance* of these representatives is defined as

$$\widetilde{S}_k^2 = \min_{(C_1, \ldots, C_k)} \sum_{u=1}^{k} \sum_{j \in C_u} d_j \left\| \mathbf{r}_j - \mathbf{c}_u \right\|^2, \tag{2}$$

where $\mathbf{c}_u = (1/\mathrm{Vol}(C_u)) \sum_{j \in C_u} d_j \mathbf{r}_j$ is the weighted center of cluster C_u. It is the *weighted k-means algorithm* that gives this minimum, and the point is that the optimum $\widetilde{S}_k$ is just the minimum distance between the eigensubspace corresponding to $\mu_0, \ldots, \mu_{k-1}$ and the one of the suitably transformed step-vectors over the k-partitions of V. In Chapter 2 of [5] we also discussed that, in view of subspace perturbation theorems, the larger the gap between $|\mu_{k-1}|$ and $|\mu_k|$, the smaller $\widetilde{S}_k$.

Note that the normalized modularity based spectral clustering is the same as the normalized Laplacian based one (see [5]) with the exception that here not only the bottom, but the large absolute value eigenvalues are considered; further, the heterogeneity of the vertex degrees is encoded into the diagonal degree matrix $\mathbf{D}$. With the above technique we embed the vertices into a low dimensional space (spectral relaxation via eigenvectors) and hence perform metric clustering. However, above the spacial location, SC does not assign any parameters to the vertices. Our method to be introduced finds parameters and classifies the vertices at the same time.

Here we propose a profoundly heterogeneous block model by carrying on the Rasch model developed more than 50 years ago for evaluating psychological tests [22, 23]. We will call it Logistic Block Model (LBM). Given the number of clusters and a classification of the vertices, we will use the Rasch model for the bipartite subgraphs but the α-β models for the subgraphs themselves and process an iteration (inner cycle) to find the ML estimate of their parameters. Then, based on their contributions to the overall likelihood, we find a new classification of the vertices via taking conditional expectation and using the Bayes rule. Eventually, the two steps are alternated, giving the outer cycle of the iteration.

Our algorithm fits into the framework of the EM algorithm [7, 24], in the context of exponential families. The method was originally developed for missing data, and the name comes from the alternating *expectation* (E) and *maximization* (M) steps, where in the E-step (assignment phase) we complete the data by substituting for the missing data via taking conditional expectation, while in the M-step (estimation phase) we find the usual ML estimate of the parameters based on the so completed data. The EM algorithm naturally extends to situations, when not the data itself is missing, but it comes from a finite mixture, and the grouping memberships are the missing parameters. This special type of the EM algorithm developed for mixtures is often called collaborative filtering [25, 26] or Gibbs sampling [27], the roots of which method can be traced back to [28].

After proving the convergence of the inner cycle to the unique solution of the likelihood equation in each block separately, the convergence of the outer cycle to a local maximum of the likelihood function is easily seen. The advantage of the LBM is that, unlike SC, above clustering the vertices, it also assigns parameters to them, where parameters depend on their cluster memberships. Therefore we call is semiparametric. In the context of social networks, the clusters can be identified with social strata and the parameters with attitudes of people of one group towards people of the other, where attitude is the same for people in the second group but depends on the individual in the first group. The number of

clusters is fixed during the iteration, but an initial number of clusters are obtained by SC, via inspecting the normalized modularity spectrum of the graph. We apply the algorithm to randomly generated and real-world data, where the initial clustering was the one obtained by SC. Then we compare the results of SC, SBM, and LBM by the Rand index, and our LBM shows a better agreement with the SC clusters than the SBM. It seems that SC gives a solution close to a local maximum of LBM, which can be regarded as fine tuning of SC. In fact, without good starting clustering, LBM can run into a local maximum (there are many) far from the global one. Therefore, it needs SC starting; however, its advantage is that, in addition, it is able to estimate parameters too.

The paper is organized as follows. In Section 2 we describe the building blocks of our model. In the context of the α-β models we refer to already proved facts about the existence of the ML estimate, and if it exists, we discuss the algorithm proposed by [9] together with convergence facts; meanwhilewhile, in the context of the β-γ model, we introduce a novel algorithm and prove the convergence of it in the appendix. In Section 3 we use both of the above algorithms for the subgraphs and bipartite subgraphs of our sample graph, and we connect them together in the framework of the EM algorithm. In Section 4, the algorithm is applied to randomly generated and real-world data, while Section 5 is devoted to a brief discussion.

2. The Building Blocks of the LBM

Log-linear and logistic type models to describe contingency tables in a combinatorial fashion were proposed, for example, by [11, 29] and widely used in statistics. Together with the Rasch model, they give the foundation of the unweighted graph and bipartite graph models which are the building blocks of our EM iteration.

2.1. α-β Models for Undirected Random Graphs. With different parameterization, [8, 9] introduced the following random graph model, where the degree sequence is a sufficient statistic. We have an unweighted, undirected random graph on n vertices without loops, such that edges between distinct vertices come into existence independently, but not with the same probability as in the classical Erdős-Renyi model [19]. This random graph can uniquely be characterized by its $n \times n$ symmetric adjacency matrix $\mathbf{A} = (A_{ij})$ which has zero diagonal and the entries above the main diagonal are independent Bernoulli random variables whose parameters $p_{ij} = \mathbb{P}(A_{ij} = 1)$ obey the following rule. Actually, we formulate this rule for the $p_{ij}/(1 - p_{ij})$ ratios, the so-called *odds*:

$$\frac{p_{ij}}{1 - p_{ij}} = \alpha_i \alpha_j \quad (1 \le i < j \le n), \tag{3}$$

where the parameters $\alpha_1, \ldots, \alpha_n$ are positive reals. This model is called α *model* in [9]. With the parameter transformation $\beta_i = \ln \alpha_i$ $(i = 1, \ldots, n)$, it is equivalent to the β *model* of [8] which applies to the *logits*:

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \beta_j \quad (1 \le i < j \le n) \tag{4}$$

with real parameters $\beta_1, \ldots, \beta_n$.

Conversely, the probabilities p_{ij} and $1 - p_{ij}$ can be expressed in terms of the parameters, like

$$\begin{aligned} p_{ij} &= \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j}, \\ 1 - p_{ij} &= \frac{1}{1 + \alpha_i \alpha_j} \end{aligned} \tag{5}$$

whose formulas will be intensively used in the subsequent calculations.

We are looking for the ML estimate of the parameter vector $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ or $\underline{\beta} = (\beta_1, \ldots, \beta_n)$ based on the observed unweighted, undirected graph as a statistical sample. (It may seem that we have a one-element sample here; however, there are $\binom{n}{2}$ independent random variables, the adjacencies, in the background.)

Let $\mathbf{D} = (D_1, \ldots, D_n)$ denote the degree vector of the above random graph, where $D_i = \sum_{j=1}^{n} A_{ij}$ $(i = 1, \ldots, n)$. The random vector $\mathbf{D}$, as a function of the sample entries A_{ij}'s, is a *sufficient statistic* for the parameter $\underline{\alpha}$, or, equivalently, for $\underline{\beta}$. Roughly speaking, a sufficient statistic itself contains all the information—which can be retrieved from the data—for the parameter. More precisely, a statistic is sufficient when the conditional distribution of the sample, given the statistic, does not depend on the parameter any more. By the *Neyman-Fisher factorization theorem* [6], a statistic is sufficient if and only if the likelihood function of the sample can be factorized into two parts: one which does not contain the parameter and the other, which includes the parameter, contains the sample entries merely compressed into this sufficient statistic. Consider this factorization of the likelihood function (joint probability of A_{ij}'s) in our case. Because of the symmetry of $\mathbf{A}$, this is

$$\begin{aligned} L_{\underline{\alpha}}(\mathbf{A}) &= \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_{ij}^{A_{ij}} \left(1 - p_{ij}\right)^{1 - A_{ij}} \\ &= \left\{ \prod_{i=1}^{n} \prod_{j=1}^{n} p_{ij}^{A_{ij}} \left(1 - p_{ij}\right)^{1 - A_{ij}} \right\}^{1/2} \\ &= \left\{ \prod_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{p_{ij}}{1 - p_{ij}} \right)^{A_{ij}} \prod_{i=1}^{n} \prod_{j=1}^{n} \left(1 - p_{ij}\right) \right\}^{1/2} \\ &= \left\{ \prod_{i=1}^{n} \alpha_i^{\sum_{j=1}^{n} A_{ij}} \prod_{j=1}^{n} \alpha_j^{\sum_{i=1}^{n} A_{ij}} \prod_{i \ne j} \left(1 - p_{ij}\right) \right\}^{1/2} \end{aligned}$$

$$= \left\{\prod_{i\neq j}\frac{1}{1+\alpha_i\alpha_j}\right\}^{1/2}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\prod_{j=1}^{n}\alpha_j^{D_j}\right\}^{1/2}$$
$$= \left\{\prod_{i<j}\frac{1}{1+\alpha_i\alpha_j}\right\}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\right\} = C_{\underline{\alpha}}\times\prod_{i=1}^{n}\alpha_i^{D_i}, \quad (6)$$

where we used (5) and the facts that $A_{ij} = A_{ji}$, $p_{ij} = p_{ji}$ ($i < j$) and $A_{ii} = 0$, $p_{ii} = 0$ ($i = 1,\ldots,n$). Here the partition function $C_{\underline{\alpha}} = \prod_{i<j}(1/(1+\alpha_i\alpha_j))$ only depends on $\underline{\alpha}$, and the whole likelihood function depends on A_{ij}'s merely through D_i's. Therefore, $\mathbf{D}$ is a sufficient statistic. The other factor is constantly 1, indicating that the conditional joint distribution of the entries—given $\mathbf{D}$—is uniform, but we will not make use of this fact. The whole model comes from the so-called log-linear way of model building; see [29]. In [8, 10], the converse statement is also proved: the above α model (reparametrized as β model) is the unique one, where the degree sequence is a sufficient statistic.

Let (a_{ij}) be the matrix of the sample realizations (the adjacency entries of the observed graph), let $d_i = \sum_{j=1}^{n} a_{ij}$ be the actual degree of vertex i ($i = 1,\ldots,n$), and let $\mathbf{d} = (d_1,\ldots,d_n)$ be the observed degree vector. The above factorization also indicates that the joint distribution of the entries belongs to the exponential family, and hence, with natural parameterization [24], the maximum likelihood estimate $\widehat{\underline{\alpha}}$ (or equivalently, $\widehat{\underline{\beta}}$) is derived from the fact that, with it, the observed degree d_i equals the expected one; that is, $\mathbb{E}(D_i) = \sum_{i=1}^{n} p_{ij}$. Therefore, $\widehat{\underline{\alpha}}$ is the solution of the following *maximum likelihood equation*:

$$d_i = \sum_{j\neq i}^{n}\frac{\alpha_i\alpha_j}{1+\alpha_i\alpha_j} \quad (i = 1,\ldots,n). \quad (7)$$

The ML estimate $\widehat{\underline{\beta}}$ is easily obtained from $\widehat{\underline{\alpha}}$ via taking the logarithms of its coordinates.

Before discussing the solution of the system of (7), let us see what conditions a sequence of nonnegative integers should satisfy so that it could be realized as the degree sequence of a graph. The sequence $d_1,\ldots,d_n$ of nonnegative integers is called *graphic* if there is an unweighted, undirected graph on n vertices such that its vertex degrees are the numbers $d_1,\ldots,d_n$ in some order. Without loss of generality, d_i's can be enumerated in nonincreasing order. The Erdősi-Gallai theorem [30] gives the following necessary and sufficient condition for a sequence to be graphic. The sequence $d_1 \geq \cdots \geq d_n \geq 0$ of integers is graphic if and only if it satisfies the following two conditions: $\sum_{i=1}^{n} d_i$ is even and

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n}\min\{k, d_i\}, \quad k = 1,\ldots,n-1. \quad (8)$$

Note that for nonnegative (not necessarily integer) real sequences, a continuous analogue of (8) is derived in [8]. For given n, the convex hull of all possible graphic degree sequences is a polytope, to be denoted by $\mathscr{D}_n$. Its extreme points are the so-called *threshold graphs* [31]. It is interesting that for $n = 3$ all undirected graphs are threshold, since there are 8 possible graphs on 3 nodes, and there are also 8 vertices of $\mathscr{D}_3$; the $n = 2$ case is also not of much interest; therefore we will treat the $n > 3$ cases only. The number of vertices of $\mathscr{D}_n$ superexponentially grows with n [32]; therefore the problem of characterizing threshold graphs has a high computational complexity. Its facial and cofacial sets are fully described in [10]. Apart from the trivial cases (when there is at least one degree equal to 0 or $n-1$), in [33], the authors give the following equivalent characterization of a threshold graph for $n \geq 4$: it has no four different vertices, a, b, c, d, such that a, b and c, d are connected by an edge, but a, c and b, d are not; that is, it has no two disjoint copies of the complete graph K_2.

The authors of [8, 9] prove that $\mathscr{D}_n$ is the topological closure of the set of expected degree sequences, and, for given $n > 3$, if $\mathbf{d} \in \mathrm{int}(\mathscr{D}_n)$ is an interior point, then maximum likelihood equation (7) has a unique solution. Later, it turned out that the converse is also true: in [10] the authors prove that the ML estimate exists if and only if the observed degree vector is an inner point of $\mathscr{D}_n$. On the contrary, when the observed degree vector is a boundary point of $\mathscr{D}_n$, there is at least one 0 or 1 probability p_{ij} which can be obtained only by a parameter vector such that at least one of β_i's is not finite. In this case, the likelihood function cannot be maximized with a finite parameter set; its supremum is approached with a parameter vector $\underline{\beta}$ with at least one coordinate tending to $+\infty$ or $-\infty$.

The authors in [9] recommend the following algorithm and prove that, provided $\mathbf{d} \in \mathrm{int}(\mathscr{D}_n)$, its iteration converges to the unique solution of system (7). To motivate the iteration, we rewrite (7) as

$$d_i = \alpha_i\sum_{j\neq i}\frac{1}{1/\alpha_j + \alpha_i} \quad (i = 1,\ldots,n). \quad (9)$$

Then starting with initial parameter values $\alpha_1^{(0)},\ldots,\alpha_n^{(0)}$ and using the observed degree sequence $d_1,\ldots,d_n$, which is an inner point of $\mathscr{D}_n$, the iteration is as follows:

$$\alpha_i^{(t)} = \frac{d_i}{\sum_{j\neq i}\left(1/\left(1/\alpha_j^{(t-1)} + \alpha_i^{(t-1)}\right)\right)} \quad (i = 1,\ldots,n) \quad (10)$$

for $t = 1, 2, \ldots$, until convergence.

2.2. β-γ Model for Bipartite Graphs. This bipartite graph model, since there is a one-to-one correspondence between bipartite graphs and 0-1 rectangular arrays, traces back to Haberman [34], Lauritzen [29], and Rasch [22, 23] who applied it for psychological and educational measurements, later market research. According to the Rasch model, the entries of an $m \times n$ binary table $\mathbf{A}$ are independent Bernoulli random variables, where for the parameter p_{ij} of the entry A_{ij} the following holds:

$$\ln\frac{p_{ij}}{1-p_{ij}} = \beta_i - \delta_j \quad (i = 1,\ldots,m;\ j = 1,\ldots,n) \quad (11)$$

with real parameters $\beta_1, \ldots, \beta_m$ and $\delta_1, \ldots, \delta_n$. As an example, Rasch in [22] investigated binary tables where the rows corresponded to persons and the columns to items of some psychological test, whereas the jth entry of the ith row was 1 if person i answered test item j correctly and 0 otherwise. He also gave a description of the parameters: β_i was the ability of person i, while δ_j was the difficulty of test item j. Therefore, in view of model equation (11), the more intelligent the person is and the less difficult the test is, the larger the success/failure ratio was on a logarithmic scale.

Given an $m \times n$ random binary table $\mathbf{A} = (A_{ij})$, or, equivalently, a bipartite graph, our model is

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \gamma_j \quad (i = 1, \ldots, m, \ j = 1, \ldots, n) \tag{12}$$

with real parameters $\beta_1, \ldots, \beta_m$ and $\gamma_1, \ldots, \gamma_n$; further, $p_{ij} = \mathbb{P}(A_{ij} = 1)$.

In terms of the transformed parameters $b_i = e^{\beta_i}$ and $g_j = e^{\gamma_j}$, model (12) is equivalent to

$$\frac{p_{ij}}{1 - p_{ij}} = b_i g_j \quad (i = 1, \ldots, m, \ j = 1, \ldots, n), \tag{13}$$

where $b_1, \ldots, b_m$ and $g_1, \ldots, g_n$ are positive reals.

Conversely, the probabilities can be expressed in terms of the parameters:

$$\begin{aligned} p_{ij} &= \frac{b_i g_j}{1 + b_i g_j}, \\ 1 - p_{ij} &= \frac{1}{1 + b_i g_j}. \end{aligned} \tag{14}$$

Observe that if (12) holds with the parameters β_i's and γ_j's, then it also holds with the transformed parameters $\beta_i' = \beta_i + c$ $(i = 1, \ldots, m)$ and $\gamma_j' = \gamma_j - c$ $(j = 1, \ldots, n)$ with some $c \in \mathbb{R}$. Equivalently, if (13) holds with the positive parameters b_i's and g_j's, then it also holds with the transformed parameters

$$\begin{aligned} b_i' &= b_i \kappa, \\ g_j' &= \frac{g_j}{\kappa} \end{aligned} \tag{15}$$

with some $\kappa > 0$. Therefore, the parameters b_i and g_j are arbitrary to within a multiplicative constant.

Here the row-sums $R_i = \sum_{j=1}^{n} A_{ij}$ and the column-sums $C_j = \sum_{i=1}^{m} A_{ij}$ are the sufficient statistics for the parameters collected in $\mathbf{b} = (b_1, \ldots, b_m)$ and $\mathbf{g} = (g_1, \ldots, g_n)$. Indeed, the likelihood function is factorized as

$$\begin{aligned} L_{\mathbf{b},\mathbf{g}}(\mathbf{A}) &= \prod_{i=1}^{m} \prod_{j=1}^{n} p_{ij}^{A_{ij}} \left(1 - p_{ij}\right)^{1 - A_{ij}} \\ &= \left\{ \prod_{i=1}^{m} \prod_{j=1}^{n} \left(\frac{p_{ij}}{1 - p_{ij}} \right)^{A_{ij}} \right\} \prod_{i=1}^{m} \prod_{j=1}^{n} \left(1 - p_{ij}\right) \\ &= \left\{ \prod_{i=1}^{m} b_i^{\sum_{j=1}^{n} A_{ij}} \right\} \left\{ \prod_{j=1}^{n} g_j^{\sum_{i=1}^{m} A_{ij}} \right\} \prod_{i=1}^{m} \prod_{j=1}^{n} \left(1 - p_{ij}\right) \\ &= \left\{ \prod_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{1 + b_i g_j} \right\} \left\{ \prod_{i=1}^{m} b_i^{R_i} \right\} \left\{ \prod_{j=1}^{n} g_j^{C_j} \right\}. \end{aligned} \tag{16}$$

Since the likelihood function depends on $\mathbf{A}$ only through its row- and column-sums, by the Neyman-Fisher factorization theorem, $R_1, \ldots, R_m$, $C_1, \ldots, C_n$ are a sufficient statistic for the parameters. The first factor (including the partition function) depends only on the parameters and the row- and column-sums, whereas the seemingly not present factor—which would depend merely on $\mathbf{A}$—is constantly 1, indicating that the conditional joint distribution of the entries, given the row- and column-sums, is uniform (microcanonical) in this model. Note that, in [35], the author characterizes random tables sampled uniformly from the set of 0-1 matrices with fixed margins. Given the margins, the contingency tables coming from the above model are uniformly distributed, and a typical table of this distribution is produced by the β-γ model with parameters estimated via the row- and column-sums as sufficient statistics. In this way, here we obtain another view of the typical table of [35].

Based on an observed binary table (a_{ij}), since we are in exponential family and $\beta_1, \ldots, \beta_m$, $\gamma_1, \ldots, \gamma_n$ are natural parameters, the likelihood equation is obtained by making the expectation of the sufficient statistic equal to its sample value. Therefore, with the notations $r_i = \sum_{j=1}^{n} a_{ij}$ $(i = 1, \ldots, m)$ and $c_j = \sum_{i=1}^{m} a_{ij}$ $(j = 1, \ldots, n)$, the following *system of likelihood equations* is yielded:

$$\begin{aligned} r_i &= \sum_{j=1}^{n} \frac{b_i g_j}{1 + b_i g_j} = b_i \sum_{j=1}^{n} \frac{1}{1/g_j + b_i}, \quad i = 1, \ldots, m; \\ c_j &= \sum_{i=1}^{m} \frac{b_i g_j}{1 + b_i g_j} = g_j \sum_{i=1}^{m} \frac{1}{1/b_i + g_j}, \quad j = 1, \ldots, n. \end{aligned} \tag{17}$$

Note that, for any sample realization of $\mathbf{A}$,

$$\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j \tag{18}$$

holds automatically. Therefore, there is dependence between the equations of system (17), indicating that the solution is not unique, in accord with our previous remark about the arbitrary scaling factor $\kappa > 0$ of (15). We will prove that, apart from this scaling, the solution is unique if it exists at all.

For our convenience, let $(\widetilde{\mathbf{b}}, \widetilde{\mathbf{g}})$ denote the equivalence class of the parameter vector $(\mathbf{b}, \mathbf{g})$, which consists of parameter vectors $(\mathbf{b}', \mathbf{g}')$ satisfying (15) with some $\kappa > 0$. So to avoid this indeterminacy, we may impose conditions on the parameters; for example,

$$\sum_{i=1}^{m} \beta_i + \sum_{j=1}^{n} \gamma_j = 0. \tag{19}$$

Like the graphic sequences, here the following sufficient conditions can be given for the sequences $r_1 \geq \cdots \geq r_m > 0$ and $c_1 \geq \cdots \geq c_n > 0$ of integers to be row- and column-sums of an $m \times n$ matrix of 0-1 entry (see, e.g., [36]):

$$\begin{aligned} \sum_{i=1}^{k} r_i &\leq \sum_{j=1}^{n} \min\left\{c_j, k\right\}, \quad k = 1, \ldots, m; \\ \sum_{j=1}^{k} c_j &\leq \sum_{i=1}^{m} \min\left\{r_i, k\right\}, \quad k = 1, \ldots, n. \end{aligned} \tag{20}$$

Observe that the $k = 1$ cases imply $r_1 \leq n$ and $c_1 \leq m$, whereas the $k = m$ and $k = n$ cases together imply $\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$. This statement is the counterpart of the Erdős-Gallai conditions for bipartite graphs, where—due to (18)—the sum of the degrees is automatically even. In fact, the conditions in (20) are redundant: one of the conditions—either the one for the rows or the one for the columns—suffices together with (18) and $c_1 \leq m$ or $r_1 \leq n$. The so obtained necessary and sufficient conditions define *bipartite realizable sequences* with the wording of [33]. Already, in 1957, the author of [37] determined arithmetic conditions for the construction of a 0-1 matrix having given row- and column-sums. The construction was given via swaps. More generally, [38] referred to the transportation problem and the Ford-Fulkerson max flow–min cut theorem [39].

The convex hull of the bipartite realizable sequences $\mathbf{r} = (r_1, \ldots, r_m)$ and $\mathbf{c} = (c_1, \ldots, c_n)$ forms a polytope in $\mathbb{R}^{m+n}$, actually, because of (18), in its $(m + n - 1)$-dimensional hyperplane. It is called *polytope of bipartite degree sequences* and denoted by $\mathscr{P}_{m,n}$ in [33]. It is the special case of the transportation polytope describing margins of contingency tables with nonnegative integer entries. There is an expanding literature on the number of such matrices, for example, [40], and on the number of 0-1 matrices with prescribed row- and column-sums, for example, [41].

Analogously to the considerations of the α-β models, and applying the thoughts of the proofs in [8–10], $\mathscr{P}_{m,n}$ is the closure of the set of the expected row- and column-sum sequences in the above model. In [33], it is proved that an $m \times n$ binary table, or equivalently a bipartite graph on the independent sets of m and n vertices, is on the boundary of $\mathscr{P}_{m,n}$ if it does not contain two vertex-disjoint edges. In this case, the likelihood function cannot be maximized with a finite parameter set; its supremum is approached with a parameter vector with at least one coordinate, β_i or γ_j, tending to $+\infty$ or $-\infty$, or, equivalently, with at least one coordinate, b_i or g_j, tending to $+\infty$ or 0. Based on the proofs of [10], and stated as Theorem 6.3 in the supplementary material of [10], the maximum likelihood estimate of the parameters of model (13) exists if and only if the observed row- and column-sum sequence $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}(\mathscr{P}_{m,n})$, the relative interior of $\mathscr{P}_{m,n}$, satisfying (18). In this case for the probabilities, calculated by formula (14) through the estimated positive parameter values $\widehat{b}_i$'s and $\widehat{g}_j$'s (solutions of (17)), $0 < p_{ij} < 1$ holds $\forall i, j$.

Under these conditions, we define an algorithm that converges to the unique (up to the above equivalence) solution of maximum likelihood equations (17).

Theorem 1. *If $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}(\mathscr{P}_{m,n})$, then the following algorithm gives a unique equivalence class of the parameter vectors as the fixed point of the iteration, which therefore provides the ML estimate of the parameters.*

Starting with positive parameter values $b_i^{(0)}$ $(i = 1, \ldots, m)$ and $g_j^{(0)}$ $(j = 1, \ldots, n)$ and using the observed row- and column-sums, the iteration is as follows:

$$(I) \quad b_i^{(t)} = \frac{r_i}{\sum_{j=1}^{n} \left(1/\left(1/g_j^{(t-1)} + b_i^{(t-1)}\right)\right)}, \quad i = 1, \ldots, m \tag{21}$$

$$(II) \quad g_j^{(t)} = \frac{c_j}{\sum_{i=1}^{m} \left(1/\left(1/b_i^{(t)} + g_j^{(t-1)}\right)\right)}, \quad j = 1, \ldots, n$$

for $t = 1, 2, \ldots$, until convergence.

3. Parameter Estimation in the LBM

In the several clusters' case, we are putting the bricks together. The above discussed α-β and β-γ models will be the building blocks of the LBM to be introduced. Here the degree sequences are not any more sufficient for the whole graph, only for the building blocks of the subgraphs.

Given $1 \leq k \leq n$, we are looking for k-partition, in other words, clusters $C_1, \ldots, C_k$ of the vertices, such that

(i) different vertices are independently assigned to a cluster C_u with probability π_u $(u = 1, \ldots, k)$, where $\sum_{u=1}^{k} \pi_u = 1$;

(ii) given the cluster memberships vertices $i \in C_u$ and $j \in C_v$ are connected independently, with probability p_{ij} such that

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_{iv} + \beta_{ju}, \tag{22}$$

for any $1 \leq u, v \leq k$ pair. Equivalently,

$$\frac{p_{ij}}{1 - p_{ij}} = b_{ic_j} b_{jc_i}, \tag{23}$$

where c_i is the cluster membership of vertex i and $b_{iv} = e^{\beta_{iv}}$.

The parameters are collected in the vector $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ and the $n \times k$ matrix $\mathbf{B}$ of b_{iu}'s ($i \in C_u$, $u = 1, \ldots, k$). The likelihood function is the following mixture:

$$\sum_{1 \le u, v \le k} \pi_u \pi_v \prod_{i \in C_u, j \in C_v} p_{ij}^{a_{ij}} \left(1 - p_{ij}\right)^{(1-a_{ij})}. \tag{24}$$

Here $\mathbf{A} = (a_{ij})$ is the incomplete data specification as the cluster memberships are missing. Therefore, it is straightforward to use the EM algorithm, proposed by [24], also discussed in [7, 42], for parameter estimation from incomplete data. This special application for mixtures is sometimes called *collaborative filtering* (see [25, 26]) which is rather applicable to fuzzy clustering.

First we complete our data matrix $\mathbf{A}$ with latent membership vectors $\mathbf{m}_1, \ldots, \mathbf{m}_n$ of the vertices that are k-dimensional i.i.d. Multy$(1, \underline{\pi})$ (multinomially distributed) random vectors. More precisely, $\mathbf{m}_i = (m_{i1}, \ldots, m_{ik})$, where $m_{iu} = 1$ if $i \in C_u$ and zero otherwise. Thus, the sum of the coordinates of any $\mathbf{m}_i$ is 1, and $\mathbb{P}(m_{iu} = 1) = \pi_u$.

Note that if the cluster memberships were known, then the complete likelihood would be

$$\prod_{u=1}^{k} \prod_{i=1}^{n} \prod_{v=1}^{k} \prod_{j=1}^{n} \left[p_{ij}^{m_{jv} a_{ij}} \cdot \left(1 - p_{ij}\right)^{m_{jv}(1-a_{ij})} \right]^{m_{iu}} \tag{25}$$

that is valid only in case of known cluster memberships.

Starting with initial parameter values $\underline{\pi}^{(0)}$ and $\mathbf{B}^{(0)}$ and membership vectors $\mathbf{m}_1^{(0)}, \ldots, \mathbf{m}_n^{(0)}$, the tth step of the iteration is the following ($t = 1, 2, \ldots$):

(i) E-step: we calculate the conditional expectation of each $\mathbf{m}_i$ conditioned on the model parameters and on the other cluster assignments obtained in step $t - 1$ and collectively denoted by $M^{(t-1)}$.

The responsibility of vertex i for cluster u in the tth step is defined as the conditional expectation $\pi_{iu}^{(t)} = \mathbb{E}(m_{iu} \mid M^{(t-1)})$, and, by the Bayes theorem, it is

$$\pi_{iu}^{(t)} = \frac{\mathbb{P}\left(M^{(t-1)} \mid m_{iu} = 1\right) \cdot \pi_u^{(t-1)}}{\sum_{v=1}^{k} \mathbb{P}\left(M^{(t-1)} \mid m_{iv} = 1\right) \cdot \pi_v^{(t-1)}} \tag{26}$$

($u = 1, \ldots, k$; $i = 1, \ldots, n$). For each i, $\pi_{iu}^{(t)}$ is proportional to the numerator; therefore the conditional probabilities $\mathbb{P}(M^{(t-1)} | m_{iu} = 1)$ should be calculated for $u = 1, \ldots, k$. But this is just the part of likelihood (25) affecting vertex i under the condition $m_{iu} = 1$. Therefore,

$$\mathbb{P}\left(M^{(t-1)} \mid m_{iu} = 1\right) = \prod_{v=1}^{k} \prod_{j \in C_v, j \sim i} \frac{b_{iv}^{(t-1)} b_{ju}^{(t-1)}}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \cdot \prod_{j \in C_v, j \not\sim i} \frac{1}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} = \prod_{v=1}^{k} \left\{ \frac{b_{iv}^{(t-1)} b_{ju}^{(t-1)}}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \right\}^{e_{vi}} \cdot \left\{ \frac{1}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \right\}^{|C_v| \cdot (|C_v| - 1)/2 - e_{vi}}, \tag{27}$$

where e_{vi} is the number of edges within C_v that are connected to i.

(ii) M-step: we update $\underline{\pi}^{(t)}$ and $\mathbf{m}^{(t)}$: $\pi_u^{(t)} := (1/n) \sum_{i=1}^{n} \pi_{iu}^{(t)}$ and $m_{iu}^{(t)} = 1$ if $\pi_{iu}^{(t)} = \max_v \pi_{iv}^{(t)}$ and 0 otherwise (in case of ambiguity, we select the smallest index for the cluster membership of vertex i). This is an ML estimation (discrete one, in the latter case, for the cluster membership). In this way, new clustering of the vertices is obtained.

Then we estimate the parameters in the actual clustering of the vertices. In the within-cluster scenario, we use the parameter estimation of model (3), obtaining estimates of b_{iu}'s ($i \in C_u$) in each cluster separately ($u = 1, \ldots, k$); as for cluster u, b_{iu} corresponds to α_i and the number of vertices is $|C_u|$. In the between-cluster scenario, we use bipartite graph model (13) in the following way. For $u < v$, edges connecting vertices of C_u and C_v form a bipartite graph, based on which the parameters b_{iv} ($i \in C_u$) and b_{ju} ($j \in C_v$) are estimated with the above algorithm; here b_{iv}'s correspond to b_i's, b_{ju}'s correspond to g_j's, and the number of rows and columns of the rectangular array corresponding to this bipartite subgraph of $\mathbf{A}$ is $|C_u|$ and $|C_v|$, respectively. With the estimated parameters, collected in the $n \times k$ matrix $\mathbf{B}^{(t)}$, we go back to the E-step and so forth.

As in the M-step we increase the likelihood in all parts, and in the E-step we relocate the vertices into the cluster where their likelihoods are maximized, the nonnegative likelihood function is increased in each iteration. Since the likelihood function is bounded from above (unless in some inner cycle we start from the boundary of a polytope of Section 2), it must converge to a local maximum.

Note that here the parameter β_{iv} with $c_i = u$ embodies the affinity of vertex i of cluster C_u towards vertices of cluster C_v; and likewise, β_{ju} with $c_j = v$ embodies the affinity of vertex j of cluster C_v towards vertices of cluster C_u. By the model, these affinities are added together on the level of the logits. This so-called k-β model, introduced in [43], is applicable to social networks, where attitudes of individuals in the same social group (say, u) are the same toward members of another social group (say, v), though this attitude also depends on the individual in group u. The model may also be applied to biological networks, where the clusters correspond, for example, to different functioning synopses or other units of the brain; see [44].

After normalizing the β_{iv} ($i \in C_u$) and β_{ju} ($j \in C_v$) to meet the requirement of (19), for any $u \neq v$ pair, the sum of the parameters will be zero, and their sign and magnitude indicate the affinity of nodes of C_u to make ties with the nodes of C_v and vice versa:

$$\sum_{i \in C_u} \beta_{iv} + \sum_{j \in C_v} \beta_{ju} = 0. \tag{28}$$

This becomes important when we want to compare the parameters corresponding to different cluster pairs. For

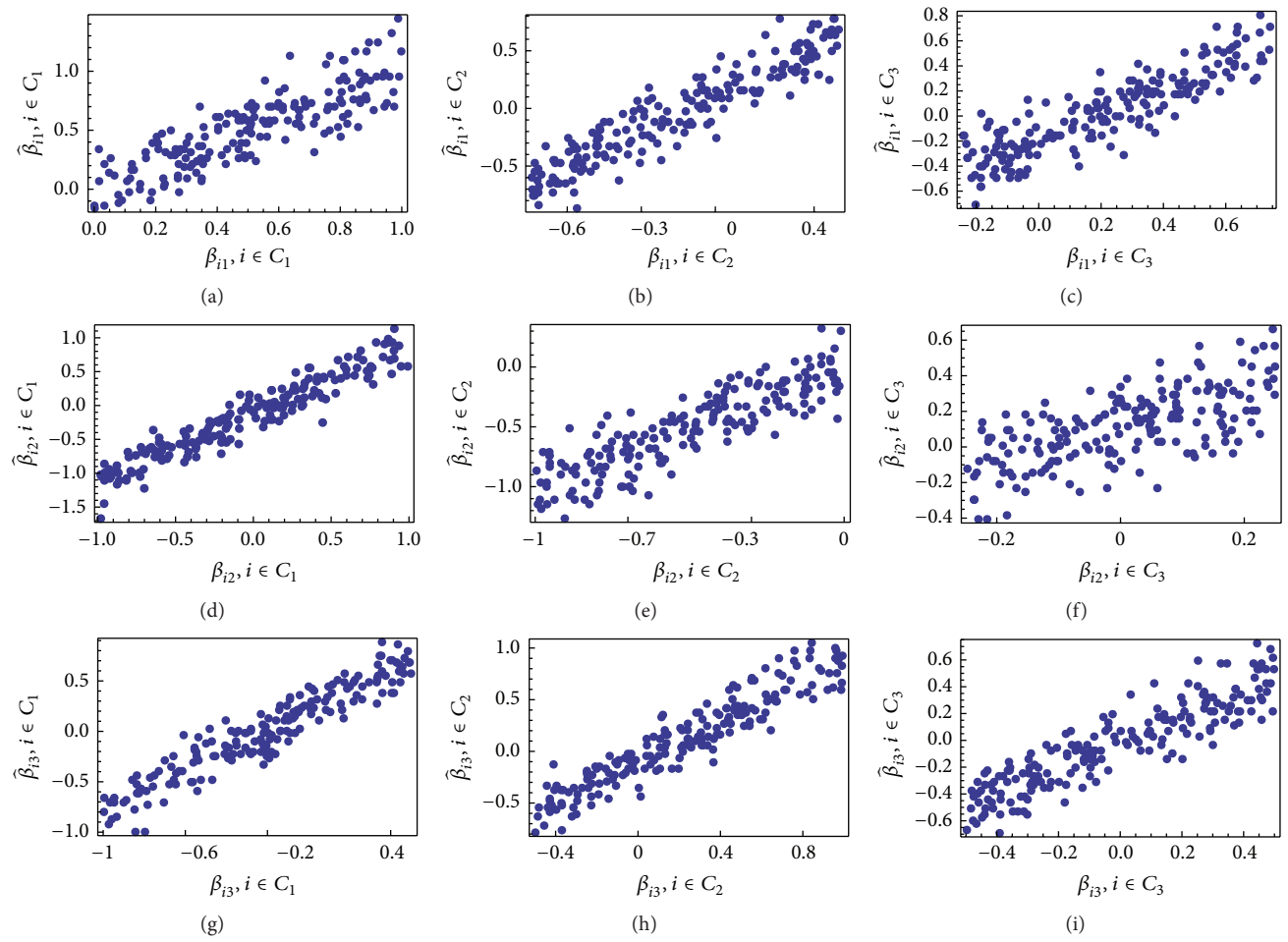

FIGURE 1: Data were generated based on parameters β_{iv}'s chosen uniformly in different intervals, $k = 3$, $|C_1| = 190$, $|C_2| = 193$, and $|C_3| = 197$. The estimated versus the original parameters β_{iv}'s are shown for $i \in C_u$ $(u, v = 1, \ldots, k)$, where $\beta_{i1} \sim \mathcal{U}[0, 1]$ $(i \in C_1)$, $\beta_{i1} \sim \mathcal{U}[-0.75, 0.5]$ $(i \in C_2)$, $\beta_{i1} \sim \mathcal{U}[-0.25, 0.75]$ $(i \in C_3)$, $\beta_{i2} \sim \mathcal{U}[-1, 1]$ $(i \in C_1)$, $\beta_{i2} \sim \mathcal{U}[-1, 0]$ $(i \in C_2)$, $\beta_{i2} \sim \mathcal{U}[-0.25, 0.25]$ $(i \in C_3)$, $\beta_{i3} \sim \mathcal{U}[-1, 0.5]$ $(i \in C_1)$, $\beta_{i3} \sim \mathcal{U}[-0.5, 1]$ $(i \in C_2)$, and $\beta_{i3} \sim \mathcal{U}[-0.5, 0.5]$ $(i \in C_3)$, respectively. MSE = 1.14634.

selecting the initial number of clusters, we can use considerations of [45], while for the initial clustering, we can use spectral clustering tools of [5].

4. Applications

Now we illustrate the performance of our algorithm via randomly generated and real-world data. Note that while processing the iteration, we sometimes run into threshold subgraphs or bipartite subgraphs on the boundary of the polytope of bipartite degree sequences. Even in this case, our iteration converged for most coordinates of the parameter vectors, while some b_{iv} coordinates tended to $+\infty$ or 0 (numerically, when stopping the iteration, they took on a very "large" or "small" value). This means that the affinity of node i towards nodes of the cluster j is infinitely "large" or "small"; that is, this node is liable to always or never make ties with nodes of cluster j.

First we generated a random graph on $n = 580$ vertices and with $k = 3$ underlying vertex-clusters C_1, C_2, C_3 in the following way. Let $|C_1| := 190$, $|C_2| := 193$, and $|C_3| := 197$. The parameters β_{i1} $(i \in C_1)$, β_{i2} $(i \in C_1)$, and β_{i3} $(i \in C_1)$ were chosen independently at uniform from the intervals $[0, 1]$, $[-1, 1]$, and $[-1, 0.5]$, respectively. The parameters β_{i1} $(i \in C_2)$, β_{i2} $(i \in C_2)$, and β_{i3} $(i \in C_2)$ were chosen independently at uniform from the intervals $[-0.75, 0.5]$, $[-1, 0]$, and $[-0.5, 1]$, respectively. The parameters β_{i1} $(i \in C_3)$, β_{i2} $(i \in C_3)$, and β_{i3} $(i \in C_3)$ were chosen independently at uniform from the intervals $[-0.25, 0.75]$, $[-0.25, 0.25]$, and $[-0.5, 0.5]$, respectively.

Starting with 3 clusters, obtained by spectral clustering tools, and initial parameter values collected in $\mathbf{B}^{(0)}$ of all 1 entries, after some outer steps, the iteration converged to $\widehat{\mathbf{B}} = (\widehat{b}_{iv})$. With $\widehat{\beta}_{iv} = \ln \widehat{b}_{iv}$, we plotted the $\beta_{iv}, \widehat{\beta}_{iv}$ pairs for $i \in C_u$, $u, v = 1, 2, 3$. Figure 1 shows a good fit with MSE = 1.14634 of the estimated parameters to the original ones. Indeed, by the general theory of the ML estimation [6], for "large" n, the ML estimate should approach the true parameter, based on which the model was generated. So the good fit means that our algorithm finds estimates close to the true parameters.

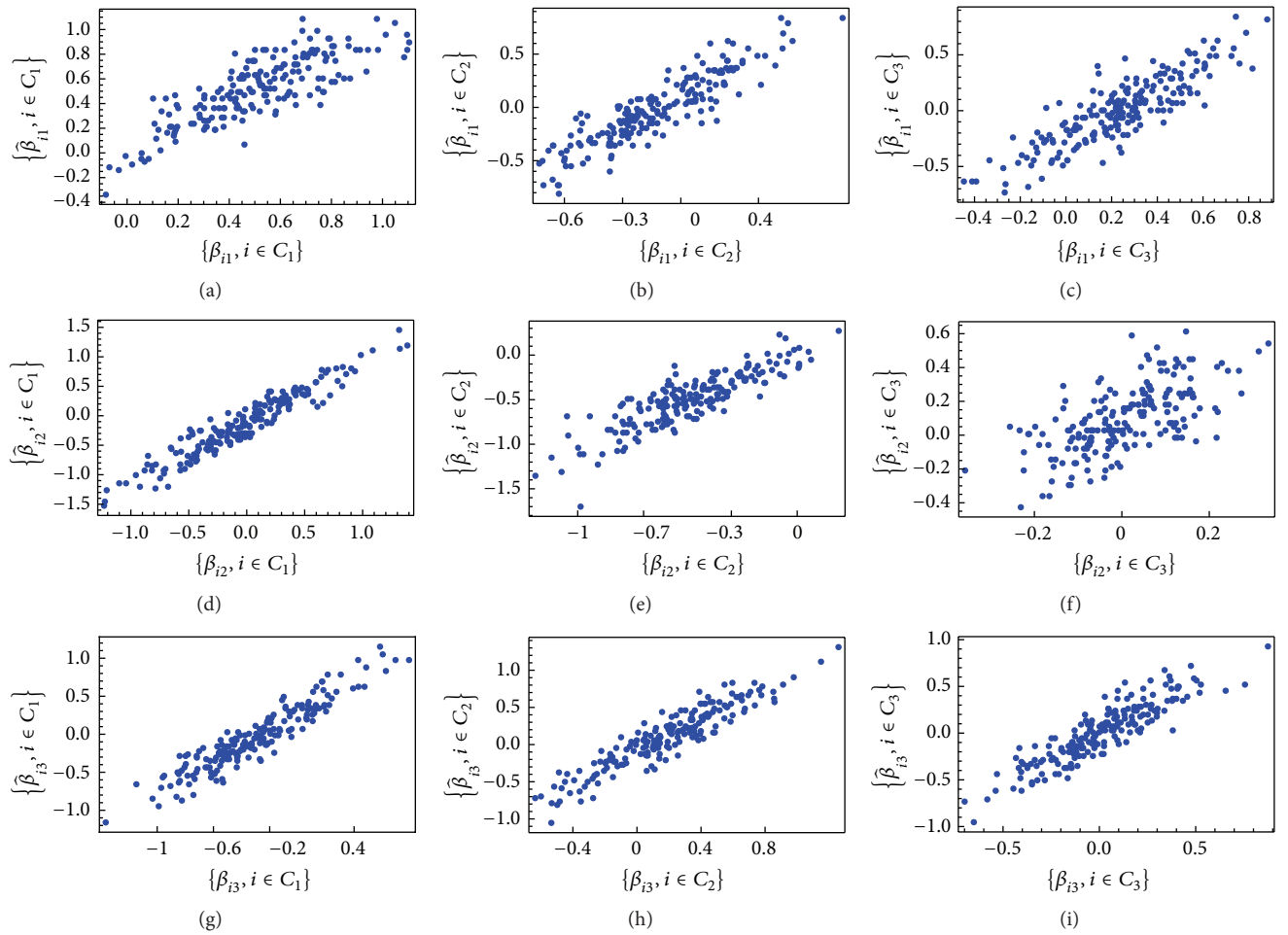

FIGURE 2: Data were generated based on parameters β_{iv}'s following Gaussian distribution with different parameters for the within- and between-cluster relations, $k = 3$, $|C_1| = 190$, $|C_2| = 193$, and $|C_3| = 197$. The estimated versus the original parameters β_{iv}'s are shown for $i \in C_u$ $(u, v = 1, \ldots, k)$, where $\beta_{i1} \sim \mathcal{N}(0.5, 0.25)$ $(i \in C_1)$, $\beta_{i1} \sim \mathcal{N}(-0.125, 0.312)$ $(i \in C_2)$, $\beta_{i1} \sim \mathcal{N}(0.25, 0.25)$ $(i \in C_3)$, $\beta_{i2} \sim \mathcal{N}(0, 0.5)$ $(i \in C_1)$, $\beta_{i2} \sim \mathcal{N}(-0.5, 0.25)$ $(i \in C_2)$, $\beta_{i2} \sim \mathcal{N}(0, 0.125)$ $(i \in C_3)$, $\beta_{i3} \sim \mathcal{N}(-0.25, 0.375)$ $(i \in C_1)$, $\beta_{i3} \sim \mathcal{N}(0.25, 0.375)$ $(i \in C_2)$, and $\beta_{i3} \sim \mathcal{N}(0, 0.25)$ $(i \in C_3)$, respectively. MSE = 1.12556.

Then we generated the same size of a random graph, where the initial parameters followed Gaussian distribution, with different parameters for the within- and between-cluster relations. Based on the parameters we calculated the edge probabilities, and we generated a random graph with them. Eventually, we estimated the parameters with our algorithm; see Figure 2. The Gaussian data are about in the same ranges as the uniform ones; however, they are better concentrated to their means. It can be the cause of a bit smaller MSE = 1.12556.

Figure 3 shows the resulting clusters obtained by applying the LBM algorithm to the B&K fraternity data [46] with $n = 58$ vertices; see also http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#bkfrat. The data, collected by Bernard and Killworth, are behavioral frequency counts, based on communication frequencies between students of a college fraternity in Morgantown, West Virginia. We used the binarized version of the symmetric edge-weight matrix. When the data were collected, the 58 occupants had been living together for at least three months, but senior students had been living there for up to three years. We used our normalized modularity based spectral clustering algorithm [4] to find the starting clusters. In the normalized modularity spectrum we found a gap after the third eigenvalue (in decreasing order of their absolute values); therefore we applied the algorithm with $k = 4$ clusters. The four groups are likely to consist of persons living together for about the same time period.

While processing the iteration, occasionally we bumped into the situation when the degree sequence lied on the boundary of the convex polytopes defined in Sections 2.1 and 2.2. Unfortunately, this can occur when our graph is not dense enough. In these situations, the iteration did not converge for some coordinates β_{iv}, but they seemed to tend to $+\infty$ or $-\infty$. Equivalently, the corresponding b_{iv} for some $i \in C_u$ and v tended to $+\infty$ or 0, yielding the situation that

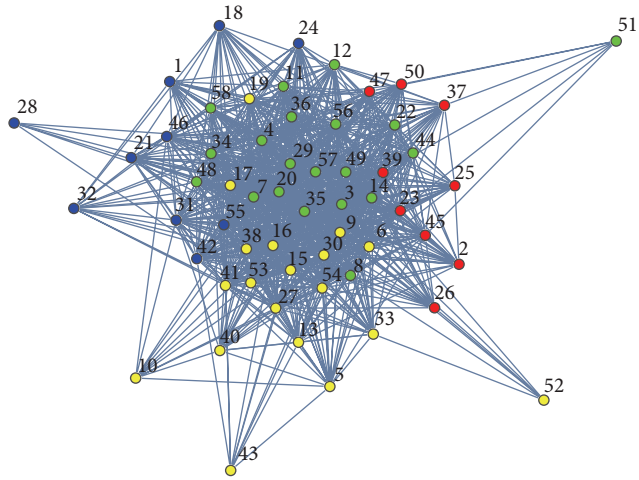

Figure 3: The 4 clusters found by the LBM algorithm in the B&K fraternity data, with 10, 9, 20, and 19 students in the clusters, respectively. RAND index = 1 between the SC and LBM.

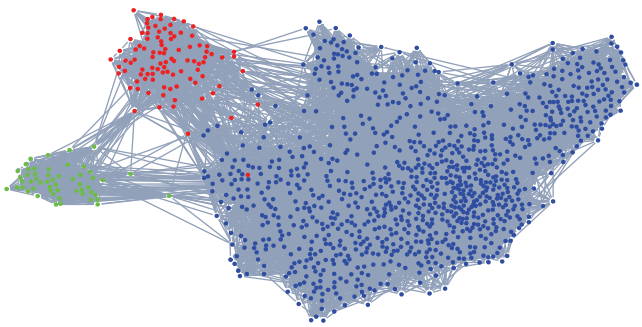

Figure 4: The 3 clusters found by the LBM algorithm in the network of the Last.fm users with 1012, 97, and 53 users in the clusters, respectively. RAND index = 0.99205 between the SC and LBM.

member $i \in C_u$ had $+\infty$ or 0 affinity towards members of C_v. Another way, recommended in [8], is to add a small amount to each degree to avoid this situation. However, we did not want to manipulate the original graph, which was too sparse to produce degree sequences in the interior of one or more polytopes.

We also analyzed the network based on the friendships between the users of the Last.fm music recommendation system [47]. Last.fm is an online service in music based social networking. Each user may have friends inside the Last.fm social network, and so they form a timestamped undirected graph. In 2012, there were 71,000 users and 285,241 edges between them. Actually, we only used the 15-core of this graph for clustering. Starting with SC, the LBM iteration refined the three underlying clusters; see Figure 4.

We clustered the vertices of the above networks with the EM algorithm and estimated the parameters in both the LBM and SBM (for the iteration of the letter one, see Chapter 5 of [5]). For the initial clustering we used SC with number of clusters k such that we found a remarkable gap in the normalized modularity spectrum between $|\mu_{k-1}|$ and $|\mu_k|$. During the iteration some clusters may become empty which reduces k; it was not the case in our iterations. It is also possible to start with different values of k; here we started with the smallest possible k which indicated a gap. In case of the B&K fraternity data, the leading eigenvalues in decreasing absolute values (apart from the trivial 1) were $\mu_1 = 0.235826$, $\mu_2 = -0.228652$, $\mu_3 = 0.223039$, $\mu_4 = -0.198867$, and $\mu_5 = 0.194783$, indicating a gap between $|\mu_3|$ and $|\mu_4|$, so we selected $k = 4$. In case of the Last.fm data, the leading eigenvalues in decreasing absolute values (apart from the trivial 1) were $\mu_1 = 0.97061$, $\mu_2 = 0.942929$, $\mu_3 = 0.892111$, and $\mu_4 = 0.862594$, indicating a gap between $|\mu_2|$ and $|\mu_3|$, so we selected $k = 3$.

After some outer iterations both the LBM and SBM converged to a local maximum. We compared the clustering obtained by SC versus LBM and SC versus SBM via the Rand index introduced in [48]. This index is between 0 and 1, and the larger it is, the better the agreement between the two clusterings is. We found a good agreement between the SC clusters and those of the LBM: RAND = 1 in Figure 3 and RAND = 0.99205 in Figure 4, whereas, between the SC clustering and SBM clustering, we obtained RAND = 0.61525 for the B&K fraternity data and RAND = 0.96912 for the Last.fm data. This shows that LBM is better fine tuning of the spectral clustering than SBM, at least, in these examples, where the diversity of the degrees is present.

5. Discussion

Our model is a profoundly heterogeneous kind of a block model, where the subgraphs and bipartite subgraphs obey parametric graph models, within which the connections are mainly determined by the degrees. The EM type algorithm introduced here finds the blocks and estimates the parameters at the same time.

When investigating controllability of large networks, the authors of [49] observe and prove that a system's controllability is to a great extent encoded by the underlying network's degree distribution. In our model, this is true only for the building blocks. Possibly, the blocks could be controlled separately, based on the degree sequences of the subgraphs.

Our model is applicable to large inhomogeneous networks, and above finding clusters of the vertices, it also assigns multiscale parameters to them. In social networks, these parameters can be associated with attitudes of persons of one group towards those in the same or another group. The attitudes are, in fact, affinities to make ties.

We prove the convergence of the inner cycle of the algorithm to the unique solution of the maximum likelihood equation within the subgraphs and bipartite subgraphs. Then, by the iteration of the EM algorithm, the convergence of the outer cycle to a local maximum of the overall likelihood is straightforward. As there can be several local maxima, good starting is important. Our final clusters show a good agreement with the spectral clusters; therefore, the algorithm can be considered as a refinement of the spectral clustering and gives estimates of the parameters which provide a local

maximum of the overall LBM likelihood with clusters near to the spectral ones.

Appendix

Proof of Theorem 1. To show the convergence, we rewrite the iteration as the series of $(\phi, \psi) : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ maps, where $\phi = (\phi_1, \ldots, \phi_m)$ and $\psi = (\psi_1, \ldots, \psi_n)$; further ψ depends on ϕ such that

$$b_i^{(t)} = \phi_i\left(\mathbf{b}^{(t-1)}, \mathbf{g}^{(t-1)}\right), \quad i = 1, \ldots, m,$$
$$g_j^{(t)} = \psi_j\left(\mathbf{b}^{(t)}, \mathbf{g}^{(t-1)}\right) = \psi_j\left(\phi\left(\mathbf{b}^{(t-1)}, \mathbf{g}^{(t-1)}\right), \mathbf{g}^{(t-1)}\right), \quad j = 1, \ldots, n. \tag{A.1}$$

We define

$$\rho\left((\mathbf{b}, \mathbf{g}), \left(\mathbf{b}', \mathbf{g}'\right)\right) = \max\left\{\max\left\{\max_{1\le i\le m}\frac{b_i}{b_i'}, \max_{1\le i\le m}\frac{b_i'}{b_i}\right\}, \max\left\{\max_{1\le j\le n}\frac{g_j}{g_j'}, \max_{1\le j\le n}\frac{g_j'}{g_j}\right\}\right\}. \tag{A.2}$$

It is easy to see that $\rho \ge 1$ and $\rho = 1$ if and only if $(\mathbf{b}, \mathbf{g}) = (\mathbf{b}', \mathbf{g}')$; further, $\log \rho$ is a metric. We will use the following lemma of [9]. □

Lemma A.1. *For any integer $n > 1$ and arbitrary positive real numbers $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$, we have*

$$\frac{u_1 + \cdots + u_n}{v_1 + \cdots + v_n} \le \max_{1\le i\le n}\frac{u_i}{v_i}, \tag{A.3}$$

and equality holds if and only if the ratios u_i/v_i have the same value.

Now we prove that the (ϕ, ψ) map is a weak contraction in the $\log \rho$ metric.

Step 1. Applying Lemma A.1 twice (first with n and then with two terms),

$$\frac{\phi_i(\mathbf{b}, \mathbf{g})}{\phi_i(\mathbf{b}', \mathbf{g}')} = \frac{r_i\left(\sum_{j=1}^n \left(1/\left(1/g_j + b_i\right)\right)\right)^{-1}}{r_i\left(\sum_{j=1}^n \left(1/\left(1/g_j' + b_i'\right)\right)\right)^{-1}} = \frac{\sum_{j=1}^n \left(1/\left(1/g_j' + b_i'\right)\right)}{\sum_{j=1}^n \left(1/\left(1/g_j + b_i\right)\right)} \le \max_{1\le j\le n}\frac{1/\left(1/g_j' + b_i'\right)}{1/\left(1/g_j + b_i\right)} = \max_{1\le j\le n}\frac{1/g_j + b_i}{1/g_j' + b_i'} \le \max_{1\le j\le n}\max\left\{\frac{g_j'}{g_j}, \frac{b_i}{b_i'}\right\} = \max\left\{\max_{1\le j\le n}\frac{g_j'}{g_j}, \frac{b_i}{b_i'}\right\}. \tag{A.4}$$

Likewise,

$$\frac{\phi_i\left(\mathbf{b}', \mathbf{g}'\right)}{\phi_i(\mathbf{b}, \mathbf{g})} \le \max_{1\le j\le n}\max\left\{\frac{g_j}{g_j'}, \frac{b_i'}{b_i}\right\} = \max\left\{\max_{1\le j\le n}\frac{g_j}{g_j'}, \frac{b_i'}{b_i}\right\}. \tag{A.5}$$

Assume that $\rho((\mathbf{b}, \mathbf{g}), (\mathbf{b}', \mathbf{g}')) = \kappa$ and $\kappa > 1$; otherwise, when $\kappa = 1$, we already have the fixed point and there is nothing to prove. In view of the above calculations and (A.2),

$$\rho\left(\left(\phi(\mathbf{b}, \mathbf{g}), \mathbf{g}\right), \left(\phi\left(\mathbf{b}', \mathbf{g}'\right), \mathbf{g}'\right)\right) \le \kappa, \tag{A.6}$$

and the inequality can be attained with equality only if at least one of the following holds:

(1)
$$\text{(a)} \quad \max_i \frac{\phi_i(\mathbf{b}, \mathbf{g})}{\phi_i(\mathbf{b}', \mathbf{g}')} = \kappa \quad \text{or} \qquad \text{(b)} \quad \max_i \frac{\phi_i\left(\mathbf{b}', \mathbf{g}'\right)}{\phi_i(\mathbf{b}, \mathbf{g})} = \kappa. \tag{A.7}$$

(2)
$$\text{(a)} \quad \max_j \frac{g_j}{g_j'} = \kappa \quad \text{or} \qquad \text{(b)} \quad \max_j \frac{g_j'}{g_j} = \kappa. \tag{A.8}$$

1(a) is equivalent to the following: there is i such that $b_i/b_i' = \kappa$ and $g_j'/g_j = \kappa$, $\forall j$; meanwhile 1(b) is equivalent to the following: there is i such that $b_i'/b_i = \kappa$ and $g_j/g_j' = \kappa$, $\forall j$. 1(a) implies 2(b) and 1(b) implies 2(a). However, it cannot be that 2(a) or 2(b) hold, but 1(a) and 1(b) do not, since $\max_i(\phi_i(\mathbf{b}, \mathbf{g})/\phi_i(\mathbf{b}', \mathbf{g}')) = \kappa'$ with $1 < \kappa' < \kappa$ would result in $g_j'/g_j = \kappa'$, $\forall j$, which is in contradiction with 2(b); likewise, $\max_i(\phi_i(\mathbf{b}', \mathbf{g}')/\phi_i(\mathbf{b}, \mathbf{g})) = \kappa'$ with $1 < \kappa' < \kappa$ would result in $g_j/g_j' = \kappa'$, $\forall j$, which is in contradiction with 2(a). Therefore, it suffices to keep condition 1.

Step 2. Again applying Lemma A.1 twice (first with m and then with two terms),

$$\begin{aligned}
&\frac{\psi_j\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)}{\psi_j\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)}\\
&\quad=\frac{c_j\left(\sum_{i=1}^{m}\left(1/\left(1/\phi_i\left(\mathbf{b},\mathbf{g}\right)+g_j\right)\right)\right)^{-1}}{c_j\left(\sum_{i=1}^{m}\left(1/\left(1/\phi_i\left(\mathbf{b}',\mathbf{g}'\right)+g_j'\right)\right)\right)^{-1}}\\
&\quad=\frac{\sum_{i=1}^{m}\left(1/\left(1/\phi_i\left(\mathbf{b}',\mathbf{g}'\right)+g_j'\right)\right)}{\sum_{i=1}^{m}\left(1/\left(1/\phi_i\left(\mathbf{b},\mathbf{g}\right)+g_j\right)\right)}\\
&\quad\le\max_{1\le i\le m}\frac{1/\left(1/\phi_i\left(\mathbf{b}',\mathbf{g}'\right)+g_j'\right)}{1/\left(1/\phi_i\left(\mathbf{b},\mathbf{g}\right)+g_j\right)}\\
&\quad=\max_{1\le i\le m}\frac{1/\phi_i\left(\mathbf{b},\mathbf{g}\right)+g_j}{1/\phi_i\left(\mathbf{b}',\mathbf{g}'\right)+g_j'}\\
&\quad\le\max_{1\le i\le m}\max\left\{\frac{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)}{\phi_i\left(\mathbf{b},\mathbf{g}\right)},\frac{g_j}{g_j'}\right\}\\
&\quad=\max\left\{\max_{1\le i\le m}\frac{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)}{\phi_i\left(\mathbf{b},\mathbf{g}\right)},\frac{g_j}{g_j'}\right\}.
\end{aligned}\tag{A.9}$$

Likewise,

$$\begin{aligned}
\frac{\psi_j\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)}{\psi_j\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)}&\le\max_{1\le i\le m}\max\left\{\frac{\phi_i\left(\mathbf{b},\mathbf{g}\right)}{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)},\frac{g_j'}{g_j}\right\}\\
&=\max\left\{\max_{1\le i\le m}\frac{\phi_i\left(\mathbf{b},\mathbf{g}\right)}{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)},\frac{g_j'}{g_j}\right\}.
\end{aligned}\tag{A.10}$$

Therefore, in view of (A.2),

$$\begin{aligned}
&\rho\left(\left(\phi\left(\mathbf{b},\mathbf{g}\right),\psi\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)\right),\right.\\
&\quad\left.\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\psi\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)\right)\right)\le\rho\left(\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right),\\
&\quad\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)\right)\le\kappa
\end{aligned}\tag{A.11}$$

and both inequalities can be attained with equality only if at least one of the following holds:

(1)

$$\begin{aligned}
&\text{(a)}\quad\max_j\frac{\psi_j\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)}{\psi_j\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)}=\kappa\quad\text{or}\\
&\text{(b)}\quad\max_j\frac{\psi_j\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)}{\psi_j\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)}=\kappa.
\end{aligned}\tag{A.12}$$

(2)

$$\begin{aligned}
&\text{(a)}\quad\max_i\frac{\phi_i\left(\mathbf{b},\mathbf{g}\right)}{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)}=\kappa\quad\text{or}\\
&\text{(b)}\quad\max_i\frac{\phi_i\left(\mathbf{b}',\mathbf{g}'\right)}{\phi_i\left(\mathbf{b},\mathbf{g}\right)}=\kappa.
\end{aligned}\tag{A.13}$$

1(a) is equivalent to the following: there is j such that $g_j/g_j'=\kappa$ and $\phi_i(\mathbf{b}',\mathbf{g}')/\phi_i(\mathbf{b},\mathbf{g})=\kappa$, $\forall i$; meanwhile, 1(b) is equivalent to the following: there is j such that $g_j'/g_j=\kappa$ and $\phi_i(\mathbf{b},\mathbf{g})/\phi_i(\mathbf{b}',\mathbf{g}')=\kappa$, $\forall i$. Here again, 1(a) implies 2(b) and 1(b) implies 2(a), and it cannot be that 2(a) or 2(b) hold, but 1(a) and 1(b) do not. Therefore, it suffices to keep condition 1 again. But conditions 1(a) and 1(b) of Steps 1 and 2 together imply that either $b_i'/b_i=\kappa$, $\forall i$, and $g_j/g_j'=\kappa$, $\forall j$, or $b_i/b_i'=\kappa$, $\forall i$, and $g_j'/g_j=\kappa$, $\forall j$. In either case, this means that $(\mathbf{b},\mathbf{g})$ and $(\mathbf{b}',\mathbf{g}')$ belong to the same equivalence class, and, in two steps, we already obtained a fixed point with due regard to the equivalence classes. But this fixed point can only be the unique solution of the system of likelihood equations (17), which is guaranteed (up to equivalence) if $(\mathbf{r},\mathbf{c})\in\mathrm{ri}(\mathscr{P}_{m,n})$.

Otherwise, both inequalities in (A.11) cannot hold with equality, but there must be a strict inequality. Consequently,

$$\begin{aligned}
&\rho\left(\left(\phi\left(\mathbf{b},\mathbf{g}\right),\psi\left(\phi\left(\mathbf{b},\mathbf{g}\right),\mathbf{g}\right)\right),\right.\\
&\quad\left.\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\psi\left(\phi\left(\mathbf{b}',\mathbf{g}'\right),\mathbf{g}'\right)\right)\right)<\rho\left(\left(\mathbf{b},\mathbf{g}\right),\right.\\
&\quad\left.\left(\mathbf{b}',\mathbf{g}'\right)\right),
\end{aligned}\tag{A.14}$$

and hence $f=(\phi,\psi)$ is a weak contraction.

Observe that $f((\mathbf{b}^{(t)},\mathbf{g}^{(t)}))=(\mathbf{b}^{(t+1)},\mathbf{g}^{(t+1)})$, and, under the condition $(\mathbf{r},\mathbf{c})\in\mathrm{ri}(\mathscr{P}_{m,n})$, the ML estimate $(\widehat{\mathbf{b}},\widehat{\mathbf{g}})$ is a unique fixed point of f; that is, $f(\widehat{\mathbf{b}},\widehat{\mathbf{g}})=(\widehat{\mathbf{b}},\widehat{\mathbf{g}})$. Therefore, we have

$$\begin{aligned}
&\ln\rho\left(\left(\mathbf{b}^{(t+1)},\mathbf{g}^{(t+1)}\right),\left(\widehat{\mathbf{b}},\widehat{\mathbf{g}}\right)\right)\\
&\quad<\ln\rho\left(\left(\mathbf{b}^{(t)},\mathbf{g}^{(t)}\right),\left(\widehat{\mathbf{b}},\widehat{\mathbf{g}}\right)\right).
\end{aligned}\tag{A.15}$$

This means that $\ln\rho((\mathbf{b}^{(t)},\mathbf{g}^{(t)}),(\widehat{\mathbf{b}},\widehat{\mathbf{g}}))$ is a monotonic decreasing sequence of nonnegative entries, and so it has a limit $c\ge 0$. But this implies that $\lim_{t\to\infty}\ln\rho((\mathbf{b}^{(t)},\mathbf{g}^{(t)}),(\mathbf{b}^{*},\mathbf{g}^{*}))=0$, where $(\mathbf{b}^{*},\mathbf{g}^{*})$ is in the equivalence class of $(\widehat{\mathbf{b}},\widehat{\mathbf{g}})$, with scaling constant $\kappa=e^{c}$.

On the contrary, when $(\mathbf{r},\mathbf{c})\notin\mathrm{ri}(\mathscr{P}_{m,n})$, the sequence cannot converge to a fixed point, since then it was the solution of the maximum likelihood equations (17). But we have seen that no finite solution can exist in this case. It means that at least one coordinate of the sequence $\{(\mathbf{b}^{(t)},\mathbf{g}^{(t)})\}$ tends to infinity. We remark that, even in this case, we obtain convergence in the other coordinates; this issue was discussed in Section 4.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Gábor Tusnády and Róbert Pálovics for fruitful discussions and making the music recommendation

data available; further, Despina Stasi for suggesting to the authors the fraternity data to be processed. Ahmed Elbanna's research was partly done under the auspices of the MTA-BME Stochastic Research Group.

References

[1] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, Article ID 066111, 2004.

[2] M. E. J. Newman, *Networks, An Introduction*, Oxford University Press, Oxford, UK, 2010.

[3] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[4] M. Bolla, "Penalized versions of the Newman-Girvan modularity and their relation to normalized cuts and k-means clustering," *Physical Review E*, vol. 84, no. 1, Article ID 016108, 2011.

[5] M. Bolla, *Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables*, Wiley, 2013.

[6] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, 1973.

[7] G. J. McLachlan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, NY, USA, 1997.

[8] S. Chatterjee, P. Diaconis, and A. Sly, "Random graphs with a given degree sequence," *Annals of Applied Probability*, vol. 21, no. 4, pp. 1400–1435, 2011.

[9] V. Csiszár, P. Hussami, J. Komlós, T. F. Móri, L. Rejtő, and G. Tusnády, "When the degree sequence is a sufficient statistic," *Acta Mathematica Hungarica*, vol. 134, no. 1-2, pp. 45–53, 2012.

[10] A. Rinaldo, S. Petrović, and S. E. Fienberg, "Maximum lilkelihood estimation in the β-model," *Annals of Statistics*, vol. 41, no. 3, pp. 1085–1110, 2013.

[11] P. W. Holland and S. Leinhardt, "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 33–50, 1981.

[12] P. J. Bickel and A. Chen, "A nonparametric view of network models and Newman-Girvan and other modularities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 50, pp. 21068–21073, 2009.

[13] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[14] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: first steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[15] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, Article ID 016107, 2011.

[16] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, "Stochastic blockmodels with a growing number of classes," *Biometrika*, vol. 99, no. 2, pp. 273–284, 2012.

[17] D. E. Fishkind, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe, "Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 23–39, 2013.

[18] B. Bollobás, S. Janson, and O. Riordan, "The phase transition in inhomogeneous random graphs," *Random Structures & Algorithms*, vol. 31, no. 1, pp. 3–122, 2007.

[19] P. Erdős and A. Renyi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.

[20] E. Szemeredi, "Regular partitions of graphs," in *Colloque Inter*, J.-C. Bermond, J.-C. Fournier, M. Las Vergnas, and D. Sotteau, Eds., CNRS. No. 260, Problémes Combinatoires et Théorie Graphes, pp. 399–401, 1976.

[21] M. Bolla, "Relating multiway discrepancy and singular values of nonnegative rectangular matrices," *Discrete Applied Mathematics*, 2015.

[22] G. Rasch, *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen and Lydiche, Oxford, UK, 1960.

[23] G. Rasch, "On general laws and the meaning of measurement in psychology," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 321–333, University of California Press, July 1961.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[25] L. H. Ungar and D. P. Foster, "A formal statistical approach to collaborativefiltering," in *Proceedings of the Conference on Automatical Learning and Discovery (CONALD '98)*, 1998.

[26] T. Hofmann and J. Puzicha, "Latent class models for collaborative filtering," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)*, T. Dean, Ed., vol. 2, pp. 688–693, Morgan Kaufmann Publications, Stockholm, Sweden, July-August 1999.

[27] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[29] S. L. Lauritzen, *Extremal Families and Systems of Sufficient Statistics*, vol. 49 of *Lecture Notes in Statistics*, Springer, Berlin, Germany, 1988.

[30] P. Erdős and T. Gallai, "Graphs with given degree of vertices," *Matematikai Lapok*, vol. 11, pp. 264–274, 1960 (Hungarian).

[31] N. V. R. Mahadev and U. N. Peled, *Threshold Graphs and Related Topics*, vol. 56 of *Annals of Discrete Mathematics*, North-Holland, Amsterdam, The Netherlands, 1995.

[32] L. P. Stanley, "A zonotope associated with graphical degree sequences," in *Applied Geometry and Discrete Mathematics*, vol. 4 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 555–570, American Mathematical Society, Providence, RI, USA, 1991.

[33] P. L. Hammer, U. N. Peled, and X. Sun, "Difference graphs," *Discrete Applied Mathematics*, vol. 28, no. 1, pp. 35–44, 1990.

[34] S. J. Haberman, "Log-linear models and frequency tables with small expected cell counts," *The Annals of Statistics*, vol. 5, no. 6, pp. 1148–1169, 1977.

[35] A. Barvinok, "What does a random contingency table look like?" *Combinatorics, Probability and Computing*, vol. 19, no. 4, pp. 517–539, 2010.

[36] A. Barvinok, "Matrices with prescribed row and column sums," *Linear Algebra and its Applications*, vol. 436, no. 4, pp. 820–844, 2012.

[37] D. Gale, "A theorem on flows in networks," *Pacific Journal of Mathematics*, vol. 7, pp. 1073–1082, 1957.

[38] H. J. Ryser, "Combinatorial properties of matrices of zeros and ones," *Canadian Journal of Mathematics*, vol. 9, pp. 371–377, 1957.

[39] J. Ford and D. R. Fulkerson, "Maximal flow through a network," *Canadian Journal of Mathematics*, vol. 8, pp. 399–404, 1956.

[40] A. Barvinok and J. A. Hartigan, "An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums," *Transactions of the American Mathematical Society*, vol. 364, pp. 4323–4368, 2012.

[41] A. Barvinok, "On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries," *Advances in Mathematics*, vol. 224, no. 1, pp. 316–339, 2010.

[42] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, NY, USA, 2001.

[43] V. Csiszar, P. Hussami, J. Komlos, T. F. Mori, L. Rejtő, and G. Tusnady, "Testing goodness of fit of random graph models," *Algorithms*, vol. 5, no. 4, pp. 629–635, 2012.

[44] L. Négyessy, T. Nepusz, L. Zalányi, and F. Bazsó, "Convergence and divergence are mostly reciprocated properties of the connections in the network of cortical areas," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1649, pp. 2403–2410, 2008.

[45] D. Yan, A. Chen, and M. I. Jordan, "Cluster forests," *Computational Statistics & Data Analysis*, vol. 66, pp. 178–192, 2013.

[46] H. R. Bernard, P. D. Killworth, and L. Sailer, "Informant accuracy in social-network data V. An experimental attempt to predict actual communication from recall data," *Social Science Research*, vol. 11, no. 1, pp. 30–66, 1982.

[47] R. Pálovics, A. Benczúr, L. Kocsis, T. Kiss, and E. Frigó, "Exploiting temporal influence in online recommendation," in *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*, pp. 273–280, ACM, Foster City, Calif, USA, October 2014.

[48] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[49] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167–173, 2011.