Proceedings of the
1st International Conference and Exhibition on Future RFID Technologies
Eszterhazy Karoly University of Applied Sciences and
Bay Zoltán Nonprofit Ltd. for Applied Research
Eger, Hungary, November 5–7, 2014. pp. 15–22

doi: 10.17048/FutureRFID.1.2014.15

Extending the Rash model to a multiclass parametric network model*

Marianna Bolla[a], **Ahmed Elbanna**[ab]

[a]Institute of Mathematics, Budapest University of Technology and Economics
[b]MTA–BME Stochastics Research Group
marib@math.bme.hu, ahmed@math.bme.hu

Abstract

The newly introduced α-β-models and the classical Rash model are united in a semiparametric multiclass graph model. We give a classification of the nodes of an observed network so that the generated subgraphs and bipartite graphs of it obey these models, where their strongly connected parameters give multiscale evaluation of the nodes. This is a heterogeneous version of the stochastic block model, built via mixtures of loglinear models, the parameters of which are estimated by collaborative filtering. In the context of social networks, the clusters can be identified with social groups and the parameters with attitudes of people of one group towards people of the other, which attitudes depend on the cluster memberships. The algorithm is applied to real-word networks.

Keywords: Rasch model, multiclass loglinear models, collaborative filtering

MSC: 62F86, 05C80

1. Introduction

Recently, α-β-models [1, 2] were developed as the unique graph models where the degree sequence is a sufficient statistic. 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 and their membership 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

*Research supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project.

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, the rectangular analogue of the α-β models.

We propose a heterogeneous block model by carrying on the Rasch model developed more than 50 years ago for evaluating psychological tests [5]. Given the number of clusters and a classification of the vertices, we will use the Rasch model for the bipartite subgraphs, whereas 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 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, the convergence of which is proved in exponential families under very general conditions [4]. This special type of the EM algorithm developed for mixtures is often called collaborative filtering [7].

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, which 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 can be obtained by spectral clustering tools. Together with the description of the algorithm and a theorem about the rank of the matrix of logits, the algorithm is applied to real-word networks.

2. The submodels used

Together with the Rasch model, loglinear type models give the foundation of our unweighted graph and bipartite graph models, the building blocks of our EM iteration.

2.1. α-β models for undirected random graphs

With different parameterization, [1] and [2] 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–Rényi model. 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 $\frac{p_{ij}}{1-p_{ij}}$ ratios, the so-called odds:

$$\frac{p_{ij}}{1-p_{ij}} = \alpha_i \alpha_j \quad (1 \le i < j \le n), \tag{2.1}$$

where the parameters $\alpha_1, \dots, \alpha_n$ are positive reals. This model is called α model in [2]. With the parameter transformation $\beta_i = \ln \alpha_i$ $(i = 1, \dots n)$, it is equivalent to the β model of [1] which applies to the log-odds:

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \beta_j \quad (1 \le i < j \le n) \tag{2.2}$$

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

We are looking for the ML estimate of the parameter vector $\underline{\alpha} = (\alpha_1, \dots, \alpha_n)$ or $\underline{\beta} = (\beta_1, \dots, \beta_n)$ 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 $\mathbf{D} = (D_1, \dots, D_n)$ denote the degree-vector of the above random graph, where $D_i = \sum_{j=1}^{n} A_{ij}$ $(i = 1, \dots n)$. 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 $\underline{\beta}$, see [2]. Let (a_{ij}) be the matrix of the sample realizations (the adjacency entries of the observed graph), $d_i = \sum_{j=1}^{n} a_{ij}$ be the actual degree of vertex i $(i = 1, \dots, n)$ and $\mathbf{d} = (d_1, \dots, d_n)$ be the observed degree-vector. The maximum likelihood estimate $\underline{\hat{\alpha}}$ (or equivalently, $\underline{\hat{\beta}}$) 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, $\underline{\hat{\alpha}}$ is the solution of the following system of maximum likelihood equations:

$$d_i = \sum_{j \ne i}^{n} \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j} \quad (i = 1, \dots, n). \tag{2.3}$$

The Erdős–Gallai conditions characterize so-called graphic degree sequences that can be realized as degree sequences of a graph. For given n, the convex hull of all possible graphic degree sequences is a polytope, to be denoted by $\mathcal{D}_n$. Its extreme points are the so-called threshold graphs. The authors of [1, 2] prove that whenever the observed degree sequence is in the interior of $\mathcal{D}_n$, the maximum likelihood equation (2.3) has a unique solution. On the contrary, when the observed degree vector is a boundary point of $\mathcal{D}_n$, 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 the β_i's is not finite.

The authors in [2] recommend the following algorithm and prove that, provided $\mathbf{d}$ is an interior point of $\mathcal{D}_n$, the iteration of it converges to the unique solution of the system (2.3). Starting with initial parameter values $\alpha_1^{(0)}, \dots, \alpha_n^{(0)}$ and using the observed degree sequence $d_1, \dots, d_n$, which is an inner point of $\mathcal{D}_n$, the iteration is as follows:

$$\alpha_i^{(t)} = \frac{d_i}{\sum_{j \ne i} \frac{1}{\frac{1}{\alpha_j^{(t-1)}} + \alpha_i^{(t-1)}}} \quad (i = 1, \dots, n) \tag{2.4}$$

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

2.2. β-γ model for bipartite graphs

This bipartite graph model traces back to Rasch [5], who investigated binary tables. Given an $m \times n$ random binary array $\mathbf{A} = (A_{ij})$, or equivalently, a bipartite graph, and using the notation $p_{ij} = \mathbb{P}(A_{ij} = 1)$, our model is

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \gamma_j \quad (i = 1, \dots, m, \, j = 1, \dots, n) \tag{2.5}$$

with real parameters $\beta_1, \dots, \beta_m$ and $\gamma_1, \dots, \gamma_n$. In terms of the transformed parameters $b_i = e^{\beta_i}$ and $g_j = e^{\gamma_j}$, it is equivalent to

$$\frac{p_{ij}}{1 - p_{ij}} = b_i g_j \quad (i = 1, \dots, m, \, j = 1, \dots, n) \tag{2.6}$$

where $b_1, \dots, b_m$ and $g_1, \dots, g_n$ are positive reals. Observe that these parameters 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 $\mathbf{b} = (b_1, \dots, b_m)$ and $\mathbf{g} = (g_1, \dots, g_n)$. Based on an observed binary table (a_{ij}), since we are in exponential family, and $\beta_1, \dots, \beta_m, \gamma_1, \dots, \gamma_n$ are natural parameters, the likelihood equation is obtained by making the expectation of the sufficient statistic equal to its sample value. Therefore, with the notation $r_i = \sum_{j=1}^{n} a_{ij}$ $(i = 1, \dots, m)$ and $c_j = \sum_{i=1}^{m} a_{ij}$ $(j = 1, \dots, n)$, the following system of likelihood equations is yielded:

$$\begin{aligned} r_i &= \sum_{j=1}^{n} \frac{b_i g_j}{1 + b_i g_j} = b_i \sum_{j=1}^{n} \frac{1}{\frac{1}{g_j} + b_i}, \quad i = 1, \dots m; \\ c_j &= \sum_{i=1}^{m} \frac{b_i g_j}{1 + b_i g_j} = g_j \sum_{i=1}^{m} \frac{1}{\frac{1}{b_i} + g_j}, \quad j = 1, \dots n. \end{aligned} \tag{2.7}$$

Note that for any sample realization of $\mathbf{A}$, $\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$ holds automatically. Therefore, there is a dependence between the equations of the system (2.7), indicating that the solution is not unique, in accord with our previous remark about the arbitrary scaling factor.

Like the graphic sequences, here we define so-called bipartite realizable sequences, the convex hull of which is the polytope $\mathcal{P}_{m,n}$. In [6] it is proved that the maximum likelihood estimate of the parameters of model (2.6) exists if and only if the observed row- and column-sum sequences are in the relative interior of $\mathcal{P}_{m,n}$. Under these conditions, we define an algorithm that converges to the unique (up to the scaling factor) solution of the maximum likelihood equation (2.7). Starting with positive parameter values $b_i^{(0)}$ $(i = 1, \dots, m)$ and $g_j^{(0)}$ $(j = 1, \dots, n)$ and using the observed row- and column-sums, the iteration is as follows:

$$b_i^{(t)} = \frac{r_i}{\sum_{j=1}^{n} \frac{1}{\frac{1}{g_j^{(t-1)}} + b_i^{(t-1)}}}, \quad i = 1, \dots m$$

$$g_j^{(t)} = \frac{c_j}{\sum_{i=1}^{m} \frac{1}{\frac{1}{b_i^{(t)}} + g_j^{(t-1)}}}, \quad j = 1, \dots n$$

for $t = 1, 2, \dots,$ until convergence. Convergence facts are obtained by the weak contraction property of the transformations yielding the sequence of the iteration.

3. The multipartite graph model

In the several clusters case, the above discussed submodels are the building blocks of a heterogeneous block model. Here the degree sequences are not any more sufficient for the whole graph, only for the building blocks of the subgraphs.

Given $1 \le k \le n$, we are looking for k-partition, in other words, clusters $C_1, \dots, C_k$ of the vertices such that different vertices are independently assigned to the clusters and, given the cluster memberships, vertices $i \in C_u$ and $j \in C_v$ are connected independently, with probability p_{ij} such that

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_{iv} + \beta_{ju}, \tag{3.1}$$

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

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

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

The parameters are collected in the $n \times k$ matrix $\mathbf{B}$ of b_{iv}'s for $i \in C_u$ $u, v = 1, \dots, k$, and are estimated via the EM algorithm for mixtures (collaborative filtering). Here $\mathbf{A} = (a_{ij})$ is the incomplete data specification as the cluster memberships are missing. First we complete our data matrix $\mathbf{A}$ with latent membership vectors $\mathbf{m}_1, \dots, \mathbf{m}_n$ of the vertices that are k-dimensional i.i.d. multinomially distributed random vectors. More precisely, $\mathbf{m}_i = (m_{i1}, \dots, m_{ik})$, where $m_{iu} = 1$ if $i \in C_u$ and zero otherwise.

Starting with initial parameter values $\mathbf{B}^{(0)}$ and membership vectors $\mathbf{m}_1^{(0)}, \dots, \mathbf{m}_n^{(0)}$, the t-th step of the iteration is the following $(t = 1, 2, \dots)$.

- **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 $t-1$, via taking conditional expectation (in the possession of binary variables, the Bayes rule is applicable).

- **M**-step: We estimate the parameters in the actual clustering of the vertices. In the within-cluster scenario, we use the parameter estimation of model (2.1), obtaining estimates of b_{iu}'s ($i \in C_u$) in each cluster separately ($u = 1, \dots, k$); here b_{iu} corresponds to α_i and the number of vertices is $|C_u|$. In the between-cluster scenario, we use the bipartite graph model (2.6) in the following way. For $u \neq v$, 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, etc.

By the general theory of the EM algorithm, since we are in exponential family, the iteration will converge. 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. For selecting the initial number of clusters we used spectral clustering tools.

Theorem 3.1. *Let the $n \times n$ symmetric matrix $\mathbf{L}$ contain the log-odds satisfying the model equation (3.1) as its entries. Then rank $\mathbf{L} \leq 2k$.*

Proof. Let $\widetilde{\mathbf{B}}$ denote the $n \times k$ matrix of β_{iv}'s for $i \in C_u$, $u, v = 1, \dots, k$. We define the $n \times n$ matrix $\mathbf{U}$ as follows: $u_{ij} := \beta_{ic_j}$ ($i, j = 1, \dots, n$). Then $\mathbf{L} = \mathbf{U} + \mathbf{U}^T$. This is obvious if we understand the structure of the matrix $\mathbf{U}$; actually, it is the one-sided blow-up of the matrix $\widetilde{\mathbf{B}}$ as the columns j_1 and j_2 of $\mathbf{U}$ contain the same entries whenever $c_{j_1} = c_{j_2}$. Therefore, there are k different types of columns of $\mathbf{U}$, as many as the number of the clusters, and the columns occur with multiplicities $n_v = |C_v|$ ($v = 1, \dots, k$). Consequently, rank $(\mathbf{U})$ = rank $(\mathbf{U}^T) \leq k$, and, by applying rank theorems, rank $\mathbf{L}$ = rank $(\mathbf{U} + \mathbf{U}^T) \leq 2k$ that finishes the proof. □

Remark 3.2. Let $\mathbf{U} = \sum_{l=1}^k s_l \mathbf{x}_l \mathbf{y}_l^T$ be SVD, where $s_1 \geq \cdots \geq s_k$ are the non-zero singular values of $\mathbf{U}$ with unit-norm singular vector pairs $\mathbf{x}_l, \mathbf{y}_l \in \mathbb{R}^n$ ($l = 1, \dots, k$). For brevity, we drop the subscripts, and consider the unit-norm pair $\mathbf{x}, \mathbf{y}$ corresponding to the singular value s. By their definition, they satisfy the equation $\mathbf{U}\mathbf{y} = s\mathbf{x}$, or equivalently,

$$\sum_{j=1}^n u_{ij} y_j = \sum_{v=1}^k \sum_{j \in C_v} u_{ij} y_j = s x_i, \quad i = 1, \dots, n \tag{3.2}$$

in entry-wise form. Because of the vertical block-form of $\mathbf{U}$, $y_{j_1} = y_{j_2}$ whenever $c_{j_1} = c_{j_2}$. Let $\tilde{\mathbf{y}} \in \mathbb{R}^k$ denote the shrunken vector $\mathbf{y}$ such that $y_j = \tilde{y}_v$ whenever $c_j = v$. Hence, Equation (3.2) has the concise form

$$\sum_{v=1}^k n_v \tilde{b}_{iv} \tilde{y}_v = s x_i, \quad i = 1, \dots, n. \tag{3.3}$$

Introducing the diagonal matrices $\mathbf{D} = \operatorname{diag}(n_1, \dots, n_k)$ and $\widetilde{\mathbf{D}} = \frac{1}{n}\mathbf{D}$, Equation (3.3) has the concise form

$$\widetilde{\mathbf{B}}\mathbf{D}\tilde{\mathbf{y}} = s\mathbf{x},$$

or equivalently,

$$(\widetilde{\mathbf{B}}\widetilde{\mathbf{D}}^{1/2})(\mathbf{D}^{1/2}\tilde{\mathbf{y}}) = \frac{s}{\sqrt{n}}\mathbf{x}$$

is the SVD equation of the matrix $\widetilde{\mathbf{B}}\widetilde{\mathbf{D}}^{1/2}$, where $\|\mathbf{D}^{1/2}\tilde{\mathbf{y}}\| = \|\mathbf{y}\| = 1$. Therefore, the non-zero singular values of this matrix are the numbers $\frac{s_1}{\sqrt{n}}, \dots, \frac{s_k}{\sqrt{n}}$, and they can be bounded from below and from above by a constant, independent of n. Consequently, $s_1, \dots, s_k = \Theta(\sqrt{n})$. Note, that denoting by $\ell_1 \geq \cdots \geq \ell_{2k}$ the positive singular values (absolute values of its eigenvalues) of $\mathbf{L}$, by simple norm inequalities, for the spectral norm of $\mathbf{L}$, $\|\mathbf{L}\| = \ell_1 \leq 2s_1$ holds, and by interlacing theorems, $\ell_{k+v} \leq s_v$ $(v = 1, \dots, k)$. Consequently, the eigenvalues of $\mathbf{L}$ are $\mathcal{O}(\sqrt{n})$.

Note that the ij entry of the symmetric matrix $\mathbf{L}$ is $\beta_{ic_j} + \beta_{jc_i}$, but it equals $\ln \frac{p_{ij}}{1-p_{ij}}$ only when $i \neq j$. For $i = j$, $p_{ii} = 0$, and the log-odds are not defined. However, by filling in the diagonal automatically, the rank of $\mathbf{L}$ cannot exceed $2k$, which gives rise to a low-rank approximation of our data in terms of the log-odds.

4. Applications

Figure 1 shows the resulting clusters obtained by applying our algorithm to the B&K fraternity data with $n = 58$ vertices. The data, collected by Bernard and Killworth, are behavioral frequency counts, based on communication frequencies between students of a college fraternity (see Bernard, H. R., Killworth, P. D. and Sayler, L., Social Science Research 11, 1982). 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. Based spectral clustering considerations, 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 Subsections 2.1 and 2.2. Unfortunately, this can occur when our graph is large but not dense enough. In these situations the iteration did not converge for some coordinates β_{iv} $(i \in C_u)$, but they seemed to tend to $+\infty$ or $-\infty$. Equivalently, the corresponding $\mathbf{b}_{iv}$ $(i \in C_u)$ tended to $+\infty$ or 0, yielding the situation that member $i \in C_u$ had $+\infty$ or 0 affinity towards members of C_v.

In Figure 1, we also enumerated the within-cluster parameter values for each cluster separately, and the parameters reflecting attitudes of students of one group towards the others were written in a concise form above the arrows, where 0, $+\infty$, or finite parameter values can occur. The many 0 affinities show that some groups are quite separated, whereas some people in some groups show infinite affinity towards persons of some specific groups.

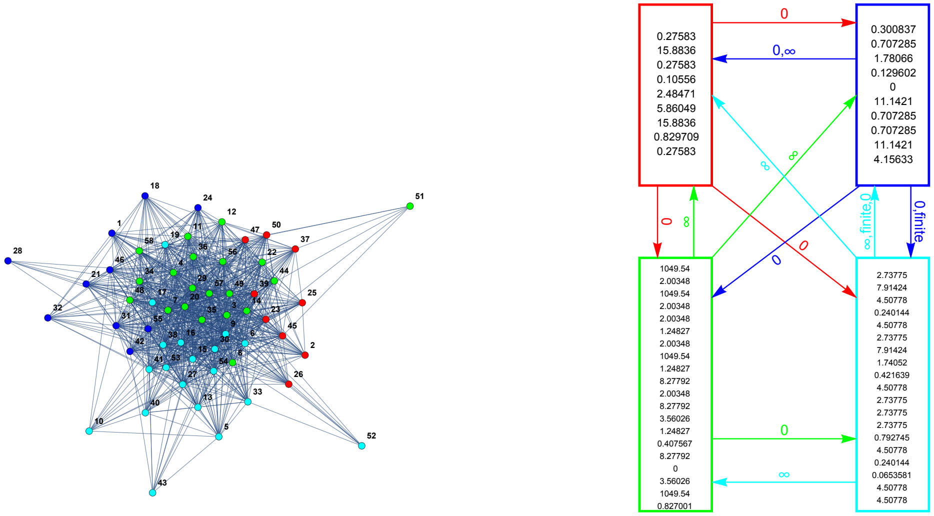

Figure 1: The 4 clusters found by the algorithm and the within- and between-cluster affinities in the B&K fraternity data.

Acknowledgements. The authors thank Gábor Tusnády for fruitful discussions; further, Róbert Pálovics and Despina Statsi for handing us real-world data to be processed.

References

[1] Chatterjee, S., Diaconis, P. and Sly, A., Random graphs with a given degree sequence, *Ann. Stat.*, Vol. 21 (2010), 1400–1435.

[2] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejtő, L. and Tusnády, G., When the degree sequence is a sufficient statistic, *Acta Math. Hung.*, Vol. 134 (2011), 45–53.

[3] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejtő, L. and Tusnády, G., Testing goodness of fit of random graph models, *Algorithms*, Vol. 5 (2012), 629–635.

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B*, Vol. 39 (1977), 1–38.

[5] Rasch, G., On general laws and the meaning of measurement in psychology. In *Proc. of the Fourth Berkeley Symp. on Math. Statist. and Probab.*, University of California Press (1961), 321–333.

[6] Rinaldo, A., Petrovic, S. and Fienberg, S. E., Maximum likelihood estimation in the β-model, *Ann. Statist.*, Vol. 41 (2013), 1085–1110.

[7] Ungar, L. H., Foster, D. P., A Formal Statistical Approach to Collaborative Filtering. In *Proc. Conference on Automatical Learning and Discovery* (1998), 1–6.