Extending the Rash model to a multiclass parametric network model

Marianna Bolla and **Ahmed Elbanna**
Institute of Mathematics
Technical University of Budapest
Bolla: also at the Center for Telecommunication and Informatics, Debrecen University
(Research supported by the TÁMOP-4.2.2.B-10/1-2010-0009 project)
Elbanna: also at the MTA-BME Stochastic Research Group
marib@math.bme.hu, ahmed@math.bme.hu

1st International Conference on Future RFID Technologies, Eger

November 6, 2014

Outline

- We will amalgamate the Rash model and the newly introduced α-β-models in the framework of a semiparametric probabilistic graph model.
- Our algorithm gives a partition of the vertices of an observed graph so that the generated subgraphs and bipartite graphs obey these models, where their strongly connected parameters give multiscale evaluation of the vertices at the same time.
- We build a heterogeneous version of the stochastic block model via mixtures of loglinear models and the parameters are estimated with a special EM iteration.
- 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 randomly generated and real-word data.

References

- Chatterjee,S., Diaconis,P. and Sly, A., Random graphs with a given degree sequence, *Ann. Statist.* (2010.)
- 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.* (2011).
- Dempster,A.P., Laird,N.M. and Rubin,D.B., Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B* (1977).
- Lauritzen,S.L., Extremal families and systems of sufficient statistics. Lecture Notes in Statistics, Springer (1988).
- Rasch,G., On general laws and the meaning of measurement in psychology. In *Proc. of 4h Berkeley Symp. on Math. Statist. and Probab.* (1961).
- Rinaldo,A., Petrovic,S. and Fienberg,S.E., Maximum likelihood estimation in the β-model, *Ann. Statist.* (2013).

α-β models for undirected random graphs

$\mathbf{A} = (A_{ij})$: $n \times n$ symmetric adjacency with zero diagonal.
$p_{ij} = \mathbb{P}(A_{ij} = 1)$ independently for all $i < j$ pairs.
α model:

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

with positive real parameters $\alpha_1, \dots, \alpha_n$.
β model: $\beta_i = \ln \alpha_i$ $(i = 1, \dots n)$.

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

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

$$p_{ij} = \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j}, \qquad 1 - p_{ij} = \frac{1}{1 + \alpha_i \alpha_j}.$$

ML estimation of the parameters

Find 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 (a_{ij}) 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.)

The degree sequence $\mathbf{D} = (D_1, \dots, D_n)$ is a sufficient statistic.

The ML estimate $\hat{\underline{\alpha}}$ (or equivalently, $\hat{\underline{\beta}}$) is derived from the fact that the observed degree $d_i = \sum_{j=1}^{n} a_{ij}$ is equal to the expected $\mathbb{E}(D_i) = \sum_{i=1}^{n} p_{ij}$. Therefore, the ML equation:

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

Graphic degree sequences

The sequence $d_1, \dots, d_n$ of nonnegative integers is called graphic if there is an unweighted, undirected graph on n vertices such that its vertex-degrees are the numbers $d_1, \dots, d_n$ in some order.
By the Erdős–Gallai theorem, the sequence $d_1 \geq \dots \geq d_n \geq 0$ of integers is graphic if and only if $\sum_{i=1}^{n} d_i$ is even and

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

For given n, the convex hull of all possible graphic degree sequences is a convex polytope $\mathcal{D}_n$. Its extreme points are the threshold graphs. For $n = 3$ all undirected graphs are threshold, therefore assume that $n > 3$. The number of vertices of $\mathcal{D}_n$ superexponentially grows with n, therefore the problem of characterizing threshold graphs has a high computational complexity.

Existence and uniqueness of the ML estimete

Chatterjee et al. and V. Csiszár et al. proved that $\mathcal{D}_n$ is the topological closure of the set of expected degree sequences, and for given $n > 3$, if $\mathbf{d} \in \mathrm{int}\,(\mathcal{D}_n)$ is an interior point, then the ML equation has a unique solution.

The converse is also true: Rinaldo et al. proved that the ML estimate exists if and only if the observed degree vector is an inner point of $\mathcal{D}_n$.

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. 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 $-\infty$. Still, the other coordinates have a unique ML estimate.

Algorithm to estimate the α parameters

V. Csiszár et al. proved that, provided $\mathbf{d} \in \operatorname{int}(\mathcal{D}_n)$, the following iteration converges to the unique solution of the ML equation. 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 $\mathcal{D}_n$:

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

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

The origins: Rasch model

Our bipartite graph model traces back to Lauritzen,Rasch, etc.
Rasch model: the entries of an $m \times n$ binary table $\mathbf{A} = (A_{ij})$ are independent Bernoulli random variables with parameters $p_{ij} = \mathbb{P}(A_{ij} = 1)$ satisfying

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

with real parameters $\beta_1, \dots, \beta_m$ and $\delta_1, \dots, \delta_n$.
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. Rasch also gave a description of the parameters: β_i was the ability of person i, while δ_j the difficulty of test item j. Therefore, the more intelligent the person and the less difficult the test, the larger the success/failure ratio was on a logarithmic scale.

β-γ model for bipartite graphs

Given an $m \times n$ random binary table $\mathbf{A} = (A_{ij})$, with $p_{ij} = \mathbb{P}(A_{ij} = 1)$, the β-γ model:

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

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 $g_j = e^{\gamma_j}$:

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

where $b_1, \dots, b_m$ and $g_1, \dots, g_n$ are positive reals.
Conversely, the probabilities can be expressed in terms of the parameters:

$$p_{ij} = \frac{b_i g_j}{1 + b_i g_j}, \qquad 1 - p_{ij} = \frac{1}{1 + b_i g_j}.$$

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

$$b_i' = b_i \kappa, \qquad g_j' = \frac{g_j}{\kappa}$$

with some $\kappa > 0$. Therefore, the parameters b_i $(i = 1, \dots, m)$ and g_j $(j = 1, \dots, n)$ are arbitrary to within a multiplicative constant.

Sufficient statistics

By the Neyman–Fisher factorization theorem, 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 for the parameters $\mathbf{b} = (b_1, \ldots, b_m)$ and $\mathbf{g} = (g_1, \ldots, g_n)$:

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

0-1 matrices with fixed margins

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 **A** – is constantly 1, indicating that the conditional joint distribution of the entries, given the row- and column-sums, is uniform in this model.

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.

System of ML equations

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 equal the expectation of the sufficient statistic 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 ML equations:

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

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

Because of $\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$, there is a dependence between the ML equations, indicating that the solution is not unique, in accord with our previous remark about the arbitrary scaling factor $\kappa > 0$.

We will prove that apart from this scaling, the solution is unique if it exists at all. So that to avoid this indeterminacy, we may impose conditions on the parameters, for example, $\sum_{i=1}^{m} \beta_i + \sum_{j=1}^{n} \gamma_j = 0$. Conditions 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 entries:

$$\sum_{i=1}^{k} r_i \leq \sum_{j=1}^{n} \min\{c_j, k\}, \quad k = 1, \dots, m;$$

$$\sum_{j=1}^{k} c_j \leq \sum_{i=1}^{m} \min\{r_i, k\}, \quad k = 1, \dots, n.$$

These conditions define bipartite realizable sequences and form a polytope in $\mathbb{R}^{m+n}$; more precisely, in an $(m + n - 1)$-dimensional hyperplane of it. It is called polytope of bipartite degree sequences and denoted by $\mathcal{P}_{m,n}$.

When the ML estimate exists

Analogously to the $\alpha - \beta$ models, $\mathcal{P}_{m,n}$ is the closure of the set of the expected row- and column-sum sequences in the above model. Hammer 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 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 $-\infty$.
Rinaldo et al.: the ML estimate of the model parameters exists if and only if the observed row- and column-sum sequence $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}\,(\mathcal{P}_{m,n})$, the relative interior of $\mathcal{P}_{m,n}$.
In this case for the probabilities, calculated through the estimated finite parameter values $\hat{b}_i$'s and $\hat{g}_j$'s, $0 < p_{ij} < 1$ holds $\forall i, j$.

Algorithm

Under these conditions, we defined an algorithm that converges to the unique (up to the above equivalence) solution of the ML equation. More precisely, we proved that if $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}\,(\mathcal{P}_{m,n})$, then our 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 $b_i^{(0)}$ $(i = 1, \dots, m)$ and $g_j^{(0)}$ $(j = 1, \dots, n)$ and using the observed row- and column-sums, the iteration:

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

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

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

The multipartite graph model

The above α-β and β-γ models will be the building blocks of a heterogeneous block model. Here the degree sequences are not any more sufficient for the whole graph, only for 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 a cluster C_u with probability π_u $(u = 1, \dots, k)$, where $\sum_{u=1}^{k} \pi_u = 1$;
- given the cluster memberships, vertices $i \in C_u$ and $j \in C_v$ are connected independently, with probability p_{ij} such that

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

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

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

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

EM algorithm

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

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

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 Dempster et al., reminiscent of collaborative filtering.
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. $Multy(1, \underline{\pi})$ random vectors. More precisely, $\mathbf{m}_i = (m_{i1}, \dots, m_{ik})$, 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 $\mathbb{P}(m_{iu} = 1) = \pi_u$.

EM iteration

Starting with initial parameter values $\underline{\pi}^{(0)}$, $\mathbf{G}^{(0)}$ 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: Calculate the conditional expectation of each $\mathbf{m}_i$ conditioned on the model parameters and on the other cluster assignments $M^{(t-1)}$, obtained in step $t-1$.

The responsibility of vertex i for cluster u in the t-th step:

$$\pi_{iu}^{(t)} = \mathbb{E}(m_{iu} \mid M^{(t-1)})$$

and by the Bayes theorem, it is

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

$(u = 1, \dots, k;\ i = 1, \dots, n)$.

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 the complete likelihood effecting vertex i under the condition $m_{iu}=1$. Therefore,

$$\mathbb{P}(M^{(t-1)}|m_{iu}=1)$$
$$=\prod_{v=1}^{k}\prod_{j\in C_v, j\sim i} p_{ij}^{(t-1)} \prod_{j\in C_v, j\nsim i}(1-p_{ij}^{(t-1)})$$
$$=\prod_{v=1}^{k}\left\{p_{ij}^{(t-1)}\right\}^{e_{vi}}\left\{(1-p_{ij})^{(t-1)}\right\}^{|C_v|\cdot(|C_v|-1)/2-e_{vi}},$$

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

$$p_{ij}^{(t-1)}=\frac{g_{ic_j}^{(t-1)}g_{jc_i}^{(t-1)}}{1+g_{ic_j}^{(t-1)}g_{jc_i}^{(t-1)}}.$$

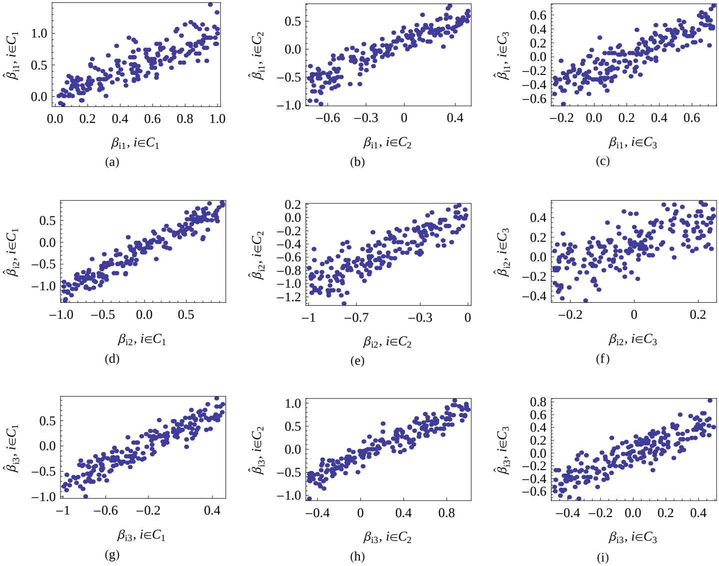
M-step: We update $\underline{\pi}^{(t)}$ and $\mathbf{m}^{(t)}$: $\pi_u^{(t)} := \frac{1}{n}\sum_{i=1}^{n} \pi_{iu}^{(t)}$ and $m_{iu}^{(t)} = 1$ if $\pi_{iu}^{(t)} = \max_v \pi_{iv}^{(t)}$ and 0, otherwise. (in case of ambiguity, we select the smallest index for the cluster membership of vertex i). This gives a new clustering of the vertices.
Then we estimate the parameters in the actual clustering of the vertices. In the within-cluster scenario, we use the parameter estimation of the $\alpha - \beta$ model, obtaining estimates of g_{iu}'s ($i \in C_u$) in each cluster separately ($u = 1, \dots, k$); as for cluster u, g_{iu} corresponds to α_i and the number of vertices is $|C_u|$. In the between-cluster scenario, we use the bipartite graph model 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 g_{iv} ($i \in C_u$) and g_{ju} ($j \in C_v$) are estimated with the above algorithm; here g_{iv}'s correspond to b_i's, g_{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 $\mathbf{G}^{(t)}$, 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.

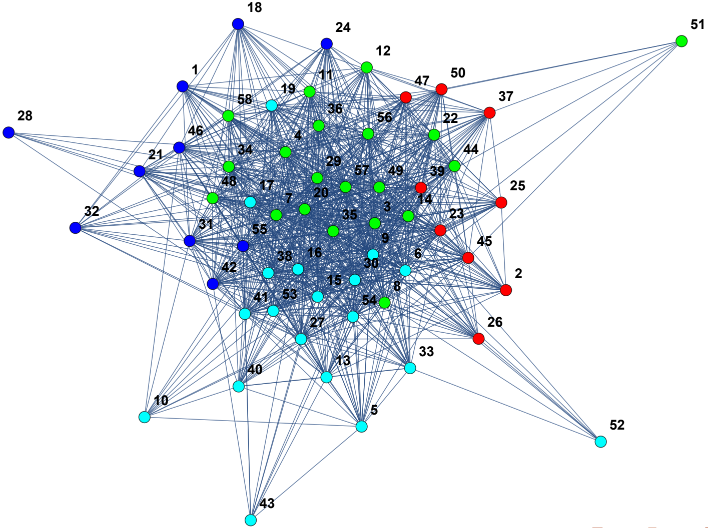
The parameter β_{iv} with $c_i = u$ embodies the affinity of vertex i of cluster C_u towards vertices of cluster C_v; and likewise, β_{ju} with $c_j = v$ embodies the affinity of vertex j of cluster C_v towards vertices of cluster C_u. By the model, this affinities are added together on the level of the log-odds.

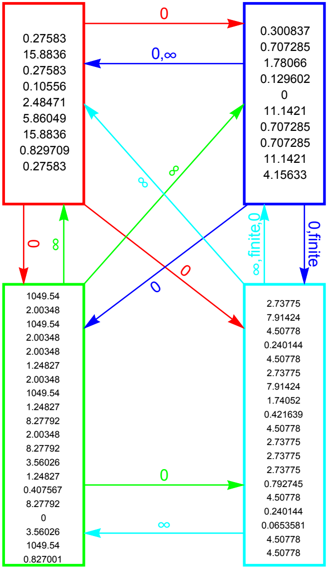
This so-called $k - \beta$ model, introduced in V. Csiszár et al., is applicable to social networks, where attitudes of individuals in the same social group (say, u) are the same toward members of another social group (say, v), though, this attitude also depends on the individual in group u. The model may also be applied to biological networks, where the clusters consist, for example, of different functioning synopses or other units of the brain.

Application to generated data



Application to the B&K (Bernard and Killworth) fraternity data




0.27583
15.8836
0.27583
0.10556
2.48471
5.86049
15.8836
0.829709
0.27583
0.300837
0.707285
1.78066
0.129602
0
11.1421
0.707285
0.707285
11.1421
4.15633
1049.54
2.00348
1049.54
2.00348
2.00348
1.24827
2.00348
1049.54
1.24827
8.27792
2.00348
8.27792
3.56026
1.24827
0.407567
8.27792
0
3.56026
1049.54
0.827001
2.73775
7.91424
4.50778
0.240144
4.50778
2.73775
7.91424
1.74052
0.421639
4.50778
2.73775
2.73775
2.73775
0.792745
4.50778
0.240144
0.0653581
4.50778
4.50778
0
0,∞
∞
∞
0
∞
0
0
∞,finite,0
0,finite
0
∞

Thank you for your attention