Spectral Clustering and Biclustering

Marianna Bolla

Institute of Mathematics
Budapest University of Technology and Economics
marib@math.bme.hu

Vienna, April 16, 2012
To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.

To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.

To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.
Motivation

- To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.
- To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.
- To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.
Motivation

- To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.

- To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.

- To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.
Motivation

Spectral clustering of graphs

Noisy random graphs

Biclustering of contingency tables

Spectral clustering of edge-weighted graphs

\[ G = (V, \mathbf{W}) \] edge-weighted graph, \(|V| = n\), \(\mathbf{W}\): weight matrix of edges
\[ w_{ij} = w_{ji} \geq 0 \quad (i \neq j) \text{ and } w_{ii} = 0 \quad (i=1, \ldots, n). \]

\[ d_i := \sum_{j=1}^{n} w_{ij} \quad (i = 1, \ldots, n) \] generalized degrees
\[ \mathbf{d} := (d_1, \ldots, d_n)^T \] degree vector, \( \sqrt{\mathbf{d}} := (\sqrt{d_1}, \ldots, \sqrt{d_n})^T \)

\[ \mathbf{D} := \text{diag}(d_1, \ldots, d_n) \] degree matrix

w.l.g. \( \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1 \) will be supposed
Laplacian and modularity matrices

\[ L = D - W: \text{ Laplacian} \]
\[ L_D = I - D^{-1/2}WD^{-1/2}: \text{ normalized Laplacian} \]
\[ \text{Spec} (L_D) \in [0, 2] \]
If \( G \) is connected (\( W \) is irreducible), then 0 is a single eigenvalue with corresponding unit-norm eigenvector \( \sqrt{d} \).

\[ M_D = D^{-1/2}WD^{-1/2} - \sqrt{d}\sqrt{d}^T: \text{ normalized modularity matrix} \]
\[ B, \text{ Phys. Rev. E (2011)} \]
\[ \text{Spec} (M_D) \in [-1, 1] \]
1 cannot be an eigenvalue if \( G \) is connected, and 0 is always an eigenvalue with eigenvector \( \sqrt{d} \).

The spectral gap of \( G \): \( 1 - \|M_D\| \) (spectral norm)
Fact: the spectral decomposition of either $L_D$ or $M_D$ solves the following \textit{quadratic placement problem}.

For a given positive integer $k$ ($1 < k < n$), minimize

$$Q_k(X) = \sum_{i<j} w_{ij} ||r_i - r_j||^2$$

on the conditions

$$\sum_{i=1}^{n} d_i r_i r_i^T = I_{k-1}, \quad \sum_{i=1}^{n} d_i r_i = 0,$$

where the vectors $r_1, \ldots, r_n$ are $(k - 1)$-dimensional representatives of the vertices, which form the row vectors of the $n \times (k - 1)$ matrix $X$. 
Normalized Laplacian eigenvalues

$G$ is connected, $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$
eigenvalues of $L_D$ with corresponding unit-norm, pairwise orthogonal eigenvectors $u_0 = \sqrt{d}, u_1, \ldots, u_{n-1}$.

In B, Tusnády, Discrete Math. (1994): the minimum of $Q_k(X)$ under the constraints for the representatives is

$$
\sum_{i=1}^{k-1} \lambda_i
$$

and is attained by the following representation:

$r_1^*, \ldots, r_n^*$ are row vectors of the matrix

$X^* = (D^{-1/2}u_1, \ldots, D^{-1/2}u_{k-1})$. 

Instead of $X$ the augmented $n \times k$ matrix $\tilde{X}$ can as well be used, which is obtained from $X$ by inserting the column $x_0 = 1$ of all 1’s. In fact, $x_0 = D^{-1/2}u_0 = 1$, where $u_0 = \sqrt{d}$ is the eigenvector belonging to the eigenvalue 0 of $L_D$. Then

$$Q_k(\tilde{X}) = Q_k(X) = \text{tr} \left( D^{1/2} \tilde{X} \right)^T (I_n - D^{-1/2}WD^{-1/2})(D^{1/2} \tilde{X}),$$

and $Q_k(X)$ is minimized on the constraint $\tilde{X}^T D \tilde{X} = I_k$, or equivalently, $D^{1/2} \tilde{X}$ is suborthogonal.
Continuous relaxation of a discrete minimization

This problem is the **continuous relaxation** of minimizing

\[ Q_k(\tilde{X}(P_k)) = \text{tr} \left( D^{1/2}\tilde{X}(P_k) \right)^T (I_n - D^{-1/2}WD^{-1/2})D^{1/2}\tilde{X}(P_k) \]

over the set of \( k \)-partitions \( P_k = (V_1, \ldots, V_k) \) of the vertices such that \( P_k \) is planted into \( \tilde{X} \) in the way that the columns of \( \tilde{X}(P_k) \) are so-called **partition-vectors** belonging to \( P_k \):

the coordinates of the \( i \)th column are zeros, except those indexing vertices of \( V_i \) which are equal to

\[ \frac{1}{\sqrt{\text{Vol}(V_i)}}, \quad i = 1, \ldots, k. \]
**Representation, spectral relaxation**

$Q_k(\tilde{X}(P_k))$ is the normalized cut of $P_k = (V_1, \ldots, V_k)$:

$$Q_k(\tilde{X}(P_k)) = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \left( \frac{1}{\text{Vol}(V_a)} + \frac{1}{\text{Vol}(V_b)} \right) w(V_a, V_b)$$

$$= \sum_{a=1}^{k} \frac{w(V_a, \bar{V}_a)}{\text{Vol}(V_a)} = k - \sum_{a=1}^{k} \frac{w(V_a, V_a)}{\text{Vol}(V_a)}$$

Minimum $k$-way normalized cut of $G = (V, W)$:

$$f_k(G) = \min_{P_k \in \mathcal{P}_k} Q_k(\tilde{X}(P_k)),$$

where $\text{Vol}(U) = \sum_{i \in U} d_i$: volume of $U \subset V$

$w(X, Y) = \sum_{i \in X} \sum_{j \in Y} w_{ij}$: weighted cut between $X, Y \subset V$
Because of the spectral relaxation:

\[ f_k(G) \geq \sum_{i=0}^{k-1} \lambda_i = \sum_{i=1}^{k-1} \lambda_i \]

B, Tusnády, Discrete Math. (1994) general \( k \), called weighted cut
Azran, Ghahramani, Siam J. Comput (2000) general \( k \)
Meila and Shi, NIPS (2001): \( k = 2 \)
Upper estimate: depends on the corresponding eigenvectors.
Point of spectral clustering: optimizing over \( \mathcal{P}_k \) is NP-hard.
Isoperimetric number

**Definition**

The Cheeger constant of the weighted graph $G = (V, W)$ is

$$h(G) = \min_{\substack{U \subset V \ \text{Vol}(U) \leq 1/2}} \frac{w(U, \bar{U})}{\text{Vol}(U)}$$

**Theorem**

(B, M-Sáska, Discrete Math. (2004)). Let $\lambda_1$ be the smallest positive eigenvalue of $L_D$. Then

$$\frac{\lambda_1}{2} \leq h(G) \leq \min\{1, \sqrt{2\lambda_1}\}.$$ 

If $\lambda_1 \leq 1$ (G is not the complete graph), then

$$h(G) \leq \sqrt{\lambda_1(2 - \lambda_1)}.$$
Normalized Newman–Girvan modularity


\[
M_k(W, P_k) = \sum_{a=1}^{k} \frac{1}{\text{Vol}(V_a)} \sum_{i,j \in V_a} (w_{ij} - d_i d_j) = \sum_{a=1}^{k} \frac{w(V_a, V_a)}{\text{Vol}(V_a)} - 1
\]

Since

\[
M_k(W, P_k) = k - 1 - Q_k(\tilde{X}(P_k)),
\]

maximizing the \(k\)-way normalized Newman-Girvan modularity is equivalent to the normalized cut problem and it can be solved by the same spectral relaxation.
Spectral gap and variance

Weighted $k$-variance of the vertex representatives:

$$S_k^2(X) = \min_{P_k=(V_1,\ldots,V_k)} \sum_{a=1}^k \sum_{j \in V_a} d_j \|r_j - c_a\|^2$$

where $c_a = \frac{1}{\text{Vol}(V_a)} \sum_{j \in V_a} d_j r_j$.


**Theorem**

*In the representation $X^* = (D^{-1/2}u_0, D^{-1/2}u_1) = (1, D^{-1/2}u_1)$:*

$$S_2^2(X^*) \leq \frac{\lambda_1}{\lambda_2}$$
$f_2(G)$ is the symmetric version of $h(G)$: $f_2(G) \leq 2h(G) \implies$

$$f_2(G) \leq 2\sqrt{\lambda_1(2 - \lambda_1)}, \quad \lambda_1 \leq 1.$$ 

**Theorem**

Suppose that $G = (V, W)$ is connected, and $\lambda_i$'s are the eigenvalues of $L_D$. Then $\sum_{i=1}^{k-1} \lambda_i \leq f_k(G)$ and in the case when the optimal $k$-dimensional representatives can be classified into $k$ well-separated clusters in such a way that the maximum cluster diameter $\varepsilon$ satisfies the relation

$$\varepsilon \leq \min\left\{ \frac{1}{\sqrt{2k}}, \sqrt{2} \min_i \sqrt{\text{Vol}(V_i)} \right\}$$

with $k$-partition $(V_1, \ldots, V_k)$ induced by the clusters above, then

$$f_k(G) \leq c^2 \sum_{i=1}^{k-1} \lambda_i,$$

where $c = 1 + \varepsilon c'/(\sqrt{2} - \varepsilon c')$ and $c' = 1/ \min_i \sqrt{\text{Vol}(V_i)}$. 
Normalized modularity eigenvalues

\[ M_D = I - L_D - \sqrt{d} \sqrt{d}^T \]

eigenvalues:
\[ 1 - \lambda_1 \geq \ldots \geq \lambda_{n-1} \geq -1 \] with the same eigenvectors and 0 with eigenvector \( \sqrt{d} \).

1 cannot be an eigenvalue if \( G \) is connected / \( W \) is irreducible.

- Large absolute value positive eigenvalues of \( M_D \) are responsible for clusters with high intra- and low inter-cluster densities.
- If we minimize \( M_k(W, P_k) \) instead of maximizing over \( P_k \): small negative eigenvalues of \( M_D \) are responsible for clusters with low intra- and high inter-cluster densities.
- If we take into account eigenvalues from both ends of the normalized modularity spectrum, we can recover so-called regular cluster pairs.
Volume regularity

Lemma

*Expander Mixing Lemma for weighted graphs:* Supposing $\text{Vol}(V) = 1$, for all $X, Y \subset V$,

$$|w(X, Y) - \text{Vol}(X)\text{Vol}(Y)| \leq \|M_D\| \cdot \sqrt{\text{Vol}(X)\text{Vol}(Y)}$$

For simple graphs: Alon, Combinatorica (1986)


For edge-weighted graphs: Chung, Graham, Random structures and algorithms (2008), in context of quasi-random properties.
What if the gap is not at the ends of the spectrum?

We want to partition the vertices into clusters so that a relation formulated in the Lemma (1-cluster case) between the edge-densities and volumes of the cluster pairs would hold. We will use a slightly modified version of the volume regularity’s notion introduced by Alon, Coja-Oghlan, Han, Kang, Rödl, and Schacht, Siam J. Comput. (2010):

**Definition**

Let $G = (V, W)$ be a weighted graph with $\text{Vol}(V) = 1$. The disjoint pair $(A, B)$ is **$\alpha$-volume regular** if for all $X \subset A$, $Y \subset B$ we have

$$|w(X, Y) - \rho(A, B)\text{Vol}(X)\text{Vol}(Y)| \leq \alpha \sqrt{\text{Vol}(A)\text{Vol}(B)}$$

where $\rho(A, B) = \frac{w(A,B)}{\text{Vol}(A)\text{Vol}(B)}$ is the relative inter-cluster density of $(A, B)$. 
Outline

For general deterministic edge-weighted graphs we’ll prove that the existence of $k - 1$ eigenvalues of $M_D$ separated from 0 by $\varepsilon$, is indication of a $k$-cluster structure, while the eigenvalues accumulating around 0 are responsible for the pairwise regularities.

The clusters themselves can be recovered by applying the $k$-means algorithm for the vertex representatives obtained by the eigenvectors corresponding to the structural eigenvalues.

Our theorem bounds the volume regularity’s constants of the different cluster pairs by means of $\varepsilon$ and the $k$-variance of the vertex representatives (based on the structural eigenvectors). Estimates for the intra-cluster densities are also given.
Theorem

\( G = (V, W) \) is edge-weighted graph on \( n \) vertices, \( \text{Vol}(V) = 1 \) and there are no dominant vertices: \( d_i = \Theta(1/n) \), \( i = 1, \ldots, n \) as \( n \to \infty \). The eigenvalues of \( M_D \) in decreasing absolute values are:

\[
1 > |\mu_1| \geq \cdots \geq |\mu_{k-1}| > \varepsilon \geq |\mu_i|, \quad i \geq k.
\]

The partition \((V_1, \ldots, V_k)\) of \( V \) is defined so that it minimizes the weighted \( k \)-variance \( s^2 = S^2_k(X^*) \) of the vertex representatives. Suppose that there is a constant \( 0 < c \leq \frac{1}{k} \) such that \( |V_i| \geq cn \), \( i = 1, \ldots, k \). Then the \((V_i, V_j)\) pairs are \( O(\sqrt{2ks} + \varepsilon) \)-volume regular \( (i \neq j) \) and for the clusters \( V_i \) \( (i = 1, \ldots, k) \) the following holds: for all \( X, Y \subset V_i \),

\[
|w(X, Y) - \rho(V_i)\text{Vol}(X)\text{Vol}(Y)| = O(\sqrt{2ks} + \varepsilon)\text{Vol}(V_i),
\]

where \( \rho(V_i) = \frac{w(V_i, V_i)}{\text{Vol}^2(V_i)} \) is the relative intra-cluster density of \( V_i \).
The **case** $k = 2$ was treated separately in


Under the same conditions and with notations $|\mu_1| = \theta$, $|\mu_2| = \varepsilon$, the $(V_1, V_2)$ pair is $O\left(\sqrt{\frac{1-\theta}{1-\varepsilon}}\right)$-volume regular.
Spectral clustering of data points

$x_1, \ldots, x_n \in \mathbb{R}^d$

Form one of the following graphs on $n$ vertices:

- $\varepsilon$-level graph
- $k$ nearest neighbor graph
- complete graph with Gaussian weights:

$$w_{ij} = K_{\text{Gauss}}(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}},$$

where $\sigma > 0$ is a parameter.

Data are mapped into a Reproducing Kernel Hilbert Space, but we are not interested in the mapped data, just in the new kernel which is the weight matrix of an edge-weighted graph.

Sometimes several positive definite kernels are multiplied together (Schur/Hadamard product) which characterize brightness, color, texture, etc. of pixels.
Random graphs, Wigner-noise

Definition

The $n \times n$ symmetric real matrix $W$ is a Wigner-noise if its entries $w_{ij}$, $1 \leq i \leq j \leq n$, are independent random variables, $\mathbb{E}w_{ij} = 0$, $\text{Var} w_{ij} \leq \sigma^2$ with some $0 < \sigma < \infty$ and the $w_{ij}$’s are uniformly bounded (there is a constant $K > 0$ such that $|w_{ij}| \leq K$).

Füredi, Komlós, Combinatorica (1981):

$$\max_{1 \leq i \leq n} |\lambda_i(W)| \leq 2\sigma \sqrt{n} + O(n^{1/3} \log n)$$

with probability tending to 1 as $n \to \infty$. 
Sharp concentration theorem

**Theorem**

\( W \) is an \( n \times n \) real symmetric matrix, its entries in and above the main diagonal are independent random variables with absolute value at most 1. \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \): eigenvalues of \( W \).

For any \( t > 0 \):

\[
\mathbb{P} (|\lambda_i - \mathbb{E}(\lambda_i)| > t) \leq \exp \left( -\frac{(1 - o(1))t^2}{32i^2} \right) \quad \text{when} \quad i \leq \frac{n}{2},
\]

and the same estimate holds for the probability

\[
\mathbb{P} (|\lambda_{n-i+1} - \mathbb{E}(\lambda_{n-i+1})| > t).
\]

Previous results imply:

**Lemma**

There exist positive constants $C_1$ and $C_2$, depending on the common bound $K$ for the entries of the Wigner-noise $W$, such that

$$
P\left( \|W\| > C_1 \cdot \sqrt{n} \right) \leq \exp(-C_2 \cdot n).$$

Borel–Cantelli Lemma $\implies$

The spectral norm of $W$ is $O(\sqrt{n})$ almost surely.
Perturbation results for weighted graphs

\( \mathbf{A} = \mathbf{B} + \mathbf{W} \), where

\( \mathbf{W} : n \times n \) Wigner-noise

\( \mathbf{B} : n \times n \) blown-up matrix of \( \mathbf{P} \) with blow-up sizes \( n_1, \ldots, n_k \),

\[ \sum_{i=1}^{k} n_i = n. \]

\( \mathbf{P} : k \times k \) pattern matrix

\( k \) is kept fixed as \( n_1, \ldots, n_k \to \infty \) “at the same rate”: there is a constant \( c \) such that

\[ \frac{n_i}{n} \geq c, \quad i = 1, \ldots, k. \]

growth rate condition: g.r.c.
Motivation

Spectral clustering of graphs

Noisy random graphs

Biclustering of contingency tables

Adjacency spectrum of a noisy graph

\[ G_n = (V, A), \ A = B + W \text{ is } n \times n, \ n \to \infty \]

\( B \) induces a planted partition \( P_k = (V_1, \ldots, V_k) \) of \( V \).

Weyl's perturbation theorem \( \implies \)

Adjacency spectrum of \( G_n \): under g.r.c. there are \( k \) structural eigenvalues of order \( n \) (in absolute value) and the others are \( \mathcal{O}(\sqrt{n}) \), almost surely.

The eigenvectors \( X = (x_1, \ldots, x_k) \) corresponding to the structural eigenvalues are “not far” from the subspace of stepwise constant vectors on \( P_k \) \( \implies \)

\[ S_k^2(X) \leq S_k^2(P_k, X) = \mathcal{O}\left(\frac{1}{n}\right), \text{ almost surely as } n \to \infty. \]

This extends over the normalized Laplacian and modularity spectra.
Noisy graph is simple with appropriate noise

The uniform bound $K$ on the entries of $W$ is such that $A = B + W$ has entries in $[0,1]$. With an appropriate Wigner-noise the noisy matrix $A$ is a generalized random graph: edges between $V_a$ and $V_b$ exist with probability $0 < p_{ab} < 1$. For $1 \leq a \leq b \leq k$ and $i \in V_a$, $j \in V_b$:

$$w_{ij} := \begin{cases} 1 - p_{ab}, & \text{with probability } p_{ab} \\ -p_{ab} & \text{with probability } 1 - p_{ab} \end{cases}$$

be independent random variables, otherwise $W$ is symmetric. The entries have zero expectation and bounded variance:

$$\sigma^2 = \max_{1 \leq a \leq b \leq k} p_{ab}(1 - p_{ab}) \leq \frac{1}{4}.$$
Generalized random graphs

Ideal $k$-cluster case: given the partition $(V_1, \ldots, V_k)$ of $V$, vertices $i \in V_a$ and $j \in V_b$ are connected with probability $p_{ab}$, independently of each other, $1 \leq a, b \leq k$.


If $k$ is fixed and $n \to \infty$ such that $\frac{|V_a|}{n} \geq c \ (a = 1, \ldots, k)$ with some $0 < c \leq \frac{1}{k}$, then there exists a positive number $0 < \theta \leq 1$, independent of $n$, such that for every $0 < \tau < 1/2$

- there are exactly $k - 1$ eigenvalues of $M_D$ greater than $\theta - n^{-\tau}$, while all the others are at most $n^{-\tau}$ in absolute value,
- the $k$-variance of the vertex representatives constructed by the $k - 1$ transformed structural eigenvectors is $O(n^{-2\tau})$,
- with any “small” $\alpha > 0$, the $V_a, V_b$ pairs are $\alpha$-volume regular, almost surely.
Motivation
Spectral clustering of graphs
Noisy random graphs
Biclustering of contingency tables

10-fold blow up
20-fold blow up
Motivation

Spectral clustering of graphs

Noisy random graphs

Biclustering of contingency tables

30-fold blow up
40-fold blow up
50-fold blow up
<table>
<thead>
<tr>
<th>Motivation</th>
<th>Spectral clustering of graphs</th>
<th>Noisy random graphs</th>
<th>Biclustering of contingency tables</th>
</tr>
</thead>
</table>

60-fold blow up
70-fold blow up
80-fold blow up
90-fold blow up
100-fold blow up
Before sorting and clustering the vertices
Bickel and Chen (PNAS, 2009) introduced a random block model which is, in fact, a generalised random graph.

- For given $k$, vertices independently belong to cluster $V_a$ with probability $\pi_a$, $a = 1, \ldots, k$; $\sum_{a=1}^{k} \pi_a = 1$.
- Vertices of $V_a$ and $V_b$ are connected independently of each other with probabilities $\mathbb{P}(i \sim j | i \in V_a, j \in V_b) = p_{ab}$, $1 \leq a, b \leq k$.

The parameters are collected in the vector $\pi = (\pi_1, \ldots, \pi_k)$ and in the $k \times k$ symmetric matrix $P$ of $p_{ab}$’s.

Our statistical sample is the $n \times n$ symmetric, 0-1 adjacency matrix $A = (a_{ij})$ of a simple graph on $n$ vertices. There are no loops, so the diagonal entries are zeroes. We want to estimate the parameters of the above block model.
By the theorem of mutually exclusive and exhaustive events, the likelihood function is

\[
\frac{1}{2} \sum_{1 \leq a, b \leq k} \pi_a \pi_b \prod_{i \in C_a, j \in C_b, i \neq j} p_{ab}^{a_{ij}} (1 - p_{ab})^{(1 - a_{ij})}
\]

\[
= \frac{1}{2} \sum_{1 \leq a, b \leq k} \pi_a \pi_b \cdot p_{ab}^{e_{ab}} \cdot (1 - p_{ab})^{(n_{ab} - e_{ab})}
\]

reminiscent of the mixture of binomial distributions, where

- \( e_{ab} \) is the number of edges connecting vertices of \( V_a \) and \( V_b \) \((a \neq b)\);
- \( e_{aa} \) is twice the number of edges with endpoints in \( V_a \);
- \( n_{ab} = |V_a| \cdot |V_b| \) if \( a \neq b \), and \( n_{aa} = |V_a| \cdot (|V_a| - 1), a = 1, \ldots, k \).
Here $A$ is the incomplete data specification, as the cluster memberships are missing. Therefore we complete our data matrix $A$ by latent membership vectors $\Delta_1, \ldots, \Delta_n$ of the vertices that are $k$-dimensional i.i.d. $Poly(1, \pi)$ random vectors.

$\Delta_i = (\Delta_{1i}, \ldots, \Delta_{ki})$, where $\Delta_{ai} = 1$ if $i \in V_a$ and zero otherwise. Thus, the sum of the coordinates of any $\Delta_i$ is 1, and $P(\Delta_{ai} = 1) = \pi_a$.

The likelihood function above is

$$
\frac{1}{2} \sum_{1 \leq a, b \leq k} \pi_a \pi_b \cdot p_{ab} \sum_{i,j: i \neq j} \Delta_{ai} \Delta_{bj} a_{ij} \cdot (1 - p_{ab}) \sum_{i,j: i \neq j} \Delta_{ai} \Delta_{bj} (1 - a_{ij})
$$

that is maximized in the alternating E and M steps of the EM-algorithm.
We remark that the complete likelihood would be the squareroot of

\[
\prod_{1\leq a,b\leq k} p_{ab}^{e_{ab}} \cdot (1 - p_{ab})^{(n_{ab} - e_{ab})}
\]

\[
= \prod_{a=1}^{k} \prod_{i=1}^{n} \prod_{b=1}^{k} \left[ p_{ab}^{\sum_{j:j\neq i} \Delta_{bij} a_{ij}} \cdot (1 - p_{ab})^{\sum_{j:j\neq i} \Delta_{bij} (1 - a_{ij})} \right] \Delta_{ai}
\]

that is valid only in case of known cluster memberships.
Starting with initial parameter values \(\pi^{(0)}, P^{(0)}\) and membership vectors \(\Delta_{1}^{(0)}, \ldots, \Delta_{n}^{(0)}\), the \(t\)-th step of the iteration is the following \((t = 1, 2, \ldots)\).
E-step

We calculate the **conditional expectation** of each $\Delta_i$ conditioned on the model parameters and on the other cluster assignments obtained in the $(t-1)$-th step (denoted by $M^{(t-1)}$).

By the Bayes theorem, the **responsibility of vertex $i$ for cluster $a$**:

$$
\pi_{ai}^{(t)} = \mathbb{E}(\Delta_{ai} | M^{(t-1)}) = \mathbb{P}(\Delta_{ai} = 1 | M^{(t-1)}) = \frac{\mathbb{P}(M^{(t-1)} | \Delta_{ai} = 1) \cdot \pi_{a}^{(t-1)}}{\sum_{l=1}^{k} \mathbb{P}(M^{(t-1)} | \Delta_{li} = 1) \cdot \pi_{l}^{(t-1)}}
$$

($a = 1, \ldots, k; \ i = 1, \ldots, n$). Thus, for each $i$, $\pi_{ai}^{(t)}$ is proportional to the numerator, where

$$
\mathbb{P}(M^{(t-1)} | \Delta_{ai} = 1) = \prod_{b=1}^{k} (p_{ab}^{(t-1)})^{\sum_{j \neq i} \Delta_{bj}^{(t-1)} a_{ij}} (1 - p_{ab}^{(t-1)})^{\sum_{j \neq i} \Delta_{bj}^{(t-1)}(1 - a_{ij})}
$$

is the part of the likelihood affecting vertex $i$ under the condition $\Delta_{ai} = 1$. 
M-step

For all $a, b$ pairs separately, we maximize the truncated binomial likelihood

$$p_{ab} = \frac{\sum i,j: i \neq j \pi^{(t)}_{ai} \pi^{(t)}_{bj} a_{ij}}{\sum i,j: i \neq j \pi^{(t)}_{ai} \pi^{(t)}_{bj}} \cdot (1 - p_{ab}) \sum i,j: i \neq j \pi^{(t)}_{ai} \pi^{(t)}_{bj} (1 - a_{ij})$$

with respect to the parameter $p_{ab}$. Obviously, the maximum is attained by the following estimators of $p_{ab}$’s comprising the symmetric matrix $P^{(t)}$:

$$p_{ab}^{(t)} = \frac{\sum i,j: i \neq j \pi^{(t)}_{ai} \pi^{(t)}_{bj} a_{ij}}{\sum i,j: i \neq j \pi^{(t)}_{ai} \pi^{(t)}_{bj}}, \quad 1 \leq a \leq b \leq k,$$

where edges connecting vertices of clusters $a$ and $b$ are counted fractionally, multiplied by the membership probabilities of their endpoints.
The ML-estimator of $\pi$ in the $t$-th step is $\pi^{(t)}$ of coordinates $\pi^{(t)}_a = \frac{1}{n} \sum_{i=1}^{n} \pi^{(t)}_{ai}$ ($a = 1, \ldots, k$), while that of the membership vector $\Delta_i$ is obtained by discrete maximization: $\Delta^{(t)}_{ai} = 1$, if $\pi^{(t)}_{ai} = \max_{b \in \{1, \ldots, k\}} \pi^{(t)}_{bi}$ and 0, otherwise. (In case of ambiguity, the cluster with the smallest index is selected.) This choice of $\pi$ will increase the likelihood.

The above algorithm is a special case of so-called Collaborative Filtering, see Hoffman, T., Puzicha, J., Ungar, L., Foster, D.

According to the general theory of EM-algorithm (Dempster, Laird, Rubin, J. R. Statist. Soc B 39, 1977), in exponential families (as in the present case), convergence to a local maximum can be guaranteed (depending on the starting values), but it runs in polynomial time in $n$. 
Testable weighted graph parameters

Both edges and vertices may have nonnegative weights.

**Convergent graph sequences:** $G_n$’s become more and more similar as $n \to \infty$ (homomorphism densities, convergence in the cut metric, graphons), Lovász, L. et al. (2006–).

$f$ is a **testable weighted graph parameter** if

- $f(G_n)$ converges whenever $G_n$ converges,
- It can be consistently estimated based on a sufficiently large sample selected from the huge graph by an appropriate randomization.
Examples:

- **Balanced minimum multiway cuts**, normalized and modularity cuts (statistical physics, Hamiltonian).
- If the vertex weights are the generalized degrees, the normalized modularity spectrum is testable:

\[
\mu_{n,i}(M_D(G_n)) \rightarrow \lambda_i(P_W) \quad (n \to \infty) \quad i = 1, 2, \ldots
\]

where \(\mu_{n,i}(M_D(G_n))\) is the \(i\)th largest absolute value eigenvalue of the normalized modularity matrix of \(G_n\), and \(\lambda_i(P_W)\) is the \(i\)th largest absolute value eigenvalue of the operator taking conditional expectation with respect to the joint measure embodied by \(W\) which is **compact** with discrete spectrum.

- For any \(k > 1\), the eigen-subspace belonging to the transformed \(\mu_{n,1}(M_D(G_n)), \ldots, \mu_{n,k-1}(M_D(G_n))\) also converges to the corresponding eigensubspace of \(P_W\).
- The \(k\)-variances also converge.
Biclustering of contingency tables

\((\text{Row}, \text{Col}, \mathbf{C})\) contingency table
Row set: \(\text{Row} = \{1, \ldots, n\}\)
Column set: \(\text{Col} = \{1, \ldots, m\}\)
\(\mathbf{C}\): \(n \times m\) matrix of entries \(c_{ij} \geq 0\).
\(c_{ij}\): some kind of interaction between the objects representing row \(i\) and column \(j\), where 0 means no interaction at all.

\[
d_{\text{row},i} = \sum_{j=1}^{m} c_{ij}, \quad i = 1, \ldots, n
\]

\[
d_{\text{col},j} = \sum_{i=1}^{n} c_{ij}, \quad j = 1, \ldots, m
\]

\(\mathbf{D}_{\text{row}} = \text{diag}(d_{\text{row},1}, \ldots, d_{\text{row},n}), \quad \mathbf{D}_{\text{col}} = \text{diag}(d_{\text{col},1}, \ldots, d_{\text{col},m})\)
Given the integer $1 \leq k \leq \min\{n, m\}$: find $k$-dimensional representatives $r_1, \ldots, r_n \in \mathbb{R}^k$ of the rows and $c_1, \ldots, c_m \in \mathbb{R}^k$ of the columns such that they minimize

$$Q_k = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \| r_i - c_j \|^2$$

under the conditions

$$\sum_{i=1}^{n} d_{row, i} r_i r_i^T = I_k, \quad \sum_{j=1}^{m} d_{col, j} c_j c_j^T = I_k$$
Equivalence to the correspondence analysis

\[ X := (r_1^T, \ldots, r_n^T)^T = (x_1, \ldots, x_k) \quad n \times k \]
\[ Y := (c_1^T, \ldots, c_m^T)^T = (y_1, \ldots, y_k) \quad m \times k \]

Constraints:

\[ X^T D_{\text{row}} X = I_k, \quad Y^T D_{\text{col}} Y = I_k. \]

\[ Q_k = \sum_{i=1}^{n} d_{\text{row},i} \| r_i \|^2 + \sum_{j=1}^{m} d_{\text{col},j} \| c_j \|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} r_i^T c_j \]

\[ = 2k - \text{tr} X^T C Y = 2k - \text{tr} (D_{\text{row}}^{1/2} X)^T (D_{\text{row}}^{-1/2} C D_{\text{col}}^{-1/2}) (D_{\text{col}}^{1/2} Y), \]

where \( D_{\text{row}}^{1/2} X \) and \( D_{\text{col}}^{1/2} Y \) are suborthogonal matrices.
Correspondence matrix / normalized contingency table

\[ C_{\text{corr}} := D_{\text{row}}^{-1/2} CD_{\text{col}}^{-1/2} \]

**SVD:**

\[ C_{\text{corr}} = \sum_{i=1}^{r} s_i v_i u_i^T, \]

where \( r \leq \min\{n, m\} \) is the rank of \( C_{\text{corr}} \), or equivalently (as there are not identically zero rows or columns), that is the rank of \( C \).

\( 1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0 \): non-zero singular values of \( C_{\text{corr}} \) with singular vector pairs \( v_i, u_i \) \((i = 1, \ldots, r)\).

1 is a single singular value if \( C_{\text{corr}} \) (or equivalently, \( C \)) is irreducible. In this case

\[ v_1 = (\sqrt{d_{\text{row},1}}, \ldots, \sqrt{d_{\text{row},n}})^T \quad \text{and} \quad u_1 = (\sqrt{d_{\text{col},1}}, \ldots, \sqrt{d_{\text{col},m}})^T. \]
Representation theorem for contingency tables

**Theorem**

Let \((\text{Row}, \text{Col}, \mathbf{C})\) be an irreducible contingency table with the above SVD of its correspondence matrix \(\mathbf{C}_{\text{corr}}\). Let \(k \leq r\) be a positive integer such that \(s_k > s_{k+1}\). Then the minimum of \(Q_k\) under the given constraints is \(2k - \sum_{i=1}^{k} s_i\) and it is attained with the optimum row representatives \(r_1^*, \ldots, r_n^*\) and column representatives \(c_1^*, \ldots, c_m^*\), the transposes of which are row vectors of \(\mathbf{X}^* = \mathbf{D}_{\text{row}}^{-1/2} (\mathbf{v}_1, \ldots, \mathbf{v}_k)\) and \(\mathbf{Y}^* = \mathbf{D}_{\text{col}}^{-1/2} (\mathbf{u}_1, \ldots, \mathbf{u}_k)\), respectively.

Remark: if 1 is a single singular value, the first columns of \(\mathbf{X}^*\) and \(\mathbf{Y}^*\): \(\mathbf{D}_{\text{row}}^{-1/2} \mathbf{v}_1\) and \(\mathbf{D}_{\text{col}}^{-1/2} \mathbf{u}_1\) are the constantly 1 vectors in \(\mathbb{R}^n\) and \(\mathbb{R}^m\), respectively.
Normalized two-way cuts of a contingency table

\((Row, Col, C)\): \(n \times m\) contingency table
\(k (0 < k \leq r)\): fixed integer

Partition simultaneously the rows and columns into disjoint, nonempty subsets

\[ Row = R_1 \cup \cdots \cup R_k, \quad Col = C_1 \cup \cdots \cup C_k \]

such that the cuts

\[ c(R_a, C_b) = \sum_{i \in R_a} \sum_{j \in C_b} c_{ij}, \quad a, b = 1, \ldots, k \]

between the row-column cluster pairs be as homogeneous as possible.
The normalized two-way cut of the contingency table with respect to the above \( k \)-partitions \( P_{\text{row}} = (R_1, \ldots, R_k) \) and \( P_{\text{col}} = (C_1, \ldots, C_k) \) of its rows and columns and to the collection of signs \( \sigma \):

\[
\nu_k(P_{\text{row}}, P_{\text{col}}, \sigma) = \sum_{a=1}^{k} \sum_{b=1}^{k} \left( \frac{1}{\text{Vol}(R_a)} + \frac{1}{\text{Vol}(C_b)} + \frac{2\delta_{ab}\sigma_{ab}}{\sqrt{\text{Vol}(R_a)\text{Vol}(C_b)}} \right) c(R_a, C_b),
\]

where

\[
\text{Vol}(R_a) = \sum_{i \in R_a} \sum_{j=1}^{m} c_{ij}, \quad \text{Vol}(C_b) = \sum_{j \in C_b} \sum_{i=1}^{n} c_{ij}
\]

are volumes of the clusters, and \( \sigma = (\sigma_{11}, \ldots, \sigma_{kk}) \) with \( \sigma_{aa} = \pm 1 \) (\( a = 1, \ldots, k \)), whereas \( \sigma_{ab} \) has no relevance if \( a \neq b \).

The objective function also penalizes clusters of extremely different volumes.
Theorem

The normalized two-way cut of the contingency table $C$:

$$
\nu_k(C) := \min_{P_{row}, P_{col}, \sigma} \nu_k(P_{row}, P_{col}, \sigma) \geq 2k - \sum_{i=1}^{k} s_i.
$$

Proof: $\nu_k(P_{row}, P_{col}, \sigma)$ is $Q_k$ in the special representation, where the column vectors of $X$ and $Y$ are partition vectors belonging to $P_{row}$ and $P_{col}$:

$$
x_{ia} := \frac{1}{\sqrt{\text{Vol}(R_a)}} \quad \text{if } i \in R_a \text{ and } 0 \text{ otherwise } (a = 1, \ldots, k)
$$

$$
y_{jb} := \frac{\sigma_{bb}}{\sqrt{\text{Vol}(C_b)}} \quad \text{if } j \in C_b \text{ and } 0 \text{ otherwise } (b = 1, \ldots, k)
$$

$X = (x_1, \ldots, x_k)$ and $Y = (y_1, \ldots, y_k)$ satisfy the conditions imposed on the representatives and

$$
\|r_i - c_j\|^2 = \frac{1}{\text{Vol}(R_a)} + \frac{1}{\text{Vol}(C_b)} + \frac{2\delta_{ab}\sigma_{bb}}{\sqrt{\text{Vol}(R_a)\text{Vol}(C_b)}} \quad \text{if } i \in R_a, j \in C_b.
$$
Symmetric contingency table = edge-weighted graph

- If the \( k - 1 \) largest absolute value eigenvalues of the normalized modularity matrix are all positive: the \( k - 1 \) largest singular values (apart of the 1) of \( C_{\text{corr}} \) are identical to the \( k - 1 \) largest eigenvalues of \( M_D \), and the left and right singular vectors are identical to the corresponding eigenvector with the same orientation \( \implies r_i = c_i \) for all \( (k - 1) \)-dimensional row and column representatives; \( \nu_k(C) = 2f_k(G) \implies \) the normalized two-way cut favors \( k \)-partitions with low inter-cluster edge-densities.

- If all the \( k - 1 \) largest absolute value eigenvalues of the normalized modularity matrix are negative: \( r_i = -c_i \), and any (but only one) of them can be the corresponding vertex representative; \( \nu_k(C) \) differs from \( f_k(G) \) in that it also counts the edge-weights within the clusters. Here, minimizing \( \nu_k(C) \), rather a so-called anti-community structure is detected.
Regular row-column cluster pairs

In the generic case, for given $k$, if the clusters are formed via applying the $k$-means algorithm for the row- and column representatives, respectively, then the so obtained row–column cluster pairs are homogeneous in the sense, that they form equally dense parts of the contingency table.

**Definition**

The row–column cluster pair $R \subset Row$, $C \subset Col$ of the contingency table $(Row, Col, C)$ (where the sum of the entries is 1) is $\gamma$-volume regular, if for all $X \subset R$ and $Y \subset C$ the relation

$$|c(X, Y) - \rho(R, C)Vol(X)Vol(Y)| \leq \gamma \sqrt{Vol(R)Vol(C)}$$

holds, where $\rho(R, C) = \frac{c(R,C)}{Vol(R)Vol(C)}$ is the relative inter-cluster density of the row–column pair $R, C$. 
Weighted $k$-variances

The weighted $k$-variance of the $k$-dimensional row representatives:

$$S_k^2(X) = \min_{P_{row}, k \in P_{row}} S_k^2(P_k, X) = \min_{(R_1, \ldots, R_k)} \sum_{a=1}^{k} \sum_{j \in R_a} d_{row,j} \| r_j - b_a \|^2,$$

where $b_a = \frac{1}{\text{Vol}(R_a)} \sum_{j \in R_a} d_{row,j} r_j \quad (a = 1, \ldots, k)$.

The weighted $k$-variance of the $k$-dimensional column representatives:

$$S_k^2(Y) = \min_{Q_{col}, k \in P_{col}} S_k^2(Q_k, Y) = \min_{(C_1, \ldots, C_k)} \sum_{b=1}^{k} \sum_{j \in C_b} d_{col,j} \| c_j - b_b \|^2,$$

where $b_b = \frac{1}{\text{Vol}(C_b)} \sum_{j \in C_b} d_{col,j} c_j \quad (b = 1, \ldots, k)$.

Observe, that the trivial vector components can be omitted, and the $k$-variance of the so obtained $(k - 1)$-dimensional representatives will be the same.
Volume regularity versus spectral properties

**Theorem**

Let \((\text{Row}, \text{Col}, \mathbf{C})\) be a contingency table of \(n\) rows and \(m\) columns, with row- and column sums \(d_{\text{row},1}, \ldots, d_{\text{row},n}\) and \(d_{\text{col},1}, \ldots, d_{\text{col},m}\), respectively. Suppose that \(\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} = 1\) and there are no dominant rows and columns: \(d_{\text{row},i} = \Theta(1/n), (i = 1, \ldots, n)\) and \(d_{\text{col},j} = \Theta(1/m), (j = 1, \ldots, m)\) as \(n, m \to \infty\).

Let the singular values of \(\mathbf{C}_{\text{corr}}\) be

\[
1 = s_1 > s_2 \geq \cdots \geq s_k > \varepsilon \geq s_i, \quad i \geq k + 1.
\]

The partition \((R_1, \ldots, R_k)\) of Row and \((C_1, \ldots, C_k)\) of Col are defined so that they minimize the weighted \(k\)-variances \(S_k^2(\mathbf{X}^*)\) and \(S_k^2(\mathbf{Y}^*)\) of the row and column representatives. Suppose that there are constants \(0 < K_1, K_2 \leq \frac{1}{k}\) such that \(|R_i| \geq K_1 n\) and \(|C_i| \geq K_2 m\) \((i = 1, \ldots, k)\), respectively. Then the \(R_i, C_j\) pairs are \(O(\sqrt{2k(S_k(\mathbf{X}^*) + S_k(\mathbf{Y}^*)) + \varepsilon})\)-volume regular \((i, j = 1, \ldots, k)\).
Noisy contingency table sequences

**Figure:** noisy table; table close to the limit; approximation by SVD

THE END