Discrete Mathematics 117 (1993) 19–39
North-Holland

Spectra, Euclidean representations and clusterings of hypergraphs

Marianna Bolla

Mathematical Institute, Hungarian Academy of Sciences, H-1053 Budapest, Reáltanoda u. 13–15. Hungary and DIMACS Center, Rutgers University, P.O. Box 1179, Piscataway, NJ 08855-1179, USA

Received 6 November 1990

Abstract

Bolla, M., Spectra, Euclidean representations and clusterings of hypergraphs, Discrete Mathematics 117 (1993) 19–39.

We would like to classify the vertices of a hypergraph in the way that 'similar' vertices (those having many incident edges in common) belong to the same cluster. The problem is formulated as follows: given a connected hypergraph on n vertices and fixing the integer k ($1 < k \leqslant n$), we are looking for k-partition of the set of vertices such that the edges of the corresponding cut-set be as few as possible. We introduce some combinatorial measures characterizing this structural property and give upper and lower bounds for them by means of the k smallest eigenvalues of the hypergraph. For this purpose the notion of *spectra of hypergraphs* — which is the generalization of C-spectra of graphs — is also introduced together with *k-dimensional Euclidean representations*. We shall show that the existence of k 'small' eigenvalues is a necessary but not sufficient condition for the existence of a good clustering. In addition the representatives of the vertices in an optimal k-dimensional Euclidean representation of the hypergraph should be well separated by means of their Euclidean distances. In this case the k-partition giving the optimal clustering is also obtained by this classification method.

1. Introduction

The C-spectrum of a graph G (e.g. in [12]) is defined by the eigenvalues of its *Laplacian*, $\boldsymbol{C}(G) = \boldsymbol{D}(G) - \boldsymbol{A}(G)$, $\boldsymbol{A}(G)$ being the adjacency matrix of G and $\boldsymbol{D}(G)$ denoting the valency matrix of the set of vertices. $\boldsymbol{C}(G)$ is symmetric, singular and positive semidefinite. Fiedler in [12] and [13] investigates the smallest positive eigenvalue of $\boldsymbol{C}(G)$ in relation to the vertex- and edge-connectivity and he calls it the *algebraic connectivity* of the graph G.

In the present paper the notion of C-spectra of graphs is extended to hypergraphs and the notion of Euclidean representation of hypergraphs is also introduced in the

Correspondence to: Marianna Bolla, Mathematical Institute, Hungarian Academy of Sciences, H-1053 Budapest, Reáltanoda u. 13–15. Hungary and DIMACS Center, Rutgers University, P.O. Box 1179, Piscataway, NJ 08855-1179, USA.

following way: given a hypergraph H on n vertices and fixing the integer k ($1 < k \leqslant n$), we are looking for representation $\phi: V(H) \to R^k$ and $\psi: E(H) \to R^k$ such that

$$\sum_{v \in V(H)} \phi(v)\phi(v)^{\mathrm{T}} = \boldsymbol{I}_k$$

and the cost function

$$Q = \sum_{e \in E(H)} K(e)$$

is minimized, where the *cost* $K(e)$ of an edge e is defined by

$$K(e) := \sum_{v \in e} \|\phi(v) - \psi(e)\|^2.$$

Minimizing the cost function Q means finding a 'minimal variance placement' of the vertices in the k-dimensional Euclidean space so that vertices having many incident edges in common be 'close' to each other in Euclidean metric. This gives rise to clustering the representatives of the vertices in a Euclidean space.

Theorem 2.2 states that the minimum of the cost function Q conditioned on $\sum_{v \in V(H)} \phi(v)\phi(v)^{\mathrm{T}} = \boldsymbol{I}_k$ is the sum of the k smallest eigenvalues of the Laplacian $\boldsymbol{B}(H)$ defined by

$$\boldsymbol{B}(H) = \boldsymbol{D}_v(H) - \boldsymbol{A}(H)\boldsymbol{D}_e^{-1}(H)\boldsymbol{A}(H)^{\mathrm{T}},$$

where $\boldsymbol{A}(H)$ is the vertex-edge incidence matrix of H, while $\boldsymbol{D}_v(H)$ and $\boldsymbol{D}_e(H)$ are diagonal matrices with the vertex- and edge-valencies of H in their main diagonals, respectively. The Laplacian is symmetric, singular and positive semidefinite.

The above minimum is attained for any pair of representations ϕ^* and ψ^* which assign the column vectors of matrices $\boldsymbol{X}^*(H)$ and $\boldsymbol{Y}^*(H)$ to the vertices and to the edges respectively, where the $k \times n$ matrix $\boldsymbol{X}^*(H)$ contains k pairwise orthonormal eigenvectors corresponding to the k smallest eigenvalues of $\boldsymbol{B}(H)$ in its rows and $\boldsymbol{Y}^*(H) = \boldsymbol{X}^*(H)\boldsymbol{A}(H)\boldsymbol{D}_e^{-1}(H)$. We speak of *optimal k-dimensional Euclidean representation* of the hypergraph H, if the vertices and edges are represented by an optimal ϕ^* and ψ^* pair.

Section 6 contains some remarks on spectra of hypergraphs and on Euclidean representations of some special graphs. In the sequel we indicate H in the spectral characteristics only if a hypergraph different from the underlying one is investigated. Similarly, if the dimension k is clear from the context, we simply speak of Euclidean representation.

Our purpose is to relate spectral characteristics of a hypergraph to its structural properties. The investigated property is the following: for fixed k there exists a k-partition (k disjoint, non-empty subsets) of the set of vertices with many of their incident edges concentrated within these subsets. A k-partition is sometimes called *k-coloring*, while the edges of the corresponding cut-set are called *multi-colored* in this

coloring. In Section 3, we define combinatorial measures characterizing the number and color-distributions of the multi-colored edges. Their minima through all of the k-partitions are called cardinality of the minimal k-sector and minimal weighted cut, respectively. They are denoted by $\theta_k(H)$ and $\nu_k(H)$, respectively (more precise definitions are given in Section 3).

Let $0=\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$ be the eigenvalues of H. Theorem 3.5 gives the following bounds for the sum of the k smallest eigenvalues by the above defined combinatorial measures:

$$c_n\theta_k(H)\leqslant\sum_{j=1}^{k}\lambda_j\leqslant\nu_k(H),$$

where the constant c_n merely depends on the number n of the vertices. The upper bound shows that existence of k small eigenvalues is a necessary condition for existence of a good clustering, and the spectrum can give us an idea about the number k of the clusters. But this condition is not sufficient, since there are graphs for which the lower bound is attained in order of magnitude of the constant c_n, but they cannot be classified into k clusters in a sensible way. Therefore, in addition the representatives of the vertices in an optimal k-dimensional representation of H have to be investigated. Provided the representatives of the vertices can be classified into k well-separated clusters by means of their Euclidean distances (see Definition 3.7), the vertices of the hypergraph can be well classified into k clusters, such that the above defined combinatorial measures be 'small'. In Theorem 3.8 it is stated that the diameters of the clusters being less than $1/2\sqrt{n}$, the relation

$$\nu_k(H)\leqslant 4\sum_{j=1}^{k}\lambda_j$$

holds. If there exists a well-separated k-partition of the set of vertices, it is uniquely determined. In this case, even the k-partition of the vertices giving the optimal classification is obtained by k-means clustering of their k-dimensional representatives.

Of course, these bounds can be reached for special hypergraphs and for specific k but in general, approximate results can only be obtained by means of these spectral techniques. However, these results are useful and well-adopted to automatic computation in case of large hypergraphs when one is not interested in strict structural properties.

Large hypergraphs often arise in statistical analysis of several mutually dependent binary variables where the vertices correspond to the variables, the edges to the objects and the incidence relation depends on, whether an object possesses the property represented by the variable in question or not. The *iterative algorithm* — introduced in Chapter 5 — applies the spectral technique in one of the steps of the iteration, in the other steps the partitions and the dimensions are determined.

2. Optimal Euclidean representations

Let $H=(V,E)$ be a hypergraph with vertex-set $V=\{v_1\cdots v_n\}$ and edge-set $E=\{e_1\cdots e_m\}$. H is given by its $n\times m$ *vertex-edge incidence matrix* $\boldsymbol{A}$ with entries $a_{ji}=\mathcal{I}(v_j\in e_i)$, where

$$\mathcal{I}(v\in e)=\begin{cases}1 & \text{if } v\in e,\\ 0 & \text{otherwise,}\end{cases}$$

and the relation $v\in e$ denotes that the vertex v is incident with the edge e.

Let k $(1<k\leqslant n)$ be a fixed integer. We are looking for k-dimensional representatives $\boldsymbol{x}_j:=\phi(v_j)$ and $\boldsymbol{y}_i:=\psi(e_i)$ of the vertices and edges respectively such that

$$\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}}=\boldsymbol{I}_k \tag{2.1}$$

and the sum of the costs of edges

$$Q=\sum_{i=1}^{m}K(e_i)=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ji}\|\boldsymbol{x}_j-\boldsymbol{y}_i\|^2 \tag{2.2}$$

is minimized, where the cost $K(e_i)$ of the edge e_i is defined by

$$K(e_i):=\sum_{j=1}^{n}a_{ji}\|\boldsymbol{x}_j-\boldsymbol{y}_i\|^2. \tag{2.3}$$

Let $\bar{\boldsymbol{x}}(e)$ denote the center of gravity of the representatives of those vertices the edge e is incident with:

$$\bar{\boldsymbol{x}}(e):=\frac{1}{|e|}\sum_{j=1}^{n}\mathcal{I}(v_j\in e)\boldsymbol{x}_j. \tag{2.4}$$

Let the $k\times n$ and $k\times m$ matrices $\boldsymbol{X}:=(\boldsymbol{x}_1\cdots\boldsymbol{x}_n)$ and $\boldsymbol{Y}:=(\boldsymbol{y}_1\cdots\boldsymbol{y}_m)$ contain the vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ and $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m$ as their columns respectively. Let $\boldsymbol{D}_v$ and $\boldsymbol{D}_e$ be $n\times n$ and $m\times m$ diagonal matrices with the vertex-valencies $s_1,\ldots,s_n$ and the edge-valencies $z_1,\ldots,z_m$ in their main diagonals, where

$$s_j=\sum_{i=1}^{m}a_{ji}\quad\text{and}\quad z_i=\sum_{j=1}^{n}a_{ji}.$$

We can suppose that $\boldsymbol{D}_e$ is not singular.

With these notations the cost $K(e)$ of an edge e is decreased, if $\bar{\boldsymbol{x}}(e)$ is substituted for $\psi(e)$:

$$K(e)\geqslant\sum_{j=1}^{n}\mathcal{I}(v_j\in e)\,\|\boldsymbol{x}_j-\bar{\boldsymbol{x}}(e)\|^2,\quad e\in E.$$

Denoting the right-hand side by $L(e, X)$ — by means of a simple geometrical argument — it can be written as

$$L(e, X)=\frac{1}{2|e|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{I}(v_i\in e)\,\mathcal{I}(v_j\in e)\,\|x_i-x_j\|^2, \quad e\in E. \tag{2.5}$$

Sometimes we shall refer to $L(e, X)$ as the *variance of the edge* e in the k-dimensional Euclidean representation X of the vertices. By setting $L(X):=\sum_{e\in E}L(e, X)$, the inequality $Q\geqslant L(X)$ always holds. But $L(X)$ is a quadratic form, since

$$\begin{aligned}L(X)&=\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{2}\sum_{e\in E}\mathcal{I}(v_i\in e)\,\mathcal{I}(v_j\in e)\frac{1}{|e|}\right]\|x_i-x_j\|^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}x_i^{\mathrm{T}}x_j\end{aligned} \tag{2.6}$$

with

$$b_{ij}=\begin{cases}-\displaystyle\sum_{e\in E}\mathcal{I}(v_i\in e)\,\mathcal{I}(v_j\in e)\frac{1}{|e|} & \text{if } i\neq j,\\ s_i-\displaystyle\sum_{e\in E}\mathcal{I}(v_i\in e)\frac{1}{|e|}=s_i'-\sum_{\substack{e\in E\\|e|>1}}\mathcal{I}(v_i\in e)\frac{1}{|e|} & \text{if } i=j,\end{cases} \tag{2.7}$$

where $s_i'=\#\{e\in E\colon v_i\in e,\ |e|>1\}$.

Definition 2.1. The matrix of the quadratic form (2.6) is called the *Laplacian* of the hypergraph H, and it is denoted by $\boldsymbol{B}$.

In matrix notation, $\boldsymbol{B}$ can be written as $\boldsymbol{D}_v-\boldsymbol{A}\boldsymbol{D}_e^{-1}\boldsymbol{A}^{\mathrm{T}}$.

The quadratic form $L(X)$ is equal to $\mathrm{tr}\,\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^{\mathrm{T}}$, and it is to be minimized on $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{I}_k$. As the $n\times n$ matrix $\boldsymbol{B}$ is symmetric and positive semidefinite, by means of a theorem for the extrema of quadratic forms (see e.g. [21, p. 51]) the following *Representation Theorem* is arrived at.

Theorem 2.2. *The minimum of the cost function* (2.2) *conditioned on* (2.1) *is*

$$\sum_{j=1}^{k}\lambda_j, \tag{2.8}$$

where $0=\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$ *are the eigenvalues of the Laplacian* $\boldsymbol{B}$ *and it is attained, when the k-dimensional Euclidean representation* $\boldsymbol{X}$ *of the vertices contains pairwise orthonormal eigenvectors corresponding to the k smallest eigenvalues of* $\boldsymbol{B}$ *in its rows. If such an* $\boldsymbol{X}$ *is denoted by* $\boldsymbol{X}^*$, *the optimal choice for the k-dimensional Euclidean representation* $\boldsymbol{Y}$ *of the edges is* $\boldsymbol{Y}^*=\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{D}_e^{-1}$.

Let $\boldsymbol{R}$ be a $k \times k$ orthogonal matrix ($\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{I}_k$). Then neither the objective function nor the constraint is effected by the substitution $\boldsymbol{X}' = \boldsymbol{R}\boldsymbol{X}$. Thus, together with an optimal $\boldsymbol{X}^*$, the matrix $\boldsymbol{R}\boldsymbol{X}^*$ is optimal too. But apart from k-dimensional rotations, in case of distinct eigenvalues the optimal $\boldsymbol{X}^*$ is uniquely determined by the Laplacian $\boldsymbol{B}$. Otherwise their rows can be chosen appropriately within the eigenspaces belonging to the multiple eigenvalues.

Definition 2.3. If k-dimensional representatives $\boldsymbol{x}^* = \phi^*(v)$ and $\boldsymbol{y}^* = \psi^*(e)$, which are the column vectors of an optimal $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ pair are assigned to the vertices and to the edges respectively, we speak of *optimal k-dimensional Euclidean representation* of the hypergraph H.

We remark that the dimension k does not play an important role here yet, since for any k ($1 \leqslant k < n$) an optimal $(k+1)$-dimensional Euclidean representation is obtained from an optimal k-dimensional one by introducing a subsequent eigenvector in the rows of $\boldsymbol{X}$.

It can be seen from formula (2.7) that the loops (edges with $|e| = 1$) do not contribute to the entries of the Laplacian, therefore in the future only hypergraphs without loops will be investigated.

Let us also notice that the Laplacian is always singular since all row sums are 0. The eigenvector corresponding to a single zero eigenvalue is a multiple of $\boldsymbol{e}$, where $\boldsymbol{e}$ is the n-dimensional vector of 1s. In this case a k-dimensional Euclidean representation is realized in the $(k-1)$-dimensional subspace of R^k orthogonal to the vector $\boldsymbol{e}$. It is well known that the multiplicity of the zero as an eigenvalue of a hypergraph without loops and isolated vertices is equal to the number of its connected components. In this case the spectrum consists of the spectra of its components, so only spectra of connected hypergraphs are of interest. But in case of connected hypergraphs one can ask how many edges must be removed so that the hypergraph be not connected or consist of k components. How the strongly connected sub-hypergraphs can be recognized on the basis of optimal Euclidean representations? Some of these problems are discussed in the subsequent sections.

3. Relations between spectral and structural properties of hypergraphs

Let $H = (V, E)$, $|V| = n$, $|E| = m$ be a hypergraph without loops and multiple edges, its eigenvalues being $0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ in increasing order. Now we shall give upper and lower bounds for combinatorial measures characterizing k-partitions of the vertex-set of H by means of the k smallest eigenvalues, where k is any natural number between 1 and n. First of all let us introduce the following notions.

Definition 3.1. A k-tuple $(V_1, \ldots, V_k)$ of non-empty subsets of V is called a k-partition of the set of vertices, if $V_i \cap V_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{k} V_i = V$. Sometimes a k-partition is

denoted by P_k, while the set of all k-partitions is denoted by $\mathscr{P}_k$. The *volume* $v(P_k)$ of the k-partition $P_k=(V_1,\dots,V_k)$ is defined by

$$v(P_k):=\sum_{e\in E}\frac{1}{|e|}\sum_{1\leqslant i<j\leqslant k}a_i(e)\,a_j(e)$$

and its *weighted volume* $u(P_k)$ by

$$u(P_k):=\sum_{e\in E}\frac{1}{|e|}\sum_{1\leqslant i<j\leqslant k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)a_i(e)\,a_j(e),$$

where $a_i(e)=|e\cap V_i|$ and $n_i=|V_i|$.

The *minimal k-cut* of H is defined by

$$\mu_k(H)=\min_{P_k\in\mathscr{P}_k} v(P_k), \tag{3.1}$$

while the *minimal weighted k-cut* by

$$\nu_k(H)=\min_{P_k\in\mathscr{P}_k} u(P_k). \tag{3.2}$$

Definition 3.2. The cut-set of the k-partition $P_k=(V_1,\dots,V_k)$ consists of those edges e for which $|e\cap V_i|\neq\emptyset$ holds for at least two different parts of P_k, and it is denoted by $H(P_k)$. The k-partition P_k defines a coloring c of the vertices in the following way: $c(v):=i$, if $v\in V_i$. An edge e is said to be multi-colored in this coloring, if it contains two different vertices v, v' such that $c(v)\neq c(v')$. Thus the cut-set $H(P_k)$ consists of the multi-colored edges. $H(P_k^*)$ is called a minimal k-sector of H, if

$$|H(P_k^*)|=\min_{P_k\in\mathscr{P}_k}|H(P_k)|,$$

and its cardinality is denoted by $\theta_k(H)$.

Remark 3.3.

$$\begin{aligned}
&\mu_2(H)\leqslant\mu_3(H)\leqslant\cdots\leqslant\mu_{n-1}(H)\leqslant\mu_n(H),\\
&\nu_2(H)\leqslant\nu_3(H)\leqslant\cdots\leqslant\nu_{n-1}(H)\leqslant\nu_n(H),\\
&\theta_2(H)\leqslant\theta_3(H)\leqslant\cdots\leqslant\theta_{n-1}(H)\leqslant\theta_n(H)=m.
\end{aligned}\tag{3.3}$$

Proposition 3.4. *Let* $\rho_k(H):=\dfrac{\mu_k(H)}{\theta_k(H)}$. *Then*

$$\frac{1}{2}\leqslant\rho_k(H)\leqslant\frac{k-1}{2k}\max_i z_i\leqslant\frac{k-1}{2k}n,\quad 2\leqslant k<n. \tag{3.4}$$

Proof. To obtain the upper bound, let $P_k^*:=(V_1^*,\dots,V_k^*)$ be a k-partition with $H(P_k^*)=\theta_k(H)$. Then by setting $a_i^*(e):=|e\cap V_i^*|$ we get

$$\mu_k(H) \leqslant v(P_k^*) = \sum_{e \in E} \frac{1}{2} \sum_{i=1}^{k} \frac{a_i^*(e)(|e| - a_i^*(e))}{|e|}$$

$$\leqslant \sum_{e \in E} \frac{1}{2|e|} \sum_{i=1}^{k} \frac{|e|}{k} \left(|e| - \frac{|e|}{k} \right)$$

$$= \sum_{e \in H(P_k^*)} \frac{k-1}{2k} |e| \leqslant \frac{k-1}{2k} \max_i z_i |H(P_k^*)|$$

$$= \frac{k-1}{2k} \max_i z_i \theta_k(H) \leqslant \frac{k-1}{2k} n \theta_k(H).$$

To obtain the lower bound, let $P_k' := (V_1', \ldots, V_k')$ be a k-partition for which $v(P_k') = \mu_k(H)$. Then with any $a_i'(e) := |e \cap V_i'|$ we have:

$$\mu_k(H) = v(P_k') \geqslant \sum_{e \in E} \frac{a_i'(e)(|e| - a_i'(e))}{|e|}$$

$$\geqslant \sum_{e \in H(P_k')} \frac{|e|-1}{|e|} \geqslant \frac{1}{2} |H(P_k')| \geqslant \frac{1}{2} \theta_k(H). \qquad \square$$

Theorem 3.5. *For the sum of the k smallest eigenvalues of the hypergraph H the following upper and lower bounds can be given:*

$$c_n \theta_k(H) \leqslant \sum_{j=1}^{k} \lambda_j \leqslant v_k(H), \tag{3.5}$$

where $c_n = 6/n(n^2-1)$.

From the upper estimation it follows that the existence of k relatively small eigenvalues is a necessary condition for the existence of a good classification (with a small minimal weighted cut). Thus the spectrum can give us some idea about the choice of the number k of the clusters. But the spectrum itself does not say anything about the optimal k-partition, moreover it does not give a sufficient condition for the existence of a good clustering. The lower estimate in (3.5) depends on the constant c_n, and there are graphs for which the lower bound is attained in order of magnitude. E.g. for lattices and spiders (see Section 6, Examples 6.12 and 6.13), which cannot be classified into k clusters in a sensible way. For $k=2$ a more precise lower bound can be obtained.

Theorem 3.6. *Let the hypergraph H be connected, and let λ_2 denote its smallest positive eigenvalue. Then*

$$\lambda_2 \geqslant \begin{cases} 2\left(1 - \cos\frac{\pi}{n}\right)\mu_2(H) & \text{if } 0 \leqslant \mu_2(H) \leqslant \frac{1}{2} s_{\max}, \\ c_1 \mu_2(H) - c_2 s_{\max} & \text{if } \frac{1}{2} s_{\max} < \mu_2(H), \end{cases} \tag{3.6}$$

where $c_1 = 2(\cos \pi/n - \cos 2\pi/n)$, $c_2 = 2 \cos \pi/n (1 - \cos \pi/n)$ and $s_{\max} = \max_j s_j$.

For a graph G it is the same estimate than that given by Fiedler in [12]. He has also given an upper bound for λ_2 by the edge-connectivity $e(G)$ of the graph G. As $\nu_2(H) \leqslant (n/n-1)\mu_2(H)$ and $\mu_2(H) = \frac{1}{2}e(G)$, for the smallest positive eigenvalue of ordinary graphs the upper bound $\nu_2(G)$ is sharper than $\frac{1}{2}e(G)$.

Now we want to recognize optimal k-partitions by means of classification of k-dimensional representatives of the vertices in an optimal k-dimensional Euclidean representation of the hypergraph. The classification is performed by k-means method (introduced by Mac Queen in [19]). We shall be confined to the case, when a 'very' well-separated k-partition of the above k-dimensional points exists.

Definition 3.7. A k-partition $P_k = (V_1, \ldots, V_k)$ is called a well-separated k-partition of the vertex-set V in the k-dimensional Euclidean representation $X = (x_1, \ldots, x_n)$ of the vertices, if for the coloring c belonging to P_k the relation $\alpha(P_k) > 1$ holds, where

$$\alpha(P_k) := \frac{\min\limits_{c(v_i) \neq c(v_j)} \|x_i - x_j\|}{\max\limits_{c(v_i) = c(v_j)} \|x_i - x_j\|} \tag{3.7}$$

(In the case when there exists a well-separated k-partition of the k-dimensional points $x_1, \ldots, x_n$, Dunn in [7–9] proved its uniqueness and gave an algorithm to determine the k well-separated clusters of x_j-s. He also proved, that the larger $\alpha(P_k)$ is, the better the separation and the quicker the algorithm is.)

Theorem 3.8. *Assume that for some $k < n$ there exists a well-separated k-partition of the vertex set V in an optimal k-dimensional Euclidean representation of the vertices, for the clusters of which the diameters are at most ε, where $\varepsilon < 1/2\sqrt{n}$ is a small positive number. Then*

$$\nu_k(H) \leqslant q^2 \sum_{j=1}^{k} \lambda_j, \tag{3.8}$$

where $q = 1 + \sqrt{n}\varepsilon/(1 - \sqrt{n}\varepsilon)$.

Comparing the results of Theorems 3.5 and 3.8, under the constraints of Theorem 3.8 we obtain that

$$\sum_{j=1}^{k} \lambda_j \leqslant \nu_k(H) \leqslant q^2 \sum_{j=1}^{k} \lambda_j, \quad \text{where } 1 < q < 2.$$

Provided ε is less than $1/2\sqrt{n}$, then q is at most 2 and the combinatorial and analytical measures of H, $\nu_k(H)$ and $\sum_{j=1}^{k} \lambda_j$ differ at most by a factor 4. Under the assumptions of Theorem 3.8

$$\nu_{k+1} - \nu_k \geqslant \sum_{j=1}^{k+1} \lambda_j - q^2 \sum_{j=1}^{k} \lambda_j$$

also holds. Therefore the larger the gap in the spectrum between λ_k and λ_{k+1} is and the better the representatives of the vertices in an optimal k-dimensional Euclidean representation are separated, the bigger the difference between v_{k+1} and v_k is.

Tusnády and Bolla have found an upper bound for the sum of the inner variances of the representatives of the vertices in terms of the gap in the spectrum (see [3]).

The proofs of the theorems are contained in the next section.

4. Proofs of theorems

Lemma 4.1. *Let the real number $\Delta>0$ and the integers n and k $(1\leqslant k<n)$ be fixed. Then any n-tuples of points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in R^k$ satisfying the property $\mathcal{T}$ can be colored with $k+1$ different colors in such a way that the minimal distance between points of different colors is at least Δ, where the property $\mathcal{T}$ is the following: projecting the points onto any line of R^k, on this line there are at least two consecutive points, whose distance is at least Δ.*

Proof. By induction, for $k=1$ the statement is straightforward. Let us suppose that for $k-1$ the lemma is proved. For k: according to the property $\mathcal{T}$ the points can be colored with 2 different colors in the requested way. On the one hand let us choose one–one points from both color-classes. Let us connect them and project the k-dimensional points onto the $(k-1)$-dimensional subspace orthogonal to this line.

On the other hand let us consider the Δ-level graph of the points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ (the ones, whose distance is less than Δ, are connected). We have to show that this graph consists of $k+1$ connected components. Since the points can be colored with two colors, we have at least two components. After the above projection these two components will be connected by an edge. Therefore the number of connected components is decreased by one with this projection. But according to the inductive supposition even after the projection there will be k components, which form connected components in R^k too. Therefore the number of connected components is exactly $k+1$. □

Lemma 4.2. *Let $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\dots,\boldsymbol{x}_n\in R^k$ be arbitrary points subject to the constraints $\sum_{j=1}^n \boldsymbol{x}_j=0$ and $\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}}=\boldsymbol{I}_k$. Then they can be colored with $k+1$ different colors in such a way that the minimal distance between points of different colors is at least $d_n=2\sqrt{3}/\sqrt{n(n^2-1)}$.*

Proof. Due to the constraints the property $\mathcal{T}$ of Lemma 4.1 is satisfied with $\Delta=d_n$. Indeed, projecting the point $\boldsymbol{x}_j$ onto the line with direction vector $\boldsymbol{f}(\|\boldsymbol{f}\|=1)$ let us denote its projected copy by $x_j:=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{x}_j$, for which

$$\sum_{j=1}^{n} x_j=0$$

and

$$\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_j \boldsymbol{x}_j^{\mathrm{T}})\boldsymbol{f}^{\mathrm{T}} = \|\boldsymbol{f}\|^2 = 1.$$

Let us denote by $x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$ the ordered set of the one-dimensional points $x_1, \ldots, x_n$ and let δ be defined by

$$\delta := \max_{1 \leqslant i < n} (x_{i+1}^* - x_i^*).$$

Then

$$2n = 2n \sum_{i=1}^{n} x_i^{*2} = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i^* - x_j^*)^2$$

$$= 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_j^* - x_i^*)^2 \leqslant 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta^2 (j-i)^2 = \frac{\delta^2}{6} n^2 (n^2 - 1).$$

Hence $\delta^2 \geqslant d_n^2$, which implies the choice of Δ. □

Proof of Theorem 3.5. *Upper bound*: Let $(V_1^*, \ldots, V_k^*)$ be a k-partition giving the minimal weighted k-cut of H, where $|V_i^*| = n_i$ $(i = 1, \ldots, n)$. Let us define the following k-dimensional Euclidean representation of the vertices:

$$x_j(i) = \begin{cases} 1/\sqrt{n_i^*} & \text{if } v_j \in V_i^*, \\ 0 & \text{otherwise,} \end{cases}$$

where $x_j(i)$ denotes the ith coordinate of the k-dimensional representative $\boldsymbol{x}_j$ of the vertex v_j. With this choice $\sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^{\mathrm{T}} = \boldsymbol{I}_k$ also holds.

By formula (2.5) the variance of the edge e in this representation is

$$L(e, \boldsymbol{X}) = \frac{1}{|e|} \sum_{1 \leqslant i < j \leqslant k} \left(\frac{1}{n_i^*} + \frac{1}{n_j^*} \right) a_i^*(e)\, a_j^*(e), \quad e \in E,$$

where $a_i^*(e) = |e \cap V_i^*|$ $(i = 1, \ldots, k)$. Since $\sum_{e \in E} L(e, \boldsymbol{X}) = u(V_1^*, \ldots, V_k^*) = v_k(H)$ and $\sum_{j=1}^{k} \lambda_j = L(\boldsymbol{X}^*)$ – where the $k \times n$ matrix $\boldsymbol{X}^*$ is an optimal k-dimensional Euclidean representation of H – as a simple consequence of the Representation Theorem we have that

$$\sum_{j=1}^{k} \lambda_j \leqslant v_k(H).$$

Lower bound: (i) The cost is monotone: $e' \subseteq e$ implies $L(e', \boldsymbol{X}) \leqslant L(e, \boldsymbol{X})$ in any Euclidean representation $\boldsymbol{X}$, since by the Steiner formula:

$$L(e', X) = \sum_{v_j \in e'} \left\| x_j - \frac{\sum_{v_i \in e'} x_i}{|e'|} \right\|^2$$

$$= \sum_{v_j \in e'} \left\| x_j - \frac{\sum_{v_i \in e'} x_i}{|e'|} \right\|^2 - |e'| \left\| \frac{\sum_{v_i \in e} x_i}{|e|} - \frac{\sum_{v_i \in e'} x_i}{|e'|} \right\|^2$$

$$\leqslant \sum_{v_j \in e'} \left\| x_j - \frac{\sum_{v_i \in e} x_i}{|e|} \right\|^2 \leqslant \sum_{v_j \in e} \left\| x_j - \frac{\sum_{v_i \in e} x_i}{|e|} \right\|^2 = L(e, X).$$

(ii) If $e = \{v_i, v_j\}$, then by formula (2.5) $L(e, X) = \frac{1}{2}\|x_i - x_j\|^2$ in any Euclidean representation X.

(iii) Let us take the vectors $x_1^*, \ldots, x_n^* \in R^k$ giving an optimal k-dimensional Euclidean representation. Let $z_j^* \in R^{k-1}$ be the vector obtained by discarding the first coordinate of x_j ($j = 1, \ldots, n$). As the discarded first coordinates constitute a vector, which is a multiple of the vector $e = (1, \ldots, 1)^T \in R^n$ (corresponding to the eigenvalue $\lambda_1 = 0$), and the other eigenvectors are orthogonal to it, $\sum_{j=1}^n z_j^* = 0$ and from the side conditions for x_j^*s the relation $\sum_{j=1}^n z_j^* z_j^{*T} = I_{k-1}$ also follows. Then according to Lemma 4.2 z_j^*s can be colored with k different colors in such a way that the minimal distance between the vectors of different colors is at least d_n. This also holds for x_j^*s. Let us denote by $(V_1, \ldots, V_k)$ the k-partition formed by this coloring, and let us choose an edge from its cut-set $H(V_1, \ldots, V_k)$. Such an edge e contains an edge $e' = \{v_i, v_j\}$, where the vertices v_i and v_j are of different colors. Thus by applying the results of (i) and (ii) for the variance of the edge e in the Euclidean representation X^* we obtain that

$$L(e, X^*) \geqslant L(e', X^*) = \frac{\|x_i^* - x_j^*\|^2}{2} \geqslant \frac{d_n^2}{2},$$

whence $c_n = d_n^2/2 = 6/(n(n^2-1))$. But X^* was an optimal k-dimensional Euclidean representation, therefore by the Representation Theorem,

$$\sum_{j=1}^k \lambda_j = \sum_{e \in E} L(e, X^*) \geqslant \sum_{e \in H(V_1, \ldots, V_k)} L(e, X^*)$$

$$\geqslant c_n |H(V_1, \ldots, V_k)| \geqslant c_n \theta_k(H). \qquad \square$$

Lemma 4.3. *Let $\mu(B)$ denote the measure of irreducibility of the Laplacian B defined by* $\min_{\emptyset \neq \mathcal{M} \subset \mathcal{N}} \sum_{i \in \mathcal{M}, j \notin \mathcal{M}} |b_{ij}|$, *where $\mathcal{N} = \{1, 2, \ldots, n\}$. Then $\mu_2(H) = \mu(B)$.*

Proof. Let $\mathcal{M} \subset \mathcal{N}$ be a fixed non-empty subset, $V_1 := \{v_j : j \in \mathcal{M}\}$ and $V_2 := V \setminus V_1$. Then by an easy counting argument,

$$\sum_{\substack{i \in \mathcal{M} \\ j \notin \mathcal{M}}} |b_{ij}| = \sum_{\substack{i \in \mathcal{M} \\ j \notin \mathcal{M}}} \sum_{e \in E} \mathcal{I}(v_i \in e)\, \mathcal{I}(v_j \in e) \frac{1}{|e|} = \sum_{\substack{e \in E \\ e \cap V_1 \neq \emptyset \\ e \cap V_2 \neq \emptyset}} \frac{a_1(e)\, a_2(e)}{a_1(e) + a_2(e)} = v(V_1, V_2).$$

By taking the minima of the two sides through all pairs $(V_1, V_2) \in \mathscr{P}_2$, the equality $\mu_2(H) = \mu(\boldsymbol{B})$ is obtained. □

Proof of Theorem 3.6. Let $\boldsymbol{B}$ be the Laplacian of H and $\boldsymbol{G} = \boldsymbol{I}_n - (1/s_{\max})\boldsymbol{B}$. As all row sums of $\boldsymbol{B}$ are 0 and $\boldsymbol{B}$ is symmetric, $\boldsymbol{G}$ is doubly stochastic. Because of $b_{jj} \leqslant s_j \leqslant s_{\max}$, all of the entries in $\boldsymbol{G}$ are nonnegative. Then by functional calculus the largest eigenvalue of $\boldsymbol{G}$ is 1, while the second largest one is $1 - (1/s_{\max})\lambda_2$. Let $\mu(\boldsymbol{B})$ denote the measure of irreducibility of H. As $\mu(\boldsymbol{B})$ only depends on the nondiagonal entries of $\boldsymbol{B}$, $\mu(\boldsymbol{G}) = (1/s_{\max})\mu(\boldsymbol{B}) = (1/s_{\max})\mu_2(H)$, where the last equality follows from Lemma 4.3. Applying a theorem of Fiedler [11], for the second largest eigenvalue of $\boldsymbol{G}$ the inequality

$$\frac{\lambda_2}{s_{\max}} \geqslant \phi_n\left(\frac{\mu_2(H)}{s_{\max}}\right)$$

holds, where

$$\phi_n(x) = \begin{cases} 2(1-\cos \pi/n)\, x & \text{if } 0 \leqslant x \leqslant \frac{1}{2}, \\ 1 - 2(1-x)\cos \pi/n - (2x-1)\cos 2\pi/n & \text{if } \frac{1}{2} < x \leqslant 1. \end{cases}$$

The function $\phi_n(x)$ is defined on the interval $[0, 1]$. But $(1/s_{\max})\mu_2(H)$ is an element of this interval, since $\mu_2(H)$ is nonnegative and for any fixed j

$$\mu_2(H) = \mu(\boldsymbol{B}) \leqslant \sum_{i \neq j} |b_{ij}| = b_{jj} \leqslant s_j,$$

consequently $\mu_2(H) \leqslant \min_j s_j \leqslant s_{\max}$. If $\mu_2(H) \leqslant \frac{1}{2} s_{\max}$, the lower bound for λ_2 depends only on $\mu_2(H)$. E.g. in case of $\min_j s_j \leqslant \frac{1}{2} s_{\max}$. □

Proof of Theorem 3.8. Let ε be less than $1/2\sqrt{n}$ and let $P_k = (V_1, \ldots, V_n)$ be a well-separated k-partition of V in the optimal k-dimensional Euclidean representation $X^* = (x_1^*, \ldots, x_n^*)$ such that the diameters of the clusters are less than ε. Because of the continuity of the outer product we can suppose that under the conditions of Theorem 3.8 there are k 'centers' of the clusters such that the representative x_j^* is allocated within one of the k-dimensional spheres with radius ε around these 'centers', and denoting by $y(x_j^*)$ the nearest 'center' to x_j^*, the relation $\sum_{j=1}^n y(x_j^*) y^{\mathrm{T}}(x_j^*) = \boldsymbol{I}_k$ also holds. As amongst the 'centers' there are exactly k different ones (let them denote by $y_1, \ldots, y_k$) $\sum_{i=1}^k n_i y_i y_i^{\mathrm{T}} = \boldsymbol{I}_k$ holds, where $n_i = |V_i|$, $\sum_{i=1}^k n_i = n$, furthermore this condition determines the y_is uniquely (apart from a $k \times k$ rotation):

$$y_i(l) = \begin{cases} \dfrac{1}{\sqrt{n_i}} & \text{if } i = l, \\ 0 & \text{otherwise.} \end{cases}$$

Here $y_i(l)$ denotes the lth coordinate of the vector y_i. Let $L(e, X^*)$ denote the variance of the edge e in an optimal Euclidean representation X^* (the representative of v_j being x_j^*) and $L(e, Y(X^*))$ be the variance of it in the Euclidean representation, where the representative of v_j is $y(x_j^*)$. Then using the equation $\sum_{e\in E} L(e, Y(X^*))=u(P_k)$, and applying twice the formula (2.5) we arrive at the following argument:

$$v_k(H)\leqslant \sum_{e\in E} L(e, Y(X^*))= \sum_{e\in H(P_k)} L(e, Y(X^*))$$

$$= \sum_{e\in H(P_k)} \frac{1}{|e|} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|y(x_i^*)-y(x_j^*)\|^2$$

$$\leqslant q^2 \sum_{e\in H(P_k)} \frac{1}{|e|} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|x_i^*-x_j^*\|^2$$

$$= q^2 \sum_{e\in H(P_k)} L(e, X^*)\leqslant q^2 \sum_{e\in E} L(e, X^*)= q^2 \sum_{j=1}^{k} \lambda_j,$$

since

$$\|y(x_i^*)-y(x_j^*)\|\leqslant q\|x_i^*-x_j^*\|,$$

where the constant q is determined from the fact that the smallest distance between the 'centers' y_is is

$$\delta= \min_{\sum_{l=1}^{k} n_l=n} \sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\geqslant \frac{2}{\sqrt{n}}$$

and x_i^* and x_j^* are within the sphere with radius ε around the 'centers' $y(x_i^*)$ and $y(x_j^*)$, respectively. Therefore

$$q=\frac{\delta}{\delta-2\varepsilon}=1+\frac{2\varepsilon}{\delta-2\varepsilon}\leqslant 1+\frac{\varepsilon\sqrt{n}}{1-\varepsilon\sqrt{n}}. \qquad \square$$

5. A heuristic classification algorithm based on Euclidean representations

Let $v_1, v_2, \ldots, v_n$ be binary random variables taking the values 0–1 and $e_1, e_2, \ldots, e_m$ be a sample for them ($n\ll m$). They form a hypergraph $H=(V, E)$ with vertex-set $V=\{v_1, v_2, \ldots, v_n\}$ and edge-set $E=\{e_1, e_2, \ldots, e_m\}$, where $\mathcal{I}(v\in e)=v(e)$, $v(e)$ being the observed value of the variable v on the object e. (When v represents some property, $v(e)=1$ means the presence, while $v(e)=0$ the absence of this property on the object e.)

Let $E'\subset E$ be a *sub-sample*. The sub-hypergraph $H'=(V, E')$ is called the *hypergraph of the edge-cluster* E'. Let us denote by $0=\lambda_1(H')\leqslant\lambda_2(H')\leqslant\cdots\leqslant\lambda_n(H')$ the spectrum of H', while the $n\times n$ matrix $X'(H')$ contains a whole system of pairwise orthonormal eigenvectors of the Laplacian of H'. According to the Representation

Theorem of Section 2, for any integer d $(1<d\leqslant n)$ the $d\times n$ matrix $X_d^*(H')$ – obtained from $X^*(H')$ by retaining the eigenvectors corresponding to $\lambda_1(H')$, $\lambda_2(H'),\dots,\lambda_d(H')$ – defines an optimal d-dimensional Euclidean representation of H'. Furthermore, the sum of the variances of the edges of E' in this representation is minimal, and it is equal to

$$L(X_d^*(H'))=\sum_{e\in E'} L(e,X_d^*(H'))=\sum_{j=1}^{d}\lambda_j(H').$$

Put $K(H'):=\min_{d=1}^{n}[c2^{n-d}+L(X_d^*(H'))]$, where $c>0$ is a constant (chosen previously according to the size of problem). The dimension d^* giving the minimum is called the *dimension of the edge-cluster* E'.

Let $\mathscr{S}$ denote the set of all partitions of E into nonempty disjoint sub-samples. Our purpose is to find a partition $S\in\mathscr{S}$ consisting of sub-samples E_i for which the objective function $K=\sum_i K(H_i)$ is minimal, where $H_i=(V,E_i)$ is the hypergraph of the edge-cluster E_i.

Now let k be a fixed integer $(1<k\leqslant n)$. We shall define a numerical algorithm converging to a local minimum of the objective function, when the minimization takes place over the set of all k-partitions $\mathscr{S}_k$. Let $(E_1,\dots,E_k)\in\mathscr{S}_k$ be a k-partition of the edge-set of H. Applying the previous notations for the hypergraphs $H_i=(V,E_i)$ $(i=1,\dots,k)$ the following cost function is constructed: $Q=\sum_{i=1}^{k} Q_{d_i}(H_i)$, where

$$Q_{d_i}(H_i):=c2^{n-d_i}+L(X_{d_i}^*(H_i))\quad (i=1,\dots,k).$$

To minimize the cost function Q — with respect to k-partitions of the edges and dimensions of the edge-clusters — the following iteration is introduced. First let us choose k disjoint clusters $E_1,\dots,E_k$ of the objects (e.g. by the k-means method, see in [19]).

(i) Fixing the clusters $E_1,\dots,E_k$: the spectra and optimal Euclidean representations of the sub-hypergraphs of the edge-clusters are calculated.

(ii) The function $Q_{d_i}(H_i)$ is minimized with respect to the dimension d_i $(1<d_i\leqslant n)$ for each i separately. A unique d_i^* giving the ith minimum always exists. As for that d_i^*

$$Q_{d_i^*}(H_i)=c2^{n-d_i^*}+\sum_{j=1}^{d_i^*}\lambda_j(H_i)\quad (i=1,\dots,k)$$

holds, in this step the cost function Q is decreased. Until this moment the minimization took place within the clusters. In the next step the objects are relocated between the clusters.

(iii) Fixing the d_i^*-dimensional optimal Euclidean representations $X_{d_i^*}^*(H_i)$: an object e is replaced into the cluster E_i, for which $L(e,X_{d_i^*}^*(H_i))$ is minimal. If the minimum is attained for more than one i, let us replace e into the cluster E_i with the smallest index i. In this step Q is also decreased. In this way a new disjoint classification $E_1^*,\dots,E_k^*$ of the objects is obtained. From now on we go back to step (i) with starting classification $E_1^*,\dots,E_k^*$, etc.

As the cost function Q is in each step decreased and in steps (ii) and (iii) discrete minimizations are performed the algorithm must converge to a local minimum of Q in finite steps. It is easy to see that for fixed k the k-partition, to which the iteration converges gives a local minimum of the objective function K too. As a new step of the iteration, a minimization with respect to k could be introduced, but it would be very time-demanding. (The optimal value of k also depends on the constant c.)

During the iteration some edge-clusters may become empty. Usually the hypergraph $H_i=(V, E_i)$ contains isolated vertices (this results in additional zero eigenvalues). Let us denote by V_i the set of the nonisolated vertices of H_i. Provided H has no isolated vertices, then $\bigcup_{i=1}^{k} V_i = V$ holds but $V_1, \ldots, V_k$ are in general not disjoint subsets of the vertices. V_i is called the *characteristic property-association* of the sub-sample E_i.

6. Some remarks concerning spectra of hypergraphs

Finally, we introduce some simple propositions on spectra of hypergraphs and on Euclidean representations of some special hypergraphs (sometimes without proofs). Unless otherwise stated, the propositions refer to the spectral characteristics of the hypergraph $H=(V, E)$ with $|V|=n$ and $|E|=m$.

Assertion 6.1.

$$\sum_{j=1}^{n} \lambda_j = \operatorname{tr} \boldsymbol{B} = \sum_{e \in E} (|e|-1) = \sum_{e \in E} |e| - m. \tag{6.1}$$

Proposition 6.2.

$$\lambda_2 \leqslant \frac{n}{n-1} \max_j b_{jj} \leqslant \frac{\sum_{e \in E} |e| - m}{n-1}$$

and

$$\lambda_n \leqslant \max_j s_j. \tag{6.2}$$

Corollary 6.3.

$$\lambda_2 \leqslant \frac{n}{n-1} \max_j s'_j \left(1 - \frac{1}{t_j}\right), \tag{6.3}$$

where $s'_j = \#\{e \in E: v_j \in e, |e|>1\}$ *and* $t_j = \max_{v_j \in e} |e|$.

Assertion 6.4. *If* $H_i=(V, E_i)$ $(i=1, \ldots, k)$ *are edge-disjoint hypergraphs, and* $H=(V, E)$, *where* $E=\bigcup_{i=1}^{k} E_i$, $E_i \cap E_j = \emptyset$ $(i \neq j)$, *then for their Laplacians the relation*

$$\boldsymbol{B}(H) = \sum_{i=1}^{k} \boldsymbol{B}(H_i) \tag{6.4}$$

holds.

Proposition 6.5. *Let $H=(V,E)$ be a hypergraph, $E=E_1\cup E_2$, $E_1\cap E_2=\emptyset$, $H_i=(V,E_i)$, $i=1,2$. Then*

$$\sum_{j=1}^{k}\lambda_j\geqslant\sum_{j=1}^{k}\lambda_j^{(1)}+\sum_{j=1}^{k}\lambda_j^{(2)}\quad(1\leqslant k\leqslant n),\tag{6.5}$$

where $\lambda_j^{(i)}$ denotes the jth eigenvalue of H_i in increasing order $(i=1,2)$.

Proof. Let $L(e,\boldsymbol{X}^*)$ denote the variance of the edge e in the optimal k-dimensional Euclidean representation $\boldsymbol{X}^*$ of H. Then according to the Representation Theorem,

$$\sum_{j=1}^{k}\lambda_j=\sum_{e\in E}L(e,\boldsymbol{X}^*)=\sum_{e\in E_1}L(e,\boldsymbol{X}^*)+\sum_{e\in E_2}L(e,\boldsymbol{X}^*)$$

$$\geqslant\sum_{j=1}^{k}\lambda_j^{(1)}+\sum_{j=1}^{k}\lambda_j^{(2)}.\qquad\square$$

Proposition 6.6. *With the notations of the previous proposition:*

$$\lambda_{j-r_i}\leqslant\lambda_j^{(i)}\leqslant\lambda_j\quad(j=1,\ldots,n),\tag{6.6}$$

where $r_i=\operatorname{rank}\boldsymbol{B}_i$, $\boldsymbol{B}_i$ being the Laplacian of H_i $(i=1,2)$ and $\lambda_l=0$, if $l<1$.

Proof. According to the Poincaré separation theorem (see [21, Theorem 2.1]): if $\boldsymbol{C}$ is an $n\times n$ symmetric, positive semidefinite matrix of rank r and $\boldsymbol{D}$ is symmetric, positive semidefinite, for which $\operatorname{rank}\boldsymbol{D}\leqslant k$ $(\leqslant r)$, then

$$\lambda_i(\boldsymbol{C})\geqslant\lambda_i(\boldsymbol{C}-\boldsymbol{D})\geqslant\lambda_{i-k}(\boldsymbol{C})\quad(i=1,\ldots,n).$$

As $\boldsymbol{B}=\boldsymbol{B}_1+\boldsymbol{B}_2$, where $\boldsymbol{B}$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ are $n\times n$ symmetric, positive semidefinite matrices respectively, the statement follows. $\square$

Corollary 6.7. *Let $H=(V,E)$ be a hypergraph, and let e be an edge of it with $|e|=z$. Then by setting $E':=E\setminus\{e\}$, $H':=(V,E')$ and denoting by λ' the eigenvalues of H', the relations*

$$\sum_{j=1}^{k-z+1}\lambda_j\leqslant\sum_{j=1}^{k}\lambda_j'\leqslant\sum_{j=1}^{k}\lambda_j\quad(1\leqslant k\leqslant n-1)\tag{6.7}$$

hold, where the first sum be zero, if $k<z$.

Proof. Let $\boldsymbol{B}_2$ denote the Laplacian of the hypergraph $H_2:=(V,\{e\})$. Then the rank r_2 of the matrix $\boldsymbol{B}_2$ is equal to $z-1$, its eigenvalues being

$$\lambda_1^{(2)}=\cdots=\lambda_{n-z+1}^{(2)}=0,\qquad\lambda_{n-z+2}^{(2)}=\cdots=\lambda_n^{(2)}=1.$$

The second inequality follows immediately from the second inequality of (6.6), while the first inequality from the first one. The lower bound is 0, if $k<z$. Otherwise:

$$\sum_{j=1}^{k}\lambda'_j \geqslant \sum_{j=1}^{k}\lambda_{j-(z-1)} \geqslant \sum_{i=1}^{k-z+1}\lambda_i. \qquad \square$$

We remark that the result of Corollary 6.7 remains valid, if $H'=(V,E')$, where the rank of the Laplacian of the hypergraph $(V,E\setminus E')$ is $z-1$.

Corollary 6.8. *For $z=2$, by the successive and alternating application of the two sides of* (6.6) *we obtain that*

$$0\leqslant\lambda'_2\leqslant\lambda_2\leqslant\lambda'_2+\lambda'_3\leqslant\lambda_2+\lambda_3\leqslant\lambda'_2+\lambda'_3+\lambda'_4\leqslant\lambda_2+\lambda_3+\lambda_4\leqslant\cdots. \tag{6.8}$$

Example 6.9. Let C_n denote the *complete hypergraph* with n vertices and without loops (it has 2^n-n-1 hyperedges). Its spectrum consists of one zero and the number $(n2^{n-1}-2^n+1)/(n-1)$ with multiplicity $n-1$. Any $n-1$ pairwise orthogonal vectors within the subspace orthogonal to the vector $e\in R^n$ are eigenvectors belonging to the multiple eigenvalue.

Example 6.10. The smallest positive eigenvalue of the *path graph* P_n having $n=2l+1$ vertices is $1-\cos\pi/n$. Labelling the vertices as $v_{-l},\ldots,v_0,\ldots,v_l$, the second coordinates of their representatives in an optimal 2-dimensional Euclidean representation of P_n are

$$x_j=\frac{\sqrt{2}}{\sqrt{n}}\sin\left(j\frac{\pi}{n}\right),\quad j=-l,\ldots,0,\ldots,l \tag{6.9}$$

while the first coordinates are all equal to $1/\sqrt{n}$.

Example 6.11. Let S_d denote the *star graph* with $n=d+1$ vertices. The smallest positive eigenvalue of S_d is $\frac{1}{2}$ with multiplicity $d-1$. An optimal d-dimensional Euclidean representation of S_d is a d-simplex in the $(d-1)$-dimensional subspace of R^d orthogonal to the vector $e\in R^d$. The center of gravity of the simplex is in the origin. The representatives of the vertices of valency 1 are the vertices, while the representative of the vertex of valency d is the center of gravity of the simplex.

Example 6.12. Let $G_{d,l}$ denote the subdivision graph of S_d, where each of the edges of S_d is divided into l parts. We call $G_{d,l}$ *spider* with d feet and l sections. The number of its vertices is $n=dl+1$. The smallest positive eigenvalue of $G_{d,l}$ is of multiplicity $d-1$ and it is equal to $1-\cos(\pi/(2l+1))$. An optimal d-dimensional Euclidean representation of the spider $G_{d,l}$ is obtained from those of S_d and P_{2l+1}, where the feet of the spider are divided according to the sine rhythm of (6.9).

Example 6.13. Let $L_{d,l}$ denote the d-dimensional *lattice* whose vertices are all d-tuples of numbers $-l, \dots, 0, \dots, l$, where two d-tuples are adjacent if and only if they differ in exactly one coordinate. The number of its vertices is $n=(2l+1)^d$. The smallest positive eigenvalue of $L_{d,l}$ is $1-\cos(\pi/(2l+1))$ with multiplicity d. An optimal $(d+1)$-dimensional Euclidean representation of $L_{d,l}$ is realized in the d-dimensional subspace of R^{d+1} orthogonal to the $e \in R^{d+1}$ vector. It is a d-dimensional lattice, its center of gravity being in the origin, and the distances between the representatives of adjacent vertices follow the sine rhythm of (6.9).

Example 6.14. Let $K_{n_1,\dots,n_k}$ be the *complete k-partite graph*, where $n=\sum_{i=1}^{k} n_i$ (n being the number of vertices). Let $(V_1, \dots, V_k)$ denote the disjoint non-empty independent sets of the vertices, where $n_i=|V_i|$ $(i=1, \dots, k)$. The spectrum of $K_{n_1,\dots,n_k}$ contains a single 0, the numbers $\frac{1}{2}(n-n_i)$ with multiplicity n_i-1 $(i=1, \dots, k)$ and $k-1$ numbers equal to $\frac{1}{2}n$. If we regard the $(k-1)$-dimensional Euclidean representation corresponding to the largest eigenvalue $\frac{1}{2}n$, the representatives of the vertices in this representation constitute k different points in the $(k-1)$-dimensional Euclidean space, where the representatives of vertices of the same color coincide.

Proof. Assume that the labelling of the vertices is such that V_1 contains the first n_1 vertices, V_2 the next n_2 ones, etc... As the graph is connected, 0 is a single eigenvalue. Let us regard the Laplacian $\boldsymbol{B}$ of $K_{n_1,\dots,n_k}$ multiplied by 2. It contains k diagonal blocks in its main diagonal. The ith block is of size $n_i \times n_i$ and it is diagonal with positive entries $n-n_i$ $(i=1, \dots, k)$. Outsides of these diagonal blocks all entries are equal to -1. For $\boldsymbol{x}=(x_1, \dots, x_n)^T \in R$ the equation $2\boldsymbol{Bx}=n\boldsymbol{x}$ results in the system of equations

$$(n-n_i)\,x_j - \sum_{v_l \notin V_i} x_l = n x_j \quad (i=1, \dots, k),$$

where the ith equation holds for any j such that $v_j \in V_i$. Therefore the system of equations

$$n_i x_j + \sum_{v_l \notin V_i} x_l = 0 \quad (i=1, \dots, k \text{ and } v_j \in V_i)$$

has to be solved for the coordinates of the vector $\boldsymbol{x}$, as unknowns. Any solution has the following form:

$$x_j = y_i, \quad \text{if } v_j \in V_i, \tag{6.10}$$

where the numbers $y_1, \dots, y_k$ satisfy the equation $\sum_{i=1}^{k} n_i y_i = 0$, which assures the orthogonality to the vector $\boldsymbol{e}=(1, \dots, 1) \in R^n$. As such vectors $\boldsymbol{x}$ constitute a $(k-1)$-dimensional subspace within R^n, the eigenspace belonging to the eigenvalue $\frac{1}{2}n$ is of dimension $k-1$, therefore this eigenvalue is of multiplicity $k-1$. Since any $(k-1)$-dimensional Euclidean representation formed by a $(k-1)$-tuple of pairwise orthonormal eigenvectors belonging to the largest eigenvalue $\frac{1}{2}n$ is of the form (6.10), the

representatives of the vertices in any of these representations constitute exactly k different points in the $(k-1)$-dimensional Euclidean space such that the representatives of vertices of the same color coincide.

For any i $(i=1,\dots,k)$ the matrix equation $2\boldsymbol{B}\boldsymbol{x}=(n-n_i)\boldsymbol{x}$ implies

$$(n-n_i)x_j-\sum_{v_l\notin V_i} x_l=(n-n_i)x_j \quad (v_j\in V_i),$$

whence for any solution $\boldsymbol{x}$, $x_j=0$ holds, if $v_j\notin V_i$ and $\sum_{v_l\notin V_i} x_l=0$ because of the orthogonality to the vector $\boldsymbol{e}\in R^n$. These conditions determine an (n_i-1)-dimensional subspace of R^n, as the eigenspace belonging to the eigenvalue $\frac{1}{2}(n-n_i)$. □

In this way we can characterize the complete k-partite graph on the basis of its optimal $(k-1)$-dimensional Euclidean representation belonging to the largest eigenvalue with multiplicity $k-1$. But how can we recognize a k-partition, the members of which are independent sets of the vertices (i.e. k-colorable hypergraphs), we do not know exactly. Recently it has turned out that these spectral techniques are not always capable for the recognition of the chromatic number.

Analogously to the derivation of the Representation Theorem the maximum of the quadratic form $L(\boldsymbol{X})=\operatorname{tr}\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^{\mathrm{T}}$ on $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{I}_k$ is the sum of the k largest eigenvalues of the hypergraph in question and the $k\times n$ matrix $\boldsymbol{X}$ giving the maximum contains the corresponding eigenvectors in its rows. In this kind of representation the sum of the variances of the edges is maximized. As a k-colorable graph has no edges within the subsets of color-partition $(V_1,\dots,V_k)$, the $(k-1)$-dimensional representatives of vertices of the same color tend to be near to each other, while the representatives of vertices of the multi-colored edges tend to be far away. Consequently, the color-partition frequently results in well-separated clusters of the representatives of vertices in this representation.

Acknowledgement

I am very grateful to Gábor Tusnády for helpful discussions and to János Komlós for directing my interest to this topic. I also wish to thank the referee of the paper for suggesting many improvements to the original text.

References

[1] N. Alon, Eigenvalues and expanders, Combinatorica 6 (2) (1986) 83–96.
[2] N.L. Biggs, Algebraic Graph Theory (Cambridge Univ. Press, Cambridge, 1974).
[3] M. Bolla and G. Tusnády, Spectra and colourings of weighted graphs, Mathematical Institute of the Hungarian Academy of Sciences, Preprint, No. 52, 1990.
[4] D.M. Cvetković, M. Doob and H. Sachs, Spectra of Graphs (Academic Press, New York, 1979).
[5] M. Doob, Eigenvalues of a graph and its imbeddings. J. Combin. Theory Ser. B 17 (1974) 244–248.

[6] M. Doob and D.M. Cvetković, On spectral characterizations and embeddings of graphs, Linear Algebra Appl. 27 (1979) 17–26.
[7] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Cybernetics 3 (3) (1973) 32–57.
[8] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, Cybernetics 4 (1) (1974) 95–104.
[9] J.C. Dunn, Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems, Cybernetics, 4 (2) (1974) 1–23.
[10] P. Erdös and J. Spencer, Probabilistic Methods in Combinatorics (Akadémiai Kiadó, Budapest, 1974).
[11] M. Fiedler, Bounds for eigenvalues of doubly stochastic matrices, Linear Algebra Appl. 5 (1972) 299–310.
[12] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak. Math. J. 23 (1973) 298–305.
[13] M. Fiedler, A property of eigenvectors of non-negative symmetric matrices and its applications to graph theory, Czechoslovak. Math. J. 25 (1975) 619–633.
[14] A.J. Hoffman, The change in the least eigenvalue of the adjacency matrix of a graph under imbedding, SIAM J. Appl. Math. 17 (1969) 664–671.
[15] A.J. Hoffman, On eigenvalues and colorings of graphs, in: B. Harris, ed., Graph Theory and Its Applications (Academic Press, New York, 1970) 79–91.
[16] A.J. Hoffman, Eigenvalues and partitionings of the edges of a graph, Linear Algebra Appl. 5 (1972) 137–146.
[17] F. Juhász and K. Mályusz, Problems of cluster analysis from the viewpoint of numerical analysis, Proc. Conf. Numerical Methods, Keszthely (1977).
[18] L. Lovász, Combinatorial Problems and Exercises (Akadémiai Kiadó–North Holland, Budapest–Amsterdam, 1979).
[19] J. Mac Queen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Math. Statist. Prob. 1 (1967) 281–297.
[20] A. Mowshowitz, The characteristic polynomial of a graph, J. Combin. Theory Ser. B 12 (1972) 177–193.
[21] C.R. Rao, Separation theorems for singular values of matrices and their applications in multivariate analysis, J. Multivariate Anal. 9 (1979) 362–377.
[22] E. Shamir and J. Spencer, Sharp concentration of the chromatic number on random graphs $G_{n,p}$, Combinatorica 7 (1) (1987) 121–129.