HIPERGRÁFOK ÖSSZEFÜGGŐSÉGÉNEK VIZSGÁLATA A SPEKTRUMON KERESZTÜL

Bolla Marianna[1] és Tusnády Gábor[2]

Hipergráfok klaszteresedési tulajdonságait vizsgáljuk lineáris algebrai segédeszközökkel. Általánosítjuk a Laplace-mátrix fogalmát és ennek sajátértékeiből vonunk le következtetéseket a hipergráf összefüggőségére és a klaszterek számára vonatkozóan. Maguknak a klasztereknek a megkonstruálásához a sajátvektorokat, ill. a hipergráf általuk definiált euklideszi reprezentációját használjuk. Egy iterációs eljárást is ismertetünk.

Bevezetés

A kromatikus szám mellett az összefüggőség a gráfok másik gyakran vizsgált tulajdonsága. Mindkettő a gráf csúcsainak particionálását (felosztását, diszjunkt lefedőrendszerét) jelenti, csak ellentétes szempontokból.

Ha egy gráf kromatikus száma k, akkor a csúcsoknak létezik olyan k-partíciója (a lefedőrendszer k elemű), hogy élek csak a partíció különböző tagjai közt futnak (nincsenek "belső élek"), és k-nál kisebb elemű, ilyen tulajdonságú partíció nem létezik. Ha egy partíció elemeinek különböző színeket feleltetünk meg, akkor a partíció a csúcsok un. k-színezésének is tekinthető. Bebizonyított tény, hogy síkgráfok kromatikus száma legfeljebb 4, így bármely térkép kiszínezhető 4 különböző szín felhasználásával.

A fenti követelménnyel szemben, ha egy gráf k összefüggő komponensből áll, akkor csúcsainak létezik olyan k-partíciója, amelyben egyáltalán nincsenek köztes (különböző színű csúcsokat összekötő) élek, azaz élek csak a partíció elemein belül futnak. Ez a tulajdonság persze nagyon könnyen felismerhető, hiszen ilyenkor a partíció elemei a bennük futó élekkel együtt külön kis autonom gráfokat alkotnak, és az egész gráf nem más, mint ezek összessége. A színezések nyelvén: itt k különböző színű un. *klaszter*t tapasztalhatunk.

[1] BME Matematika Intézet, Sztochasztika Tanszék, marib@math.bme.hu

[2] MTA Matematikai Kutató Intézete, tusnady@circle.math-inst.hu

Az összefüggőségi tulajdonságot most statisztikusan fogjuk tekinteni. Ez azt jelenti, hogy egy gráfot jól klaszteresíthetőnek nevezünk akkor is, ha ugyan nem összefüggő, de valamely k pozitív egészre csúcsainak létezik olyan k-partíciója, hogy a partíció által definiált részgráfok "tömörek" (sok a belső él), a köztes élek száma pedig ehhez képest "elenyészően kevés". A köztes élek mennyiségét a gráf különböző módon súlyozott vágásaival lehet mérni. Ez azonban bonyolult leszámlálási feladatokhoz vezetne, ráadásul az összes szóbajövő k-ra ($1 < k < n$, ahol n a csúcsok száma) ki kellene számolni ezeket a mennyiségeket és minimalizálni az összes lehetséges partícióra. Ezért ezirányú vizsgálatainkat a spektrumon keresztül fogjuk végezni.

A gráfon belüli összefüggőség szorosságának mérésére Fiedler [10] 1973-ban bevezette az incidenciamátrixból számolt un. Laplace-mátrixot, és ennek legkisebb pozitív sajátértékét vizsgálta. Juhász és Mályusz [12] az előbb említett sajátértékhez tartozó sajátvektor kordinátáinak előjele alapján meg is találta a két lazán összefüggő komponenst. Ennek általánosításaként mi a k-klaszteresíthetőség mérésére a Laplace mátrix $k-1$ legkisebb pozitív sajátértékét fogjuk vizsgálni, maguknak a klaszteresnek a megtalálásához pedig a hozzájuk tartozó sajátvektorok alapján meghatározzuk a csúcsok un. *euklideszi reprezentánsait*. Ezek $(k-1)$-dimenziós vektorok, melyek metrikus klaszteresítése a matematikai statisztika klaszteranalízis című fejezetének jól körüljárt problémája, ld. McQueen [14], Dunn [8], Lengyel [13].

Vizsgálatainkat persze bármely k-ra ($1 < k < n$) elvégezhetjük, de k értékének növelése csak a dimenzió növelését jelenti olyan módon, hogy újabb komponenseket adunk hozzá a csúcsok reprezentánsaihoz, a már meglevő komponensek változatlanok maradnak. Azért előzőleg érdemes a spektrumbeli rés(ek) alapján tájékozódni az optimális k-ról.

A kromatikus számot is szokták a gráf incidenciamátrixának legnagyobb sajátértékével kapcsolatba hozni. A Laplace-mátrix alapján is kijön, hogy ha van k db. elég jól elkülönülő nagy sajátérték (itt a legnagyobb sajátértéktől lefelé haladva keressük a rést), akkor a gráf kromatikus száma ugyan nem biztosan k, de néhány belső él elvételével ez megoldható, azaz van olyan k-partíció, hogy "kevés" él fut a partíció elemein belül és sok a "köztes" él. Ha a kromatikus számot egy ilyen statisztikus "kvázi-kromatikus számmal" helyettesítenénk, akkor azt jellemezhetnénk a Laplace-mátrix nagy sajátértékeivel, a hozzájuk tartozó sajátvektorok segítségével pedig az optimális partíció metrikus szemlélet alapján lenne nyerhető. A térképészeti problémát ez természetesen nem oldaná meg, de nyilván vannak olyan particionálási feladatok a gyakorlatban, ahol azt szeretnénk, hogy az egy csoportba tartozás laza, a különbözőkbe való tartozás pedig szorosabb kapcsolatot fejezzen ki (pl. úgy akarunk klubokat létrehozni, hogy az egy klubba tartozó emberek lehetőleg ne ismerjék egymást, viszont sok legyen a kivezető szál). Ezzel a kérdéskörrel, mivel kevésbé természetes mint a másik, itt nem akarunk foglalkozni. Megjegyezzuk még, hogy a statisztikus vizsgálatokhoz mind a csúcsok, mind az élek számát tekintve nagyméretű gráfokra van szükség.

A klaszteresítési feladat lépten-nyomon felvetődik a gyakorlatban, pl. ismerősöket vagy hasonló tulajdonságokat szeretnénk egy klaszterbe sorolni, egyáltalán felismerni a hasonló tulajdonságokat. Számunkra egy ilyen probléma újszülöttek veleszületett rendellenességeinek statisztikai analízisekor állt elő, amikor kapcsolódási csoportokat (un. szindrómákat) szerettünk volna találni a sokféle (csaknem 50)

rendellenesség közt. Mintául mintegy 10000 rendellenesen született újszülöttön regisztrált megfigyelések szolgáltak, melyek nem is egy közönséges gráffal, hanem egy hipergráffal írhatók le. Míg egy gráfnál mindig két csúcs van éllel összekötve, addig egy hipergráf hiperéle több csúcsot is magában foglalhat, a továbbiakban ezeket is egyszerűen csak élnek nevezzük. Esetünkben a hipergráf csúcsai a rendellenességek, élei pedig az egyes újszülötteken regisztrált rendellenességeknek megfelelő csúcsrészhalmazok voltak.

Így a Laplace-mátrix fogalmát hipergráfokra általánosítjuk, és az összefüggőség kérdését is hipergráfokra fogjuk vizsgálni tetszőleges k ($1 < k < n$) egész szám esetén. A $k = 2$ eset, vagy közönséges gráfok esete ebből speciálisan adódik. Eredményeink fejezetek szerint a következőkben foglalhatók össze:

1. Definiálunk egy célfüggvényt, melynek minimalizálásával a csúcsok és élek olyan reprezentánsait kapjuk, hogy egy él reprezentánsa euklideszi távolságban mérve közel van az általa tartalmazott csúcsok reprezentánsaihoz. A célfüggvény lineáris algebrai segédeszközökkel minimalizálható, természetes módon adódik a Laplace-mátrix és annak spektrálfelbontása.
2. Definiálunk a k-partíciókat jellemző különbözőképpen súlyozott vágásokat, és ezekkel hozzuk összefüggésbe a Laplace-mátrix sajátértékeit. Nervezetesen, a $k-1$ legkisebb pozitív sajátérték összegére adunk alsó- és felső becslést a fenti kombinatorikus mérőszámok segítségével. A felső becslés valójában azt fejezi ki, hogy k "kicsi" sajátérték megléte szükséges feltétele a "jó" klaszteresíthetőségnek. Az alsó becslés más konstansokat is tartalmaz, és tudunk konstruálni példát arra, hogy bár az alsó becslés eléretik, a gráf mégsem klaszteresedik jól. Így ebből az alsó becslésből nem vonható le a következtetés, hogy a k "kicsi" sajátérték megléte elégséges is lenne.
3. Ugyanakkor vizsgáljuk a megfelelő sajátvektorokból adódó reprezentánsokat. Ezek metrikus "jó" klaszteresíthetősége (a klaszterátmérőkre tett korlátozásokkal) már elégséges a hipergráf jó klaszteresíthetőségéhez.
4. Létezik egy sejtés, hogy a spektrumbeli "rés" megléte önmagában is elégséges, de ehhez nem tudunk belátni egy lemmát, csak $k = 2$-re. Örömmel vennénk, ha $k > 2$-re valaki bebizonyítaná vagy ellenpéldát konstruálna.
5. Végül egy számítógépes algoritmust javaslunk nagyméretű hipergráfok klaszteresítésére. Az algoritmus az eredeti particionálási feladatot csak részként tartalmazza, eredményként pedig nem feltétlenül diszjunkt csúcs-klasztereket produkál.

A 6. fejezet néhány érdekes, de felhasználásra nem kerülő tényt közöl hipergráfok sajátértékeivel kapcsolatban, és megadjuk néhány speciális gráf euklideszi reprezentációját. A 7. fejezet a főbb tételek bizonyítását tartalmazza, ami érdekes lehet azok számára, akik lineáris algebrával foglalkoznak.

1. Hipergráfok euklideszi reprezentációja

Jelölje a $H = (V, E)$ hipergráf csúcsainak és éleinek halmazát $V = \{v_1 \dots v_n\}$ és $E = \{e_1 \dots e_m\}$. H egyértelműen megadható az A $n \times m$-es incidenciamátrixszal, melynek általános eleme $a_{ji} = \mathcal{I}(v_j \in e_i)$, ahol a $v \in e$ reláció azt jelöli, hogy a v csúcs benne van az e hiperélben:

$$\mathcal{I}(v \in e) = \begin{cases} 1, & \text{ha} \quad v \in e \\ 0, & \text{különben.} \end{cases}$$

Legyen k ($1 < k \leq n$) rögzített egész. Keressük a a csúcsok $\mathbf{x}_1, \dots, \mathbf{x}_n$ és az élek $\mathbf{y}_1, \dots, \mathbf{y}_m$ k-dimenziós reprezentánsait a

$$\sum_{j=1}^{n} \mathbf{x}_j \, \mathbf{x}_j^T \; = \mathbf{I}_k \tag{1.1}$$

kényszerfeltétel mellett úgy, hogy az élek költségösszegét kifejező

$$Q \; = \sum_{i=1}^{m} K(e_i) \tag{1.2}$$

célfüggvény minimális legyen, ahol

$$K(e_i) := \sum_{j=1}^{n} a_{ji} \|\mathbf{x}_j - \mathbf{y}_i\|^2. \tag{1.3}$$

az e_i hiperél reprezentálásának költsége.

A célfüggvény konstrukciója olyan, hogy a hipergráf élei össze szeretnék húzni a csúcsok reprezentánsait, míg a kényszerfeltétel kifeszíti azokat. Az optimális reprezentáció bizonyos kompromisszum: azok a csúcsok kerülnek közel egymáshoz, amelyek sok közös élben benne vannak.

Ezekután a célfüggvényt a következő lépésekben minimalizáljuk. Jelölje $\bar{\mathbf{x}}(e)$ az e hiperélet alkotó csúcsok reprezentánsainak átlagát:

$$\bar{\mathbf{x}}(e) := \frac{1}{|e|} \sum_{j=1}^{n} \mathcal{I}(v_j \in e) \mathbf{x}_j. \tag{1.4}$$

Jelölje $\mathbf{X} := (\mathbf{x}_1 \cdots \mathbf{x}_n)$ és $\mathbf{Y} := (\mathbf{y}_1 \cdots \mathbf{y}_m)$ a csúcsok és élek reprezentánsaiból, mint oszlopvektorokból álló $k \times n$-es ill. $k \times m$-es mátrixokat, továbbá $\mathbf{D}_v$ ill. $\mathbf{D}_e$ a csúcs- ill. él-fokszámokat tartalmazó $n \times n$-es ill. $m \times m$-es diagonálmátrixokat (a diagonálisban álló elemek valójában az incidenciamátrix marginálisai). Feltehető, hogy $\mathbf{D}_e$ nem szinguláris (nincsen üres él).

Ezekkel a jelölésekkel $K(e)$ csökkenthető és a Steiner-formula alapján a következő átalakítás végezhető:

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

A jobb oldalt $L(e,\mathbf{X})$-szel jelölve egy kis számolással

$$L(e,\mathbf{X}) = \frac{1}{2|e|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|\mathbf{x}_i-\mathbf{x}_j\|^2,\quad e\in E \tag{1.5}$$

adódik. Az $L(\mathbf{X}) := \sum_{e\in E} L(e,\mathbf{X})$ jelöléssel a $Q \geq L(\mathbf{X})$ egyenlőtlenség a csúcsok bármely $\mathbf{X}$ reprezentációjára fennáll. $L(\mathbf{X})$ viszont általánosított kvadratikus alakká alakítható:

$$L(\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{2}\sum_{e\in E}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\frac{1}{|e|}\right]\|\mathbf{x}_i-\mathbf{x}_j\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\mathbf{x}_i^T\mathbf{x}_j, \tag{1.6}$$

ahol

$$c_{ij} = \begin{cases} -\sum\limits_{e\in E}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\frac{1}{|e|}, & \text{if}\quad i\neq j,\\ s_i - \sum\limits_{e\in E}\mathcal{I}(v_i\in e)\frac{1}{|e|} = s_i' - \sum\limits_{\substack{e\in E\\ |e|>1}}\mathcal{I}(v_i\in e)\frac{1}{|e|}, & \text{if}\quad i=j, \end{cases} \tag{1.7}$$

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

1.1. Definíció. A (1.6)-beli kvadratikus alak mátrixát a H hipergráf *Laplace-mátrix*ának nevezzük és $\mathbf{C}$-vel jelöljük.

$\mathbf{C}$ elemei (1.7) alapján számolhatók, mátrix jelöléssel pedig $\mathbf{C} = \mathbf{D}_v - \mathbf{A}\mathbf{D}_e^{-1}\mathbf{A}^T$.

Minimalizáljuk a fenti $L(\mathbf{X})$-szel jelölt általánosított kvadratikus alakot az (1.1) kényszerfeltétel mellett! Könnyen látható, hogy a minimalizálandó kifejezés nem más, mint $\operatorname{tr}\mathbf{X}\mathbf{C}\mathbf{X}^T$, a kényszerfeltétel pedig $\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$. Mivel az $n\times n$-es $\mathbf{C}$ mátrix szimmetrikus és pozitív szemidefinit, egy – homogén kvadratikus alakok szélsőértékeire vonatkozó – tétel szerint (megtalálható pl. Rao [15]-ben, az 51. oldalon) bebizonyítottuk a következő un. *reprezentációs tételt*:

1.2. Tétel. *A* (1.2) *célfüggvény minimuma a* (1.1) *kényszerfeltétel mellett*

$$\sum_{j=1}^{k}\lambda_j, \tag{1.8}$$

ahol $0=\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ *jelöli a* $\mathbf{C}$ *Laplace-mátrix sajátértékeit. A minimum arra az* $\mathbf{X}$ *k-dimenziós euklideszi reprezentációra éretik el, amely a* $\mathbf{C}$ *mátrix k legkisebb pozitív sajátértékéhez tartozó páronként ortogonális, normált sajátvektorait tartalmazza soraiban, a sajátértékek növekvő sorrendje szerint. Egy ilyen optimális* $\mathbf{X}$*-et* $\mathbf{X}^*$*-gal jelölve, az élek optimális reprezentációjára* $\mathbf{Y}^* = \mathbf{X}^*\mathbf{A}\mathbf{D}_e^{-1}$ *adódik.* □

Az optimális reprezentáció egyértelműségéről a következők mondhatók el. Ha $\mathbf{R}$ tetszőleges $k\times k$-as ortogonális mátrix ($\mathbf{R}\mathbf{R}^T = \mathbf{I}_k$), akkor sem a célfüggvény, sem a kényszer nem változik az $\mathbf{X}' = \mathbf{R}\mathbf{X}$ helyettesítés hatására (nyilvánvaló is, hogy a reprezentások rendszerét elforgatva, kölcsönös helyzetük változatlan marad). Így

$\mathbf{X}^*$ csak forgatás erejéig egyértelmű, amennyiben a sajátvektorok azok (vagyis, ha a $\mathbf{C}$ mátrix sajátértékei mind egyszeres multiplicitásúak). Egyébként, ha többszörös sajátérték is van, akkor $\mathbf{X}^*$ megfelelő sorai tetszőlegesen választhatók (persze ortogonális módon) a megfelelő sajátaltéren belül.

Megjegyezzük, hogy k értéke itt még nem játszik különösebb szerepet, mivel bármely k-ra ($1 < k < n$) egy optimális $(k+1)$-dimenziós euklideszi reprezentáció könnyen nyerhető a k-dimenziósból egy újabb sajátvektor hozzávételével $\mathbf{X}^*$ soraihoz.

Az (1.7) képletből az is látszik, hogy a hurkok (élek, melyekre $|e| = 1$) nem járulnak hozzá a Laplace-mátrixhoz, így a továbbiakban csak hurokmentes hipergráfokkal foglalkozunk.

Vegyük észre, hogy a Laplace-mátrix mindig szinguláris (ui. sorösszegei 0-k). A 0 sajátértékhez tartozó normált sajátvektor az $\frac{1}{\sqrt{n}}\mathbf{e}$, vektor, ahol $\mathbf{e}$ jelöli az összes koordinátájában 1-t tartalmazó n-dimenziós vektort. Így ennek a koordinátának alapján a reprezentánsok nem különíthetők el, a k-dimenziós reprezentáció valójában az $\mathbf{e}$-re ortogonális $(k-1)$-dimenziós altérben történik.

Az is nyilvánvaló, hogy a H hipergráf összefüggő komponenseinek száma megegyezik a 0 sajátérték multiplicitásával (ha H összefüggő, csak egy 0 sajátértéke van), hisz az összefüggő komponensek $\mathbf{C}$-t blokkmátrixokra osztják, és mindegyiknek lesz egy 0 sajátértéke. A komponensek külön-külön vizsgálhatók, így a továbbiakban elég összefüggő hipergráfok vizsgálatára szorítkoznunk.

Összefüggő hipergráfok esetén természetesen merül fel a kérdés, vajon hány él elmozdításával szüntethető meg ez az állapot, és hogyan derül ki az összefüggés lazasága magából a spektrumból, továbbá hogyan tudjuk konstruktíve megtalálni az egymással lazán összefüggő komponenseket. Már Fiedler [10] megmutatta közönséges gráfokra, hogy λ_2 "kicsisége" két összefüggő komponens meglétét jelzi. Juhász és Mályusz [12] meg is konstruálták a két komponenst a λ_2-höz tartozó sajátvektor koordinátáinak előjele alapján. Mi a továbbiakban megmutatjuk, hogy a két komponens szétválásába λ_2 és λ_3 aránya is beleszól, több komponens szétválasztásához pedig a rákövetkező sajátértékeket is vizsgáljuk.

2. Hipergráfok spektruma és kombinatorikus tulajdonságai közti összefüggések

Legyen $H = (V, E)$, $|V| = n$, $|E| = m$ egy összefüggő, hurokmentes hipergráf, Laplace-mátrixának sajátértékeit jelölje $0 = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$. Bevezetünk néhány kombinatorikus mérőszámot, amelyek a hipergráf összefüggőségét karakterizálják.

H csúcsainak $(V_1, \ldots, V_k)$ nem-üres, diszjunkt részhalmazokra való felosztását *k-partíció*nak nevezzük és röviden P_k-val jelöljük. Jelölje $\mathcal{P}_k$ a hipergráf összes k-partícióinak halmazát.

2.1. Definíció. A $P_k = (V_1, \ldots, V_k)$ partíció $v(P_k)$ *sűrűség*ét a

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

$u(P_k)$-val jelölt *súlyozott sűrűség*ét pedig az

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

összefüggés definiálja, ahol $a_i(e) = |e \cap V_i|$ és $n_i = |V_i|$.

Ezek után legyen a H hipergráf *minimális k-sűrűség*e

$$\mu_k(H) = \min_{P_k \in \mathcal{P}_k} v(P_k), \tag{2.1}$$

*minimal súlyozott k-sűrűség*e pedig

$$\nu_k(H) = \min_{P_k \in \mathcal{P}_k} u(P_k). \tag{2.2}$$

2.2.Definíció. A $P_k = (V_1, \ldots, V_k)$ partíció *k-vágás*a azoknak az e éleknek az összessége, melyekre $|e \cap V_i| \neq \emptyset$ teljesül legalább két különböző V_i-vel. Ezt a halmazt $H(P_k)$-val jelöljük.

A P_k partíció tekinthető a csúcsok *k-színezés*ének is: a v csúcs színe $c(v) := i$, ha $v \in V_i$. Egy e hiperélet "tarkának" nevezünk ebben a színezésben, ha van legalább két különböző színű csúcs benne. Így a $H(P_k)$ halmaz éppen az ilyen tarka élekből áll.

A H hipergráf *minimális k-vágás*a az, amelynek számossága (a benne levő hiperélek száma) a legkisebb. Egy ilyen minimumot adó partíciót P_k^*-gal jelölve (nem biztos, hogy egyértelmű), $\theta_k(H) := |H(P_k^*)|$.

2.3. Megjegyzés. A fent definiált mennyiségekre nyilvánvalóan

$$\begin{aligned} &\mu_2(H) \leq \mu_3(H) \leq \cdots \leq \mu_{n-1}(H) \leq \mu_n(H), \\ &\nu_2(H) \leq \nu_3(H) \leq \cdots \leq \nu_{n-1}(H) \leq \nu_n(H), \\ &\theta_2(H) \leq \theta_3(H) \leq \cdots \leq \theta_{n-1}(H) \leq \theta_n(H) = m. \end{aligned} \tag{2.3}$$

□

2.4. Tétel. *A H hipergráf k legkisebb sajátértékének összegére a fenti kombinatorikus mennyiségekkel*

$$c_n\theta_k(H) \le \sum_{j=1}^{k} \lambda_j \le \nu_k(H) \tag{2.4}$$

teljesül, ahol $c_n = \frac{6}{n(n^2-1)}$.

A tétel bizonyítása a 7. fejezetben található.

Ha $\nu_k(H)$ "kicsi", ez azt jelenti, hogy a minimális súlyozott k-sűrűséget adó k-partíció olyan tulajdonságú, hogy "kevés" benne a tarka él és azok is viszonylag kisméretű csúcs-klaszterek között futnak. Ezért a fenti felső becslés azt jelzi, hogy ilyen esetben a k legkisebb sajátérték összege is kicsi. Azaz k viszonylag kis sajátérték megléte szükséges feltétele a jó klaszteresíthetőségnek.

Az alsó becslés egy, a csúcsok számától függő konstanst is tartalmaz. Vannak gráfok (ld. a 6. fejezet szalagjai, hálói és pókjai, 6.7.–6.10. példák), melyekre az alsó becslés nagyságrendileg eléretik, mégsem klaszteresednek jól semmilyen k-ra. Persze, itt a sajátértékek eloszlása egyenletes, nincsen rés a spektrumban.

$k = 2$ esetén precízebb becslés adható λ_1 és λ_2 összegére, azaz λ_2-re:

2.5. Tétel. *Legyen H összefüggő hipergráf és jelölje λ_2 a legkisebb pozitív sajátértékét. Akkor*

$$\lambda_2 \ge \begin{cases} 2(1-\cos\frac{\pi}{n})\mu_2(H), & \text{ha } 0 \le \mu_2(H) \le \frac{1}{2}s_{\max} \\ c_1\mu_2(H) - c_2 s_{\max}, & \text{ha } \frac{1}{2}s_{\max} < \mu_2(H), \end{cases} \tag{2.5}$$

ahol $c_1 = 2(\cos\frac{\pi}{n} - \cos\frac{2\pi}{n})$, $c_2 = 2\cos\frac{\pi}{n}(1-\cos\frac{\pi}{n})$ *és* $s_{\max} = \max_j s_j$.

Közönséges gráfokra ez az alsó becslés megegyezik a Fiedler által [10]-ben adottal.

3. A reprezentánsok klaszterei

Most az optimális k-partíciókat szeretnénk felismerni. Ehhez válasszunk egy olyan k-t, melyre λ_k és λ_{k+1} között viszonylag nagy rés van a spektrumban. Tekintsük a hipergráf optimális k-dimenziós euklideszi reprezentációját, ahol az $\mathbf{x}_1^*, \ldots, x_n^*$ csúcs-reprezentánsok valójában $(k-1)$-dimenziós vektorok. Klaszteresítsük őket a k-közép módszerrel (ld. McQueen [14]). Tegyük fel, hogy a fenti pontoknak létezik az alább definiált jó tulajdonságokkal rendelkező k-partíciója.

3.6.Definíció. *A $P_k = (V_1, \ldots, V_k)$ partíciót a csúcsok $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ euklideszi reprezentációjában jól-szeparáltnak nevezzük, ha vele $\alpha(P_k) > 1$ teljesül, ahol*

$$\alpha(P_k) := \frac{\min\limits_{c(v_i)\neq c(v_j)} \|\mathbf{x}_i - \mathbf{x}_j\|}{\max\limits_{c(v_i)=c(v_j)} \|\mathbf{x}_i - \mathbf{x}_j\|} \tag{3.7}$$

Ez azt jelenti, hogy az egy klaszteren belüli reprezentánsok maximális távolsága is kisebb, mint a különböző klaszterbeliek közti minimális távolság. (Ha ilyen jól-szeparált k-partíció létezik, akkor Dunn [8], [9]-ben bebizonyította, hogy az egyértelmű, és algoritmust adott a meghatározására.)

A következő tétel arról szól, hogy amennyiben létezik a reprezentánsoknak egy "nagyon" jól-szeparált k-partíciója, (ezalatt a klaszterátmérőkre tett további korlátozásokat értünk), akkor ugyanaz a mennyiség, amivel az előző fejezetben a k legkisebb sajátérték összegét felülről becsültük, most alsó becslést ad a k legkisebb sajátérték összegének konstansszorosára.

3.7. Tétel. *Tegyük fel, hogy valamely $1 < k < n$ egészre létezik az optimális k-dimenziós reprezentánsoknak olyan jól-szeparált klaszteresítése, melyben a klaszterátmérők ε-nál kisebbek, ahol $\varepsilon < \frac{1}{2\sqrt{n}}$. Akkor*

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

ahol $q = 1 + \frac{2\varepsilon}{1-\varepsilon\sqrt{n}}$.

A tétel bizonyítása szintén a 7. fejezetben található.

Összehasonlítva a 2.4 és 3.7 tételek eredményeit azt kapjuk, hogy

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

Tehát az előbbi tétel feltételei mellett $\sum_{j=1}^{k} \lambda_j$ és $\nu_k(H)$ legfeljebb csak egy 4-es fatorban különböznek, azaz a reprezentánsok metrikusan jó klaszteresedése esetén $\sum_{j=1}^{k} \lambda_j$ kicsisége elégséges is a kombinatorikus értelemben vett jó klaszteresíthetőséghez.

4. Becslések súlyozott gráfokra

További vizsgálatainkat a spektrumbeli rés elégségességével kapcsolatban egyszerűbb súlyozott gráfokra megfogalmazni, mivel ott bizonyos folytonosság teljesül (a súlyok ui. tetszőleges valós számok, melyekkel könnyebb perturbációs eredményeket bizonyítani). Az eredmények érvényben maradnak hipergráfokra is, hiszen minden hipergráfhoz egyértelműen hozzárendelhető egy, az alábbiakban definiált súlyozott gráf.

Legyen tehát $G = (V, \mathbf{W})$, egy súlyozott gráf a $V := \{v_1, \dots v_n\}$ csúcshalmazon a $\mathbf{W}$ súlymátrixszal megadva. $\mathbf{W}$ egy $n \times n$-es szimmetrikus mátrix, diagonálisa azonosan zéró, míg a nemdiagonális $w_{ij} = w_{ji} \geq 0,\ i \neq j$ elem a $\{v_i, v_j\}$ él súlya (ha v_i és v_j nincsenek éllel összekötve, akkor ez a súly 0). Egy közönséges gráf ennek speciális esete, ahol a súlymátrix a szokásos 0–1 adjacencia-mátrix.

A G súlyozott gráf euklideszi reprezentációjában csak a csúcsokhoz rendelünk reprezentánsokat. Az $\mathbf{x}_1, \dots, \mathbf{x}_n$ $(k-1)$-dimenziós reprezentánsokra a $\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T = \mathbf{I}_{k-1}$ kényszerfeltétel mellett még a $\sum_{j=1}^{n} \mathbf{x}_j = \mathbf{0}$ kikötést is tesszük, így az első koordináták egyenlőségét kizárjuk. (Az 1. fejezetbeli k-dimenziós reprezentánsokról is láttuk, hogy valójában egy $(k-1)$-dimenziós altérben helyezkednek el.)

A minimalizálandó célfüggvény most

$$Q := \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \operatorname{tr} \mathbf{X}\mathbf{C}\mathbf{X}^T, \tag{4.1}$$

ahol az $\mathbf{X}$ $(k-1) \times n$-es mátrix a reprezentánsokat tartalmazza oszlopaiban, a $\mathbf{C}$ $n \times n$-es mátrix pedig a súlyozott gráf Laplace-mátrixa. Könnyű látni, hogy $\mathbf{C} = \mathbf{D} - \mathbf{W}$, ahol a $\mathbf{D}$ diagonálmátrix diagonális elemei $d_i = \sum_{j=1}^{n} w_{ij}$, $(i = 1, \dots, n)$. Az is könnyen látható, hogy az ugyanezen a csúcshalmazon definiált $H = (V, E)$ hipergráfhoz

$$w_{ij} = w_{ji} = \sum_{e \in E} \mathcal{I}(v_i \in e)\mathcal{I}(v_j \in e)\frac{1}{|e|}, \qquad (1 \leq i < j \leq n).$$

súlyokkal hozzárendelt súlyozott gráf Laplace-mátrixa azonos a hipergráféval. Egy összefüggő súlyozott gráf (súlymátrixa nem hasad blokkokra) Laplace-mátrixának pontosan egy 0 sajátértéke van és rendelkezik a jól ismert tulajdonságokkal.

Q minimuma itt is a Laplace-mátrix k legkisebb (vagy ami ekvivalens, $k-1$ legkisebb pozitív) sajátértékének összege, és eléretik, ha az $\mathbf{X}$ mátrix soraiba a hozzájuk tartozó sajátvektorokat tesszük (itt a 0 sajátértékhez tartozó sajátvektort kihagyjuk a reprezentációból).

A csúcsok egy $P_k = (V_1, \dots, V_n)$ partíciójának sűrűsége és súlyozott sűrűsége itt is hasonlóan definiálható:

$$v(P_k) := \sum_{l=1}^{k-1} \sum_{m=l+1}^{k} w'_{lm}, \quad u(P_k) := \sum_{l=1}^{k-1} \sum_{m=l+1}^{k} \left(\frac{1}{n_l} + \frac{1}{n_m}\right) w'_{lm},$$

ahol $w'_{lm} = \sum_{v_i \in V_l} \sum_{v_j \in V_m} w_{ij}, \quad (1 \le l < m \le k)$ és $n_p = |V_p|$. A μ_k, ν_k mennyiségek pedig ezek $\mathcal{P}_k$-n vett minimumai.

A perturbációs vizsgálatokhoz rögzítsük a P_k partíciót. Ennek megfelelően a G súlyozott gráf két, éldiszjunkt részre bontható: Az egyik tartalmazza az egyszínű, a másik pedig a tarka (itt kétszínű) éleket a P_k partíció által meghatározott színezésben. Mindkét részt ugyanazon a csúcshalmazon tekintve a két súlyozott gráf Laplace-mátrixa összeadódik (ld. 6.1 állítás): $\mathbf{C} = \mathbf{B}+\mathbf{P}$. Mivel az egyszínű élekből álló rész k összefüggő komponensre bontható (az egyes színeknek megfelelően), $\mathbf{B}$-nek pontosan k db. 0 sajátértéke lesz. Jelölje ϱ a $\mathbf{B}$ mátrix legkisebb pozitív sajátértékét (ez azonos valamely egyszínű blokk legkisebb pozitív sajátértékével). Legyen továbbá $\varepsilon := ||\mathbf{P}||$ (ε annál kisebb, minél kevesebb a tarka él). Tegyük fel, hogy $\varepsilon < \varrho$.

4.1. Definíció. *Az $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^{k-1}$ vektorok k-varianciája a P_k partícióban*

$$S_k^2(P_k, \mathbf{X}) := \sum_{i=1}^{k} \sum_{j:c(j)=i} \left\| \mathbf{x}_j - \frac{\sum_{l:c(l)=i} \mathbf{x}_l}{n_i} \right\|^2,$$

ahol c a megfelelő színezés és $n_i = |V_i|$. Az $\mathbf{x}_1, \dots, \mathbf{x}_n$ vektorok k-varianciája

$$S_k^2(\mathbf{X}) := \min_{P_k \in \mathcal{P}_k} S_k^2(P_k, \mathbf{X}).$$

A fenti jelölésekkel a következő állítás látható be:

4.2. Tétel. *Az $\varepsilon < \varrho$ feltevés mellett az*

$$S_k^2(P_k, \mathbf{X}^*) \le k\frac{\varepsilon}{\varrho}$$

felső becslés adható az $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ optimális $(k-1)$-dimenziós reprezentánsok P_k-beli k-varianciájára.

A bizonyítást ld. a 7. fejezetben. Megjegyezzük, hogy

$$\varepsilon = \|\mathbf{P}\| \le \operatorname{tr} \mathbf{P} = \sum_{\substack{i,j \\ c(i)\neq c(j)}} w_{ij} = v(P_k)$$

és

$$\varrho = \min_i \lambda_1(\mathbf{B}_i) \ge \begin{cases} 2(1-\cos\frac{\pi}{n\,i})\mu_2(G_i), & \text{if } 0 \le \mu_2(G_i) \le \frac{1}{2} d_{i\max} \\ c_{i1}\mu_2(G_i) - c_{i2} d_{i\max}, & \text{if } \frac{1}{2} d_{i\max} < \mu_2(G_i), \end{cases}$$

ahol $c_{i1} = 2(\cos\frac{\pi}{n\,i} - \cos\frac{2\pi}{n\,i})$, $c_{i2} = 2\cos\frac{\pi}{n\,i}(1-\cos\frac{\pi}{n\,i})$, $d_{i\max} = \max_{j\in V_i} d_j$ – ld. 2.5 Tétel – és $\mathbf{B}_i$ a V_i csúcshalmaz által indukált, G_i-vel jelölt súlyozott részgráf $n_i \times n_i$-es Laplace-mátrixa. $\mathbf{B}_i$ a $\mathbf{B}$ mátrix i-edik diagonális blokkja. Ezért minél kisebb a P_k partíció sűrűsége és minél nagyobb az egyszínű részgráfok 2-vágása (azaz a G_i részgráfok erősen összefüggőek), annál jobban klaszteresíthetők az optimális $(k-1)$-dimenziós reprezentánsok k klaszterbe.

A fenti becslés egy adott k-partícióra vonatkozik, ϱ és ε az adott P_k-tól függ. A következő állítás az optimális $(k-1)$-dimenziós reprezentánsok k-varianciájára vonatkozik.

4.3. Állítás. *Legyen $\mathbf{X}^*$ optimális $(k-1)$-dimenziós reprezentációja a fenti súlyozott gráfnak. A csúcs-reprezentánsok k-varianciájára az*

$$S_k^2(\mathbf{X}^*) \le S_k^2(P_k, \mathbf{X}^*) \le \frac{\lambda_1 + \cdots + \lambda_{k-1}}{\varrho(P_k)}$$

összefüggés teljesül bármely P_k partícióval, ahol $\varrho(P_k)$ a P_k partíció által indukált részgráfok tömörségére jellemző, az előbbiekben bevezetett konstans.

Szeretnénk most a fenti k-varianciát közvetlenül a λ_k és λ_{k+1} sajátértékek hányadosával felülbecsülni. Ez $k = 2$ esetén sikerülni is fog. Az eredményt még általánosabban, súlyozott csúcsú súlyozott gráfokra mondjuk ki.

Legyenek a $G = (V, \mathbf{W})$ súlyozott gráf csúcsai az $d_1, \ldots, d_n$ súlyokkal ellátva, ahol $d_j = \sum_{i \ne j} w_{ij}$. Jelölje $\mathbf{D}$ az ezeket a súlyokat ilyen sorrendben diagonálisában tartalmazó diagonálmátrixot. Most a (4.1)-beli Q célfüggvényt a $\sum_{j=1}^n d_j \mathbf{x}_j \mathbf{x}_j^T = \mathbf{X}\mathbf{D}\mathbf{X}^T = \mathbf{I}_k$ és $\sum_{j=1}^n d_j \mathbf{x}_j = \mathbf{0}$ kényszerfeltételek mellett minimalizáljuk. Mivel Q most

$$\operatorname{tr} \mathbf{X}\mathbf{D}\mathbf{X}^T = \operatorname{tr} (\mathbf{X}\mathbf{D}^{1/2})[\mathbf{D}^{-1/2}\mathbf{C}\mathbf{D}^{-1/2}](\mathbf{X}\mathbf{D}^{1/2})^T$$

alakban írható, és az $\mathbf{X}\mathbf{D}$ mátrix sorai a kényszer miatt ortonormáltak, Q minimumát most a szögletes zárójelben álló, a továbbiakban $\mathbf{C}_D$-vel jelölt *súlyozott Laplace-mátrix* (amely szintén szimmetrikus, szinguláris, pozitív szemidefinit) $k-1$ legkisebb pozitív sajátértékének összege adja, az optimális $(k-1)$ dimenziós reprezentánsok pedig az $\mathbf{X}^*\mathbf{D}^{1/2}$ mátrixból nyerhetők, mely a megfelelő sajátvektorokat tartalmazza soraiban. $\mathbf{C}_D$ sajátértékeire egyébként 2 felső korlát, így ha G még összefüggő is, akkor

$$0 = \lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \le \lambda_n \le 2.$$

4.4 Tétel. *Legyen $\mathbf{X}^*$ optimális 1-dimenziós reprezantációja a fenti módon súlyozott gráfnak ($\mathbf{X}^*$ a λ_2-höz tartozó sajátvektorból nyerhető transzformációval). Akkor az optimális reprezentánsok 2-varianciájára a spektrumbeli réssel a következő becslés adható:*

$$S_2^2(\mathbf{X}^*) \le \frac{\lambda_2}{\lambda_3}.$$

Felmerül a kérdés, vajon a $(k-1)$-dimenziós optimális euklideszi reprezentánsok k-varianciáját nem lehetne-e felülről becsülni a λ_k és λ_{k+1} közti spektrumbeli réssel vagy annak esetleg k-tól függő konstansszorosával. Sejtésünk a következő:

4.5. Sejtés.

$$S_k^2(\mathbf{X}^*) \le (k-1) \cdot \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_{k+1}}, \qquad 1 \le k < n-1.$$

A 4.4 tétel bizonyítása és a 4.5 sejtés bizonyításához szükséges lemma a 7.fejezetben van leírva.

5. Egy ad hoc algoritmus hipergráfok klasztereinek megállapítására

Mintául a $v_1, v_2, \ldots, v_n$ bináris (0–1) változókra tett $e_1, e_2, \ldots, e_m$ megfigyelések szolgálnak ($n \ll m$). Ezek a $H = (V, E)$ hipergráfot alkotják, ahol $V = \{v_1, v_2, \ldots, v_n\}$ és $E = \{e_1, e_2, \ldots, e_m\}$, $\mathcal{I}(v \in e) = v(e)$ pedig 1 vagy 0 aszerint, hogy az e objektumon a v tulajdonságot megfigyelték-e vagy sem.

Legyen $E' \subset E$ egy al-minta, amely a $H' = (V, E')$ hipergráfot generálja. Jelölje $0 = \lambda_1(H') \leq \lambda_2(H') \leq \cdots \leq \lambda_n(H')$ ta H' hipergráf Laplace-spektrumát és az $n \times n$-es $\mathbf{X}^*(H')$ mátrix tartalmazza soraiban a hozzájuk tartazó teljes ortonormált sajátvektorrendszert. Az 1.2 reprezentációs tétel szerint bármely d egészre ($1 \leq d \leq n$) a $d \times n$-es $\mathbf{X}^*_d(H')$ mátrix, amely $\mathbf{X}^*(H')$ első d sorát tartalmazza, a H' hipergráf optimális d-dimenziós reprezentációját adja. Az E'-beli élek összvarianciája ebben a reprezentációban

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

A H' hipergráf beágyazásának költségét a

$$K(H') := \min_{d \in \{1,\ldots,n\}} [c2^{n-d} + L(\mathbf{X}^*_d(H'))]$$

célfüggvénnyel definiáljuk, ahol a c konstanst előre választjuk meg (a probléma méretének megfelelően), és a $c2^{n-d}$ tag a túlságosan nagy dimenziókat bünteti (az élek összvarianciáját kifejező $L(\mathbf{X}^*_d(H'))$ tag – épp ellenkezőleg – a dimenzió növelésével csökkenthető). A minimumot adó d^* dimenziót az E' *él-klaszter dimenzió*jának nevezzük.

Jelölje $\mathcal{S}$ az E élhalmaz összes lehetséges partícióit. Keressük azt az $S \in \mathcal{S}$ partíciót, melyre a $K = \sum_i K(H_i)$ célfüggvény minimális, ahol $H_i = (V, E_i)$. Most válasszunk és rögzítsünk egy k egészet ($1 \leq k \leq n$). Definiálunk egy iterációt, amely a fenti célfüggvény egy relatív minimumához vezet, ha csak az $\mathcal{S}_k$-val jelölt k-partíciók körében keressük a minimumot. Legyen tehát $(E_1, \ldots, E_k) \in \mathcal{S}_k$ az E élhalmaz egy k-partíciója. Az előző jelöléseket alkalmazva az indukált $H_i = (V, E_i)$, ($i = 1, \ldots, k$) rész-hipergráfokra a

$$Q_{d_i}(H_i) := c2^{n-d_i} + L(\mathbf{X}^*_{d_i}(H_i)), \quad (i = 1, \ldots, k)$$

jelöléseket bevezetve a $Q = \sum_{i=1}^{k} Q_{d_i}(H_i)$ költségfüggvényt fogjuk minimalizálni a $\mathcal{S}_k$-beli partíciók és a $d_1, \ldots, d_k$ dimenziók körében.

A minimumot kereső iteráció a következő lépésekből áll:

0. Kiindulásul tekintsük az E élhalmaz tetszőleges $E_1, \ldots, E_k$ partícióját (egy ilyet nyerhetünk pl. a k-közép módszerrel, ld. [13], [14]).
1. Az $E_1, \ldots, E_k$ klasztereket rögzítve: meghatározzuk a $H_i = (V, E_i)$ hipergráfok Laplace-mátrixainak spektrálfelbontását. Ezután $Q_{d_i}(H_i)$-t a d_i dimenzióban minimalizáljuk (minden i-re külön). Mivel $1 \leq d_i \leq n$ egész, ez egy diszkrét

minimalizálási feladat. Jelölje d_i^* a minimumot adó (nem feltétlenül egyértelmű) dimenziót, amellyel tehát

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

2. Most a d_i^*-dimenziókat rögzítve az objektumokat átsoroljuk a klaszterek közt: az e objektumot abba az E_i klaszterbe helyezzük, amelyben a hozzá tartozó $L(e, \mathbf{X}_{d_i^*}^*(H_i))$ variancia minimális (ha több klaszterre is minimális, akkor vegyük a legkisebb ilyen i-t). Az objektumok így nyert új klaszteresítését $E_1^*, \ldots, E_k^*$-gal jelölve, ezekkel megismételjük az 1. és 2. lépéseket, amíg csak Q csökkenthető.

Triviális, hogy a fenti lépések Q értékét csökkentik, s mivel az objektumok száma véges, az algoritmus véges lépésben Q relatív minimumához vezet. (Esetünkben, $n = 50$, $m = 10000$ értékekkel az iteráció 5–6 lépésben véget ért.)

Bevezethetnénk egy k-ban minimalizáló lépést is, így azonban az algoritmus nagyon hosszadalmas lenne. Inkább végigcsináljuk néhány kiválasztott k-ra (pl. a k-közép eljárást lefuttatva kaphatunk k-ra ötletet), és összehasonlítjuk a minimumként kapott Q értékeket.

Az iteráció során kiürülhetnek, és általában ki is ürülnek él-klaszterek. k értéke természetesen ezzel csökken. A $H_i = (V, E_i)$ hipergráfok általában tartalmaznak izolált csúcsokat (nem összefüggőek), jelölje V_i a nem izolált csúcsok halmazát. Ekkor $\cup_{i=1}^k V_i = V$, de a $V_1, \ldots, V_k$ rendszer nem feltétlenül diszjunkt. Ezek a diszjunkt él-klaszterekre jellemző tulajdonság-asszociációkat tartalmazzák. A mi mintánkon, ahol a tulajdonságok veleszületett rendellenességek, az objektumok pedig érintett újszülöttek voltak, ezek az asszociációk épp a rendellenességek speciális kapcsolódási csoportjait, un. szindrómáit adták meg.

6. Néhány megjegyzés hipergráfok spektrumához

Az első fejezet jelöléseit használjuk, ill. ha nem egyértelmű, hogy melyik hipergráfról van szó, akkor a Laplace-mátrix, sajátértékek és egyéb mennyiségek argumentumaiban feltüntetjük a hipergráfot.

6.1. Megjegyzés. Legyenek a $H_i = (V, E_i)$, $(i = 1, \ldots, k)$ hipergráfok él-diszjunktak ugyazon a csúcshalmazon. Az $E = \cup_{i=1}^{k} E_i$, $E_i \cap E_j = \emptyset$ $(i \neq j)$, $H = (V, E)$ jelöléssel a H hipergráf Laplace-mátrixára

$$\mathbf{C}(H) = \sum_{i=1}^{k} \mathbf{C}(H_i) \tag{6.1}$$

teljesül. □

6.2. Megjegyzés. Legyenek a $H_i = (V, E_i)$, $i = 1, 2$ hipergráfok él-diszjunkt részhipergráfjai a $H = (V, E)$ hipergráfnak, azaz $E = E_1 \cup E_2$, $E_1 \cap E_2 = \emptyset$. Akkor

$$\sum_{j=1}^{k} \lambda_j \geq \sum_{j=1}^{k} \lambda_j^{(1)} + \sum_{j=1}^{k} \lambda_j^{(2)}, \quad (1 \leq k \leq n), \tag{6.2}$$

ahol $\lambda_j^{(i)}$ a H_i hipergráf j-edik sajátértékét jelöli nagyság szerint növekvő sorrendben (i=1,2).

6.3. Megjegyzés. Az előző megjegyzés jelöléseivel

$$\lambda_{j-r_i} \leq \lambda_j^{(i)} \leq \lambda_j, \quad (j = 1, \ldots, n), \tag{6.3}$$

ahol r_i a H_i hipergráf $\mathbf{C}_i$ Laplace-mátrixának a rangja $(i = 1, 2)$, továbbá $\lambda_l = 0$, ha $l < 1$.

6.4. Megjegyzés. Legyen e a $H = (V, E)$ hipergráf egy éle, $|e| = z$. Az $E' := E \backslash \{e\}$, $H' := (V, E')$ jelölésekkel

$$\sum_{j=1}^{k-z+1} \lambda_j \leq \sum_{j=1}^{k} \lambda'_j \leq \sum_{j=1}^{k} \lambda_j, \quad (z \leq k \leq n-1) \tag{6.4}$$

ahol λ'-k jelölik a H' hipergráf Laplace-mátrixának sajátértékeit.

6.5. Következmény. z=2 esetén a (7.3) összefüggés két egyenlőtlenségét egymás után váltakozva alkalmazva adódik, hogy

$$0 \leq \lambda'_2 \leq \lambda_2 \leq \lambda'_2 + \lambda'_3 \leq \lambda_2 + \lambda_3 \leq \lambda'_2 + \lambda'_3 + \lambda'_4 \leq \lambda_2 + \lambda_3 + \lambda_4 \leq \cdots . \tag{6.5}$$ □

Az alábbi megjegyzések az 1. fejezet definíciói és néhány, a pozitív definit mátrixok sajátértékeire vonatkozó tétel (pl. [15]-ben) alapján könnyen bizonyíthatók,

a bizonyítások megtalálhatók [6]-ban. A továbbiakban megadjuk néhány jól ismert gráf ill. hipergráf optimális euklideszi reprezentációját.

6.6. Példa. Legyen C_n az n csúcsú *teljes hipergráf* $2^n - n - 1$ hiperéllel. Laplace-mátrixának legkisebb sajátértéke 0, a többi sajátérték pedig mind egyenlő az

$$\frac{n2^{n-1} - 2^n + 1}{n-1} \tag{6.6}$$

számmal. Ehhez az $(n-1)$-szeres multiplicitású sajátértékhez tartozó sajátvektorok tetszőlegesen választhatók (persze azért úgy, hogy ortonormált rendszert alkossanak) az $\mathbf{e} = (1, \dots, 1)$ vektorra merőleges altérben. □

6.7. Példa. Legyen P_n a *szalag-gráf.* Tegyük fel, hogy a csúcsok száma páratlan $(n = 2l+1)$, és $v_{-l}, \dots, v_0, \dots, v_l$-ként indexeljük őket. A legkisebb pozitív sajátérték $1-\cos\frac{\pi}{n}$, az ehhez tartozó sajátvektor koordinátái – melyek az optimális 2-dimenziós reprezentációban lépnek fel – pedig

$$x_j = \frac{\sqrt{2}}{\sqrt{n}} \sin(j\frac{\pi}{n}), \quad j = -l, \dots, 0, \dots, l. \tag{6.7}$$ □

6.8. Példa. Legyen S_d az $n = d + 1$ csúcsú *csillag.* Ennek legkisebb pozitív sajátértéke $1/2$, multiplicitása $d-1$. Egy optimális d-dimenziós euklideszi reprezentációt ad egy d-csúcsú szabályos poliéder, ahol az 1-fokú csúcsok reprezentánsai a poliéder csúcsai, a d-fokú csúcs reprezentánsa pedig a poliéder szimmetriacentruma. □

6.9. Példa. Jelölje $G_{d,l}$ az S_d gráf finomítását úgy, hogy S_d éleit további csúcsok beiktatásával l élre osztjuk fel, így a csúcsok száma $n = dl + 1$. $G_{d,l}$-t egy d-lábú, minden lábán l ízt tartalmazó *pók*nak nevezzük. A fenti pók Laplace-mátrixának legkisebb pozitív sajátértéke $1 - \cos\frac{\pi}{2l+1}$ és multiplicitása $d-1$. $G_{d,l}$ egy optimális d-dimenziós euklideszi reprezentációja az S_d és P_{2l+1} gráfok reprezentációjából adódik, ahol a szabályos poliéder szimmetriacentrumát a csúcsaival összekötő szakaszok vannak (6.7) szerint szinuszosan felosztva. □

6.10. Példa. Jelölje $L_{d,l}$ a d-dimenziós *rács*ot. Csúcsainak koordinátái a $-l, \dots, 0, \dots, l$ számokból kiválasztott d-esek, ahol két ilyen d-es akkor van összekötve éllel, ha pontosan egy koordinátájukban különböznek. A csúcsok száma így $n = (2l+1)^d$. Az $L_{d,l}$ gráf legkisebb pozitív sajátértéke $1-\cos\frac{\pi}{2l+1}$, multiplicitása d. Egy optimális $(d+1)$-dimenziós euklideszi reprezentáció egy d-dimenziós rács, szimmetriacentruma az origó, oldalai pedig (6.7) szerint szinuszosan vannak felosztva. □

6.11. Példa. Legyen $K_{n_1,\dots,n_k}$ a *Turán-gráf*, amely a csúcsok $V_1, \dots, V_k$-val jelölt un. független részhalmazaiból áll (a függetlenség azt jelenti, hogy a részhalmazokon belül nem futnak élek). Legyen $|V_i| = n_i$, $(i = 1, \dots, k)$ és $n = \sum_{i=1}^{k} n_i$ a csúcsok száma. A színezések nyelvén, ebben a gráfban nincsenek egyszínű élek, viszont az összes lehetséges fajta kétszínű él megtalálható. $K_{n_1,\dots,n_k}$ spektruma: egy db. 0, n_i db. $\frac{1}{2}(n - n_i)$, $(i = 1, \dots, k)$, a legnagyobb sajátérték pedig $\frac{1}{2}n$, multiplicitása

$k-1$. Ha ez utóbbihoz tartozó sajátvektorokkal adjuk meg a $K_{n_1,\ldots,n_k}$ euklideszi reprezentációját, akkor az azonos színű csúcsok reprezentánsai ugyanabba a $(k-1)$-dimenziós pontba esnek össze, így a reprezentáció k különböző pontból áll. Megjegyezzük, hogy a Turán-gráf kromatikus száma k.

A másik véglet az az eset, mikor a hipergráf k összefüggő komponensből áll. Ilyenkor a 0 sajátérték multiplicitása k, és a hozzá tartozó sajátvektorok által meghatározott euklideszi reprezentációban az azonos komponensbe tartozó csúcsok reprezentánsai esnek egybe.

Az első három fejezetben láttuk, hogy ha a hipergráf összefüggő (csak egy 0 sajátérték van), de van $k-1$ "kicsi" pozitív sajátértéke, akkor várható, hogy a hozzájuk tartozó sajátvektorok által meghatározott reprezentációban a hipergráf jól klaszteresedik. Mikor az 1. fejezetbeli Q célfüggvény maximumát keressük, analóg módon megkérdezhetnénk, vajon $k-1$ elkülönülten "nagy" sajátérték a spektrumban jelzi-e mindig k db. közel független csúcsklaszter meglétét, és következik-e ebből a megfelő sajátvektorok által történő reprezentációban a reprezentánsok metrikus értelemben vett jó klaszteresíthetősége? Tudunk-e ebből következtetni a kromatikus számra? Valószínűleg nem, mert a kromatikus szám egy szigorúan kombinatorikus mennyiség. Viszont, ha egy olyan kvázi-kromatikus számot tekintenénk, ahol nem baj, ha van néhány egyszínű él is, akkor ezt tudná jelezni a nagy sajátértékekre adaptált algoritmus, sőt kiadná a közel független csúcsklasztereket is.

7. A fontosabb tételek bizonyítása

7.1. Lemma. Legyen $\Delta > 0$ valós, n és k pedig rögzített egészek ($1 \le k < n$). Akkor tetszőleges, a $\mathcal{T}$ tulajdonságot teljesítő $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^k$ pont-n-es kiszínezhető $k+1$ különböző színnel oly módon, hogy a különböző színűek közti minimális távolság legalább Δ, ahol a $\mathcal{T}$ tulajdonság a következő: R^k bármely egyenesére vetítve a pontokat, ezen az egyenesen van legalább két szomszédos pont, melyek távolsága legalább Δ.

Bizonyítás. Teljes indukcióval, $k = 1$-re az állítás maga a $\mathcal{T}$ tulajdonság. Tegyük fel, kogy $(k-1)$-re már bebizonyítottuk az állítást. k-ra: a $\mathcal{T}$ tulajdonság szerint a pontok kiszínezhetők 2 különböző színnel a kívánt módon. Válasszunk egy-egy pontot mindegyik színből. Kössük össze őket és és vetítsük pontjainkat az összekötő egyenesre merőleges $(k-1)$-dimenziós altérre.

Ugyanakkor tekintsük az $\mathbf{x}_1, \ldots, \mathbf{x}_n$ pontok Δ-szintgráfját (az egymástól Δ-nál közelebb levőket kötjük össze). Azt kell megmutatnunk, hogy ez a gráf $k+1$ összefüggő komponensből áll. Mivel a pontok a fentiek miatt 2 színnel kiszínezhetők, legalább 2 komponensünk van. A vetítés után azonban ez a 2 komponens össze lesz kötve. Ezért az összefüggő komponensek száma a vetítés után eggyel csökken. De az indukciós feltevést a vetületekre alkalmazva, $k-1$ dimenzióban bennük már van k összefüggő komponens. Ezek összefüggőek maradnak k-dimenzióban is. Viszont a vetítés során, mint láttuk, az összefüggő komponensek száma eggyel csökkent, tehát eredetileg $k+1$ volt. $\square$

7.2. Lemma. Legyenek $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^k$ tetszőleges pontok a $\sum_{j=1}^n \mathbf{x}_j = \mathbf{0}$ és a $\sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T = \mathbf{I}_k$ kényszerfeltételekkel. Ezek kiszínezhetők $k+1$ különböző színnel

úgy, hogy a különböző színűek közti minimális távolság legalább $\dfrac{2\sqrt{3}}{\sqrt{n(n^2-1)}}$.

Bizonyítás. Az előbbi lemmát alkalmazzuk $\Delta = d_n$-nel. Azt kell csak belátnunk, hogy a $\mathcal{T}$ tulajdonság teljesül ezzel a Δ választással. Ui. legyen egy tetszőlegesen választott egyenes irányvektora $\mathbf{f}$. Az általánosság megszorítása nélkül feltehető, hogy $\|\mathbf{f}\| = 1$. Akkor a vetületek az $x_j = \mathbf{f}^T\mathbf{x}_j$ pontok lesznek, melyekre az eredeti pontokra tett kényszerfeltételek miatt

$$\sum_{j=1}^{n} x_j = 0 \quad \text{és} \quad \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} \mathbf{f}^T(\mathbf{x}_j\mathbf{x}_j^T)\mathbf{f}^T = \|\mathbf{f}\|^2 = 1.$$

Jelölje $x_1^*, \ldots, x_n^*$ a nagyság szerint rendezett $x_1, \ldots, x_n$ egydimenziós pontokat és legyen

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

Ezzel

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

Így $\delta^2 \ge d_n^2$, amiből már következik, hogy $\Delta = d_n$ jó választás.

A 3.4. Tétel bizonyítása.

A felső becslés:

Legyen $(V_1^*, \ldots, V_k^*)$ a minimális súlyozott k-sűrűséget adó partíciója a V csúcshalmaznak, legyen $|V_i^*| = n_i^*$, $(i = 1, \ldots, n)$. Definiáljuk a csúcsok következő k-dimenziós euklideszi reprezentációját:

$$x_j(i) := \begin{cases} \frac{1}{\sqrt{n_i^*}}, & \text{ha } v_j \in V_i^* \\ 0 & \text{különben,} \end{cases}$$

ahol $x_j(i)$ jelöli az $\mathbf{x}_j$ reprezentáns i-edik koordinátáját. Ezekre a reprezentánsokra $\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j^T = \mathbf{I}_k$ szintén teljesül. Az 1. fejezet jelöléseivel az e él varianciája ebben a reprezentációban

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

ahol $a_i^*(e) = |e \cap V_i^*|$, $(i = 1, \ldots k)$. Mivel $\sum_{e\in E} L(e, \mathbf{X}) = u(V_1^*, \ldots, V_k^*) = \nu_k(H)$ és $\sum_{j=1}^{k}\lambda_j = L(\mathbf{X}^*)$ – ahol $\mathbf{X}^*$ az optimális reprezentációja H-nak –, az optimalitás miatt $\sum_{j=1}^{k}\lambda_j \le \nu_k(H)$.

Az alsó becslés:

1. A költség bármely reprezentációban monoton: az $e' \subseteq e$ relációból következik, hogy $L(e', \mathbf{X}) \leq L(e, \mathbf{X})$. Ez egyszerű geometriai meggondolásokkal adódik. Részletesen ld. [6]-ban.
2. Ha $e = \{v_i, v_j\}$, akkor az (1.5) összefüggéssel $L(e, \mathbf{X}) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2$ bármely $\mathbf{X}$ reprezentációban.
3. Tekintsük az optimális k-dimenziós euklideszi reprezentációt adó $\mathbf{x}_1^*, \ldots \mathbf{x}_n^*$ vektorokat (az első koordináták – mivel egyenlők – el is hagyhatók, a maradék $\hat{\mathbf{x}}_j^*$ vektorok teljesítik a $\sum_{j=1}^n \hat{\mathbf{x}}_j^* (\hat{\mathbf{x}}_j^*)^T = \mathbf{I}_{k-1}$ kényszert). A 7.2. Lemma szerint $\hat{\mathbf{x}}_j^*$-ok, és így $\mathbf{x}_j^*$-ok is kiszínezhetők k különböző színnel úgy, hogy a különböző színűek közti távolság legalább d_n. Jelölje $(V_1, \ldots V_k)$ az ez által a színezés által indukált k-partíciót, az általa generált vágást pedig jelölje $H(V_1, \ldots, V_k)$. Egy vágásbeli e él tartalmaz egy kétszínű $e' = \{v_i, v_j\}$ élet, így az 1. és 2. rész eredményeit alkalmazva

$$L(e, \mathbf{X}^*) \geq L(e', \mathbf{X}^*) = \frac{\|\mathbf{x}_i^* - \mathbf{x}_j^*\|^2}{2} \geq \frac{d_n^2}{2}$$

adódik, ahonnan $c_n = \frac{d_n^2}{2} = \frac{6}{n(n^2-1)}$. Mivel azonban $\mathbf{X}^*$ az optimális k-dimenziós euklideszi reprezentáció volt, a 1.2 reprezentációs tétel alapján adódik, hogy

$$\sum_{j=1}^{k} \lambda_j = \sum_{e \in E} L(e, \mathbf{X}^*) \geq \sum_{e \in H(V_1, \ldots, V_k)} L(e, \mathbf{X}^*) \geq c_n |H(V_1, \ldots, V_k)| \geq c_n \theta_k(H).$$

A 3.7. Tétel bizonyítása. Legyen $\varepsilon < \frac{1}{2\sqrt{n}}$ és legyen $P_k = (V_1, \ldots V_k)$ a csúcsok olyan jól-szeparált k-partíciója az $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)$ optimális k-dimenziós euklideszi reprezentációban, hol a klaszterátmérők ε-nál kisebbek. (Az első koordináták egyenlősége miatt ezek helyett megint csak a $\hat{\mathbf{x}}_j^*$ $(k-1)$-dimenziós pontokat tekintjük, a metrikus távolságok ugyanazok.) A skalárszorzat folytonossága miatt feltehetjük, hogy a 3.7.tétel feltételei mellett van k olyan "centrum", hogy a reprezentánsok az ezek körüli ε sugarú gömbökben tömörülnek, és – $\mathbf{y}(\hat{\mathbf{x}}_j^*)$-gal jelölve az $\hat{\mathbf{x}}_j^*$-hoz legközelebb eső "centrumot" – a $\sum_{j=1}^n \mathbf{y}(\hat{\mathbf{x}}_j^*)\mathbf{y}^T(\hat{\mathbf{x}}_j^*) = \mathbf{I}_{k-1}$ összefüggés a "centrumokra" is fennáll. Mivel a "centrumok" közt k különböző van (jelölje ezeket $\mathbf{y}_1, \ldots, \mathbf{y}_k$) ezekre $\sum_{i=1}^k n_i \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}_{k-1}$ teljesül, ahol $n_i = |V_i|$, $\sum_{i=1}^k n_i = n$, továbbá ez a feltétel $\mathbf{y}_i$-ket egyértelműen meghatározza:

$$y_i(l) = \begin{cases} \frac{1}{n_i}, & \text{ha } i = l, \\ 0, & \text{különben.} \end{cases}$$

Itt az argumentumban a koordinátát jelöltük. Jelölje $\mathbf{Y}(\mathbf{X}^*)$ az $\mathbf{y}(\hat{\mathbf{x}}_j^*)$ vektorok által meghatározott $(k-1)$-dimenziós euklideszi reprezentációját a csúcsoknak. Könnyen látható, hogy $\sum_{e \in E} L(e, \mathbf{Y}(\mathbf{X}^*)) = u(P_k)$, így kétszer is alkalmazva az

(1.5) összefüggést a következő adódik:

$$\begin{aligned}\nu_k(H) &\le \sum_{e\in E} L(e,\mathbf{Y}(\mathbf{X}^*)) = \sum_{e\in H(P_k)} L(e,\mathbf{Y}(\mathbf{X}^*)) = \\ &= \sum_{e\in H(P_k)} \frac{1}{|e|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|\mathbf{y}(\hat{\mathbf{x}}_i^*)-\mathbf{y}(\hat{\mathbf{x}}_j^*)\|^2 \le \\ &= q^2\sum_{e\in H(P_k)} \frac{1}{|e|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|\hat{\mathbf{x}}_i^*-\hat{\mathbf{x}}_j^*\|^2 = \\ &= q^2\sum_{e\in H(P_k)} \frac{1}{|e|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{I}(v_i\in e)\mathcal{I}(v_j\in e)\|\mathbf{x}_i^*-\mathbf{x}_j^*\|^2 = \\ &= q^2\sum_{e\in H(P_k)} L(e,\mathbf{X}^*) \le= q^2\sum_{e\in E} L(e,\mathbf{X}^*) == q^2\sum_{j=1}^{k}\lambda_j.\end{aligned}$$

Itt q onnan határozható meg, hogy a legkisebb távolság a "centrumok" közt

$$\delta = \min\sum_{l=1}^{k} n_l = n\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \ge \frac{2}{\sqrt{n}},$$

továbbá az $\hat{\mathbf{x}}_i^*$ ill. $\hat{\mathbf{x}}_j^*$ reprezentánsok az $\mathbf{y}(\hat{\mathbf{x}}_i^*)$ ill. $\mathbf{y}(\hat{\mathbf{x}}_j^*)$ centrumok körüli ε-sugarú gömbökben ülnek, és ezért

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

7.3. Lemma. Rendelkezzen az $n\times n$-es, szimmetrikus $\mathbf{B}$ mátrix a következő tulajdonsággal: van olyan k-dimenziós $F\subset R^n$ altér, hogy $\mathbf{x}^T\mathbf{B}\mathbf{x}\notin(a,b)$ minden $\mathbf{x}\in F$ ($\|\mathbf{x}\|=1$) vektorra, ahol $a<b$ valós számok (esetleg $a=-\infty$ vagy $b=\infty$). Akkor $\mathbf{B}$-nek lagalább k sajátértéke az (a,b) intervallumon kívül esik.

Bizonyítás. Jelölje m a $\mathbf{B}$ mátrix (a,b)-beli sajátértékeinek számát, H pedig a hozzájuk tartozó sajátvektorok által kifeszített alteret. Akkor

$$k+m = \dim(F)+\dim(H)\le n,$$

amivel be is bizonyítottuk az állítást, hiszen a $\mathbf{B}$ mátrix (a,b)-n kívüli sajátértékeinek száma $n-m\ge k$. □

7.4. Következmény. Az előbbi jelölésekkel,

[a] ha van olyan k-dimenziós $F\subset R^n$ altér, hogy $\mathbf{x}^T\mathbf{B}\mathbf{x}\ge a$ minden $\mathbf{x}\in F$ ($\|\mathbf{x}\|=1$) vektorra, akkor $\mathbf{B}$-nek legalább k db. legalább a-nyi sajátértéke van;

[b] ha van olyan k-dimenziós $F\subset R^n$ altér, hogy $\mathbf{x}^T\mathbf{B}\mathbf{x}\le b$ minden $\mathbf{x}\in F$ ($\|\mathbf{x}\|=1$) vektorra, akkor $\mathbf{B}$-nek legalább k db. legfeljebb b-nyi sajátértéke van;

[c] ha van olyan k-dimenziós $F\subset R^n$ altér, hogy $\mathbf{x}^T\mathbf{B}\mathbf{x}\in[a,b]$ minden $\mathbf{x}\in F$ ($\|\mathbf{x}\|=1$) vektorra, akkor $\mathbf{B}$-nek legalább k db. sajátértéke van $[a,b]$-ben. □

7.5. Lemma. Legyen az $n \times n$-es, szimmetrikus, pozitív szemidefinit $\mathbf{B}$ mátrix 0 sajátértéke k multiplicitású, a pozitív sajátértékekre pedig legyen $\varrho > 0$ alsó korlát. Legyen a $\mathbf{P}$ $n \times n$-es pozitív szemidefinit "perturbációs" mátrix normája $\|\mathbf{P}\| = \varepsilon$. Akkor az $n \times n$-es pozitív szemidefinit $\mathbf{C} := \mathbf{B} + \mathbf{P}$ mátrixnak legalább k db. legfeljebb ε-nyi sajátértéke van. Továbbá $\mathbf{y}_1, \ldots, \mathbf{y}_k$-val jelölve a k legkisebb sajátértékhez tartozó sajátvektorokat, és felbontva őket

$$\mathbf{y}_i = \mathbf{u}_i + \mathbf{z}_i, \qquad \mathbf{u}_i \in F, \quad \mathbf{z}_i \perp F, \tag{7.1}$$

alakban, ahol F a $\mathbf{B}$ magtere, a merőleges komponensekre

$$\|\mathbf{z}_i\|^2 \leq \frac{1}{\varrho}\varepsilon, \quad (i = 1, \ldots, k)$$

teljesül.

Bizonyítás. Legyen $\mathbf{u} \in F$, $\|\mathbf{u}\| = 1$ tetszőleges. Akkor egyrészt

$$\mathbf{u}^T\mathbf{C}\mathbf{u} = \mathbf{u}^T\mathbf{P}\mathbf{u} \leq \|\mathbf{P}\| \cdot \|\mathbf{u}\|^2 \leq \varepsilon. \tag{7.2}$$

Mivel F k-dimenziós altere R^n-nek, a 7.2 következmény [b] része alapján a $\mathbf{C}$ mátrixnak legalább k db. ε-nál nem nagyobb sajátértéke van.

Másrészt, bármely $\mathbf{y} \in R^n$ vektor egyértelműen felbontható $\mathbf{y} = \mathbf{u} + \mathbf{z}$ alakban, ahol $\mathbf{u} \in F$ és $\mathbf{z} \perp F$. Tehát

$$\mathbf{y}^T\mathbf{C}\mathbf{y} = \mathbf{y}^T\mathbf{B}\mathbf{y} + \mathbf{y}^T\mathbf{P}\mathbf{y} = \mathbf{z}^T\mathbf{B}\mathbf{z} + \mathbf{y}^T\mathbf{P}\mathbf{y} \geq \varrho\|\mathbf{z}\|^2. \tag{7.3}$$

Legyen $\mathbf{y}_1, \ldots, \mathbf{y}_k$ a $\mathbf{C}$ mátrix k legkisebb sajátértékéhez tartozó ortonormált sajátvektorrendszer. Akkor (7.2) és (7.3) szerint

$$\varepsilon \geq \mathbf{y}_i^T\mathbf{C}\mathbf{y}_i \geq \varrho\|\mathbf{z}_i\|^2, \quad (i = 1, \ldots, k)$$

teljesül, amivel be is láttuk az állítást. □

7.6. Lemma. Legyen $\mathbf{A}$ rögzített $k \times k$-as mátrix, $\mathbf{R}$ pedig egy $k \times k$-as ortogonális mátrix. $\operatorname{tr}\mathbf{A}\mathbf{R}$ akkor lesz maximális, ha $\mathbf{A}\mathbf{R}$ szimmetrikus. Ezesetben a maximum az $\mathbf{A}$ mátrix szinguláris értékeinek összege. (A bizonyítást ld. [5] 67. oldalán.)

7.7. Lemma. Legyen $F \subset R^n$ a $\mathbf{B}$ mátrix magtere (az előzőek szerint F dimenziója k) és $\mathbf{y}_1, \ldots, \mathbf{y}_k$ egy tetszőleges R^n-beli ortonormált vektorrendszer. Bontsuk fel az $\mathbf{y}_i$ vektorokat

$$\mathbf{y}_i = \mathbf{v}_i + \mathbf{z}_i, \qquad \mathbf{v}_i \in F, \quad \mathbf{z}_i \perp F, \quad (i = 1, \ldots, k)$$

alakban. Akkor létezik az F altéren belül olyan $\mathbf{u}_1, \ldots, \mathbf{u}_k$ ortonormált rendszer, hogy

$$\sum_{i=1}^{k} \|\mathbf{y}_i - \mathbf{u}_i\|^2 \leq 2\sum_{i=1}^{k} \|\mathbf{z}_i\|^2.$$

Bizonyítás. Jelölje $\mathbf{Y}$ és $\mathbf{U}$ az $\mathbf{y}_i$ és $\mathbf{u}_i$ vektorokból, mint oszlopvektorokból alkotott $n \times k$-as mátrixokat, ahol $\mathbf{u}_1, \dots, \mathbf{u}_n$ tetszőleges F-beli ortonormált rendszer. Azt kell megmutatnunk, hogy van olyan $k \times k$-as ortogonális $\mathbf{R}$ mátrix, hogy

$$\|\mathbf{Y} - \mathbf{UR}\|^2 \le 2\Delta,$$

ahol $\Delta := \sum_{i=1}^{k} \|\mathbf{z}_i\|^2$. A bal oldalt $L(\mathbf{R})$-rel jelölve,

$$\begin{aligned} L(\mathbf{R}) &= \operatorname{tr}(\mathbf{Y} - \mathbf{UR})^T(\mathbf{Y} - \mathbf{UR}) \\ &= \operatorname{tr}\mathbf{Y}^T\mathbf{Y} + tr\mathbf{R}^T\mathbf{U}^T\mathbf{UR} - 2\operatorname{tr}\mathbf{Y}^T\mathbf{UR} \\ &= \operatorname{tr}\mathbf{Y}^T\mathbf{Y} + \operatorname{tr}(\mathbf{U}^T\mathbf{U})(\mathbf{R}\mathbf{R}^T) - 2\operatorname{tr}\mathbf{Y}^T\mathbf{UR} = 2(k - \operatorname{tr}\mathbf{Y}^T\mathbf{UR}). \end{aligned}$$

Másrészt $L(\mathbf{R})$ minimális, ha $\operatorname{tr}\mathbf{Y}^T\mathbf{UR}$ maximális.

A 7.6 lemmát az $\mathbf{Y}^T\mathbf{U}$ mátrixra alkalmazzuk:

$$\min_{\mathbf{R}\text{ortogonális}} L(\mathbf{R}) = 2\sum_{i=1}^{k}(1 - s_i),$$

ahol $0 \le s_1 \le \cdots \le s_k$ az $\mathbf{Y}^T\mathbf{U}$ mátrix szinguláris értékei. Másrészt

$$\begin{aligned} \Delta = \|\mathbf{Y} - \mathbf{U}\mathbf{U}^T\mathbf{Y}\|^2 &= \operatorname{tr}(\mathbf{Y} - \mathbf{U}\mathbf{U}^T\mathbf{Y})^T(\mathbf{Y} - \mathbf{U}\mathbf{U}^T\mathbf{Y}) \\ = \operatorname{tr}\mathbf{Y}^T\mathbf{Y} - \operatorname{tr}\mathbf{Y}^T\mathbf{U}\mathbf{U}^T\mathbf{Y} &= k - \|\mathbf{Y}^T\mathbf{U}\|^2 = \sum_{i=1}^{k}(1 - s_i^2). \end{aligned}$$

Már csak azt kell belátnunk, hogy $1 - s_i \le 1 - s_i^2$, $(i = 1, \dots, k)$. Ez az

$$s_i \le s_k \le s_k(\mathbf{Y}) \cdot s_k(\mathbf{U}) = 1,$$

összefüggésből következik, mivel mind $\mathbf{Y}$, mind $\mathbf{U}$ legnagyobb szinguláris értéke 1 (sőt, mivel oszlopaik ortogonormáltak, pontosan k db. 1-gyel egyenlő szinguláris értékük van). □

7.8. Következmény. A 7.1. 7.3. és 7.7 lemmák jelölései és feltételei mellett van olyan $\mathbf{u}_1, \dots, \mathbf{u}_n \in F$ ortonormált rendszer, hogy

$$\sum_{i=1}^{k} \|\mathbf{y}_i - \mathbf{u}_i\|^2 \le 2k\frac{\varepsilon}{\varrho}.$$

A 4.2. Tétel bizonyítása. Rögzítsük a P_k partíciót. Legyen $F \subset R^n$ a $\mathbf{B}$ mátrix k-dimenziós magtere. Jelölje $\mathbf{y}_1, \dots \mathbf{y}_k$ a $\mathbf{C} = \mathbf{B} + \mathbf{P}$ Laplace-mátrix k legkisebb sajátértékéhez tarozó ortonormált sajátvektorrendszert. Mivel a $\mathbf{C}$ mátrix $\mathbf{y}_1, \dots, \mathbf{y}_k$-hoz tartozó sajátértékei legfeljebb ε-nyiak, a 7.6 lemma alkalmazásával

$$d^2(\mathbf{y}_i, F) \le \frac{\varepsilon}{\varrho}, \quad (i = 1, \dots, k).$$

Összegezve $i = 1, \dots, k$-ra, a bizonyítás kész. □

A 4.4. Tétel bizonyítása. A feltételek miatt a súlyok összegére

$$\sum_{i=1}^{n} d_i = 1. \tag{7.4}$$

A súlyozott csúcsú és élű gráfok definíciója miatt (ld. 4. fejezet) a λ_2 sajátértékhez tartozó $\mathbf{u}$ sajátvektor koordinátáira a

$$\sum_{i=1}^{n} d_i u_i = 0 \quad \text{és} \quad \sum_{i=1}^{n} d_i u_i^2 = 1$$

feltételek teljesülnek. Most konstruálunk egy olyan $\mathbf{y} \in R^n$ vektort (y_i koordinátákkal), amely kielégíti a következő feltételeket:

$$\sum_{i=1}^{n} d_i y_i = 0 \tag{7.5}$$

és

$$\sum_{i=1}^{n} d_i u_i y_i = 0 \tag{7.6}$$

A következő alakban keressük $\mathbf{y}$-t:

$$y_i := |u_i - a| - b, \quad (i = 1, \ldots, n) \tag{7.7}$$

ahol a és b valós számok.

Csak egzisztenciát bizonyítunk. Megmutatjuk, hogy léteznek olyan a and b számok, hogy a velük előállított $\mathbf{y}$-ra a (7.5) és (7.6) feltételek teljesülnek. Érvelésünk a következő: tegyük fel, hogy már megtaláltuk a-t. Akkor (7.4) és (7.5) miatt b-re

$$b = \sum_{i=1}^{n} d_i |u_2(i) - a|. \tag{7.8}$$

adódik. Ezzel a b választással a (7.6) feltétel teljesítése azt jelenti, hogy

$$\sum_{i=1}^{n} d_i u_2(i) |u_2(i) - a| = 0.$$

Mivel a bal oldal a folytonos függvénye és 1-gyel egyenlő, ha $a \leq \min_i u_i$, -1-gyel pedig, ha $a \geq \max_i u_i$, ezért a Bolzano—Weierstrass tétel miatt a bal oldali függvénynek van legalább egy gyöke $\min_i u_i$ és $\max_i u_i$ között. Nevezzünk ki egy ilyen gyököt a-nak, ezután a megfelelő b (7.8) szerint adódik, $\mathbf{y}$ koordinátáit pedig egyértelműen meghatározza a (7.7) összefüggés. Legyen $c_1 := a - b$ és $c_2 := a + b$. Könnyen látható, hogy

$$y_i = |u_i - a| - b = \begin{cases} c_1 - u_i, & \text{ha} \quad u_i < a, \\ u_i - c_2, & \text{ha} \quad u_i \geq a, \end{cases}$$

ezért

$$|y_i| = \min\{|u_i - c_1|, |u_i - c_2|\} \tag{7.9}$$

teljesül minden i-re. Jelölje

$$\sigma^2(\mathbf{y}) := \sum_{i=1}^{n} d_i y_i^2$$

az $\mathbf{y}$ vektor koordinátáinak varianciáját. Mivel az $\mathbf{u}$ vektor 2-varianciája

$$S_2^2(\mathbf{u}) = \min_{c,\alpha} \sum_{i=1}^{n} d_i [u_i - c_{\alpha_i}]^2, \tag{7.10}$$

ahol u_i jelöli az $\mathbf{u}$ vektor i-edik koordinátáját, továbbá $c_{\alpha_i} \in R$ és α_i vagy 1 vagy 2 (annak megfelelően, hogy melyik klasztercentrumtól való távolságot nézzük), ezért $\sigma^2(\mathbf{y})$ egyike azoknak a (7.10)-beli kifejezéseknek, amelyek minimuma $S_2^2(\mathbf{u})$. Így $\sigma^2(\mathbf{y}) \geq S_2^2(\mathbf{u})$.

$\sigma(\mathbf{y}) = 0$ esetén $S_2^2(\mathbf{u})$ szintén 0, amikoris a tétel állítása triviálisan teljesül. Ezért feltehető, hogy $\sigma(\mathbf{y}) > 0$. Legyen $z_i := y_i/\sigma(y)$, $(i = 1, \ldots, n)$ és jelölje $\mathbf{z}$ a z_i-kből, mint koordinátákból álló vektort, $\mathbf{x}_i$ pedig legyen az a 2-dimenziós vektor, melynek első koordinátája u_i, második pedig z_i. Jelölje továbbá $\mathbf{X}$ az $\mathbf{u}$ és $\mathbf{z}$ vektorokat soraiban tartalmazó $2 \times n$-es mátrixot, míg $\mathbf{X}^*$ az $\mathbf{u}$ és $\mathbf{v}$ vektorokat hasonlóan tartalmazó $2 \times n$-es mátrixot, ahol $\mathbf{v}$ a $\mathbf{C}_D$ Laplace-mátrix λ_3 sajátértékéhez tartózó sajátvektorának transzformáltja (a 4. fejezetbeli feltételek szerint). Akkor egyrészt

$$\max_{u_i \neq u_j} \frac{|z_i - z_j|}{|u_i - u_j|} \leq \frac{1}{\sigma(\mathbf{y})}, \tag{7.11}$$

mivel $y(i)$ definíciójából következik, hogy

$$|y_i - y_j| \leq |u_i - u_j|, \quad (i \neq j),$$

azaz $\mathbf{y}$ (mint $\mathbf{u}$ függvénye) teljesíti a Lipschitz-feltételt. Másrészt a sajátértékek extremális tulajdonsága (ld. [15]-ben) és a (4.1) összefüggés alapján a következő becslés végezhető:

$$\begin{aligned}
\frac{\lambda_1 + \lambda_2}{\lambda_1} &= \frac{\operatorname{tr} \mathbf{X}^* \mathbf{C} \mathbf{X}^{*T}}{\mathbf{u}^T \mathbf{C} \mathbf{u}} \leq \frac{\operatorname{tr} \mathbf{X} \mathbf{C} \mathbf{X}^T}{\mathbf{u}^T \mathbf{C} \mathbf{u}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} (u_i - u_j)^2} \\
&= \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \left[(u_2(i) - u_2(j))^2 + (z(i) - z(j))^2 \right]}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} (u_i - u_j)^2} \\
&\leq 1 + \max_{u_i \neq u_j} \frac{(z_i - z_j)^2}{(u_i - u_j)^2} \leq 1 + \frac{1}{\sigma^2(\mathbf{y})} \leq 1 + \frac{1}{S_2^2(\mathbf{u})},
\end{aligned}$$

amelyből átrendezéssel rögtön adódik a kívánt állítás. □

A 4.5. Sejtés bizonyításához szükséges lenne a következő lemmára:

7.9. Lemma. Van olyan $y_i = f(\mathbf{x}_i^*)$ transzformáció, melyre az f függvény teljesíti a Lipschitz-feltételt, $\sum_{i=1}^{n} d_i y_i = 0$, $\sum_{i=1}^{n} d_i \mathbf{x}_i^* y_i = \mathbf{0}$ és $\sigma^2(\mathbf{y}) := \sum_{i=1}^{n} d_i y_i^2 \geq S_{k+1}^2(\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)$.

Azt gondoljuk, hogy ilyen $\mathbf{y}$ konstruálható $\sqrt{k}$ Lipschitz-konstanssal. Ezután a sejtés már bizonyítható lenne (ld. [6]-ban).

Irodalom

[1] Alon, N., Eigenvalues and Expanders. Combinatorica **6** (2) (1986), 83-96.

[2] Biggs, N.L., Algebraic Graph Theory. Cambridge University Press, Cambridge 1974.

[3] Bolla, M., Spectra, Euclidean Representations and Vertex-Colorings of Hypergraphs. Discrete Mathematics 117 (1993), 13-39.

[4] Bolla, M., Tusnády, G., Spectra and Optimal Partitions of Weighted Graphs. Discrete Mathematics 128 (1994), 1-20.

[5] Bolla, M., Mátrixok szinguláris felbontásának módszerei és statisztikai alkalmazásai. Egyetemi doktori értekezés és MTA SZTAKI Working Paper MS/11, Budapest, 1982.

[6] Bolla, M., Relations between spectral and classification properties of multigraphs. Kandidátusi értekezés, Budapest, 1993.

[7] Cvetković, D.M., Doob, M., Sachs, H., Spectra of Graphs. Academic Press, New York-San Francisco-London, 1979.

[8] Dunn, J.C., Well-Separated Clusters and Optimal Fuzzy Partitions. J. Cybernetics **4**, No.1 (1974), 95-104.

[9] Dunn, J.C., Some Recent Investigations of a New Fuzzy Partitioning Algorithm and its Application to Pattern Classification Problems. **4**, No.2 (1974), 1-23.

[10] Fiedler, M., Algebraic Connectivity of Graphs. Czechoslovak Math. J. 23 (1973), 298-305.

[11] Hoffman, A.,J., On Eigenvalues and Colorings of Graphs. In: Graph Theory and Its Applications (ed. B. Harris), Academic Press, New York-London (1970), 79-91.

[12] Juhász, F., Mályusz, K., Problems of Cluster Analysis from the Viewpoint of Numerical Analysis. In: Proc. Conf. Numerical Methods, Keszthely, 1977.

[13] Lengyel, T., A klaszteranalízis néhány kombinatorikai és valószínűségszámítási problémája. MTA SZTAKI Tanulmányok, Budapest, No. 188, 1986, 1-173.

[14] Mac Queen, J., Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5-th Berkeley Symp. Math. Statist. Prob. **1** (1967), 281-297.

[15] Rao, C.R., Separation Theorems for Singular Values of Matrices and their Applications in Multivariate Analysis. J. Multivariate Analysis **9** (1979), 362-377.

[16] Tusnády, G., Sztochasztikus számítástechnika. Debrecen, KLTE, egyetemi jegyzet, 1996-1997.

Marianna Bolla, Gábor Tusnády:
Investigating the connectivity of hypergraphs via their spectra

Some clustering properties of hypergraphs are investigated by means of linear algebraic tools. The notion of the Laplacian is generalized for hypergraphs and we conclude for the connectivity of the hypergraph by means of its spectrum. Some conclusions for the number of clusters are also made. The clusters themselves can be constructed by means of the Euclidean representation of the hypergrap via the eigenvectors of its Laplacian. An iteration algorithm is also introduced.