MTA DOKTORI ÉRTEKEZÉS TÉZISEI

Gráfok és kontingenciatáblák klaszterezése spektrális módszerekkel

BOLLA MARIANNA

BME TTK Matematika Intézet, Sztochaszika Tanszék

Budapest, 2016.

I. A kitűzött kutatási feladat

A spektrális klaszterezés az 1990-es években elterjedt fogalom arra, hogy adatpontokat vagy egy gráf csúcsait osztályozzuk egy megfelelően konstruált mátrix sajátértékei és sajátvektorai segítségével. A bevezető irodalomban (pl. [Lux]) azonban csak gyakorlati útmutatások találhatók arra nézve, hogyan építsünk hasonlósági mátrixot az adatokra vagy a gráfra, és heurisztikus algoritmusokat definiálnak. Nem mondanak sokat a célhoz adaptált mátrix választásáról vagy a kapott optimális klaszterezés minőségéről, egyáltalán a klaszterek számáról. Talán Ravi Kannan (Microsoft Research, India) videoelőadása (Simons Institute, Berkeley, 2013. December 9.) világít rá legjobban az elérendő célra. Idemásolom Clustering – Does Theory Help? című előadásának kivonatát: „*Theoretical Computer Science has brought to bear powerful ideas to find nearly optimal clusterings, while Statistics mixture models of data have been useful in understanding the structure of data and in developing clustering algorithms. However, in practice many heuristics (e.g., dimension reduction and the k-means algorithm) are widely used. The talk will describe some aspects of the Theoretical Computer Science and Statistics approaches, and attempt to answer the question: is there a happy marriage of these approaches with practice?*" Jelen dolgozatban mindkét megközelítést használom, és megpróbálom azokat összeegyeztetni a gyakorlati alkalmazók igényeivel. Mivel kutatási célom hosszú idő folyamán alakult ki és módosult is közben, először szeretnék rövid áttekintést adni a témáról és kapcsolatomról azzal.

Az 1980-as évek végén kandidátusi dolgozatom témavezetőjével, Tusnády Gáborral spektrális módszereket használtunk egy, a veleszületett rendellenességek vizsgálata kapcsán felmerült bináris klaszterezési probléma megoldására (Czeizel Endre akkori adatain). A SZTAKI-ban Prékopa András és Juhász Ferenc felhívták figyelmemet arra, hogy a vizsgálatainkban kulcsszerepet játszó mátrix az egyszerű gráfokra jól ismert Laplace-mátrix természetes általánosítása lehet hipergráfokra. Ezután öszefüggéseket állapítottunk meg a Laplace-mátrix spektruma és a hipergráf klaszteresedését kifejező vágások mérőszámai közt. Később beláttam, hogy minden hipergráfhoz hozzárendelhető egy élsúlyozott gráf, melynek Laplace-mátrixa azonos a hipergráféval. Bevezettük a Laplace-mátrix fokszámokkal normált változatát is, melynek sajátértékei egyértelműen megfeleltethetők a bolyongásoknál vizsgát átmenetvalószínűség mátrix sajátértékeinek. Az ehhez köthető spektrális rés (az átmenetvalószínűség mátrix triviális 1 és második legnagyobb, vagy ami ezzel ekvivalens, a Laplace-mátrix 0 és második legkisebb sajátértéke közt) és az izoperimetrikus szám kapcsolatát kifejező Cheeger-egyenlőtlenség már régóta ismert volt. Azonban a témakörhöz kapcsolódó cikkek egyértelműen az első nem-triviális sajátértékre fókuszáltak, amelynek, ha nagy a triviálistól való elválása, akkor a gráf jó expander, és hasonló viselkedést mutat, mint az Erdős–Rényi típusú véletlen gráf (egyosztályos eset, kvázirandomság). Mindez megfogalmazható a gráfon az élsúlyokkal arányos valószínűségekkel történő véletlen bolyongás keverési és lefedési ideje, továbbá rezisztencia, konduktancia segítségével is, melyről áttekintést nyújt pl. [Lov93, Chu]. Ha az elválás a triviális sajátértéktől nem nagy, hanem utána van a rés, akkor jön képbe a Laplace-mátrix legkisebb pozitív sajátértékéhez tartozó sajátvektor, az ún. Fiedler-vektor, melynek koordinátái alapján két, egymással „lazán" összefüggő klaszter bontakozik ki, l. [Fid72, Fid73, Hof72, Juh-Mály]. Mondhatnánk, hogy a csúcsok Fiedler-vektor alapján történő kettéosztása folytatható az osztályokon belül, azonban engem ez nem elégített ki. Szerettem volna a spektrum belsejében található rések alapján k elváló sajátérték segítségével megtalálni a csúcsok optimális k-partícióját, mely minimalizálja a klaszterpárok közti súlyozott vágások összegét. Ehhez az elváló sajátértékekhez tartozó sajátvektorok alapján konstruált ún. csúcs-reprezentánsokat és azokra a k-közép algoritmust, ill. annak súlyozott változatát használtuk [Bol93, Bol-Tus94]. Közben a spektrális klaszterezés elterjedt, ekkor vezették be a fogalmat is, és rengeteg cikk született a témában: sokan újra definiálták (partition cut, normalized cut néven) az általunk már bevezetett mennyiségeket, és hasonló vizsgálatokat folytattak a spektrummal való kapcsolatukkal, de vizsgálataik néhány kivétellel (pl. [Ng-Jo-We]) csak a sajátértékekkel való alsó becslésre terjedtek ki. A reprezentációs technika, mint *spektrális relaxáció* vonult be a köztudatba anélkül, hogy a klaszterek számára vonatkozó kritériumokat vagy a k-közép algoritmus célfüggvényének kapcsolatát a minimalizálandó többszempontú vágásokkal vizsgálták volna. Így az 1990-es évek végén célul tűztem ki, hogy általánosabb optimalizálási feladatokkal (pl. [Boletal98]) és matematikus hallgatók bevonásával a többszempontú vágások felső becslésével foglalkozom. Az ezredfordulón egyéb szempontból is új lendületet vettek a dolgok, így újabb célok fogalmazódtak meg bennem:

- A Word Wide Web rohamos bővülésével fizikusok a klasszikus Erdős–Rényi modelltől [Erd-Reny] eltérő modelleket fedeztek fel, melyekben a sajátértékek empirikus eloszlása eltér a megszokott Wigner-féle félkörtől. Egyéb *sztochasztikus blokkmodellek* is előtérbe kerültek evolválódó szociális és biológiai hálózatok leírására, elsősorban nem spektrális módszerekkel, l. [Hol-Las-Lei, McSh]. Ezért véletlen perturbációk hatását kezdtem vizsgálni azzal a reménnyel, hogy amennyiben megértem néhány speciális struktúra sajátértékeinek és a hozzájuk tartozó sajátaltereknek a viselkedését, akkor ilyen struktúrákat könnyebben fedezhetek fel valós életbeli adatokban.

- Ugyancsak fizikusok szociális hálózatok klasztereinek (náluk inkább modul vagy community) feltárására az ún. Newman–Girvan modularitást [New-Gir, New] maximalizálták, azonban nem állapítottak meg precíz összefüggéseket a modularitás-mátrix sajátértékei és a modularitást maximalizáló többszempontú vágások közt. Célul tűztem ki ennek vizsgálatát, illetve bevezettem a *normált modularitás-mátrix*ot is, melynek nagy abszolút értékű, ún. *strukturális sajátérték*ei alkalmasnak tűntek a klasztereken belüli és klaszterpárok közötti diszkrepanciák becslésére.

- Hatással volt rám Lovász Lászlónak és munkatársainak *tesztelhető gráfparaméterek*kel kapcsolatos elmélte [Borgsetal1, Borgsetal2]. Célul tűztem ki a minimális többszempontú vágások, továbbá a normált modularitás-mátrix strukturális sajátértékeinek és a hozzájuk tartozó sajátalterek tesztelhetőségének vizsgálatát. Ez lényegében azt jelenti, hogy bizonyos kiegyensúlyozottsági feltételek mellett a klaszterek ún. k-varianciája konzisztensen becsülhető.

- A humán genom projekt eredményeként az ezredfordulón a genetikai vizsgálatok középpontjába kerültek a *microarray*-ek, melyek statisztikai szempontból átskálázott kontingenciatáblák (téglalap alakú, nem-negatív elemű mátrixok), soraik a géneknek (ezek száma nagyon nagy), oszlopaik a vizsgálati feltételeknek (ezek száma sokkal kevesebb) felelnek meg, az egyes mátrixelemek pedig megmutatják, hogy az adott sorbeli gén az oszlopbeli feltétel mellett milyen mértékben van kifejezve (ez egy nem-negatív valós szám, bináris esetben 0 vagy 1). A biológusok célja általában a gének és feltételek jellegzetes kapcsolódási csoportjainak a megtalálása úgy, hogy az egy osztályba tartozó gének hasonlóan befolyásolják az egy osztályba tartozó feltételeket (pl. betegségeket), l. [Klugetal]. Elhatároztam, hogy *korrespondanciaanalízis* technikákat használok a sorok és oszlopok alacsony dimenziós reprezentációjához, és a reprezentánsok szimultán klaszterezésével nyerem ki a feltételeknek eleget tevő klaszterpárokat. Ez az ún. *spektrális biklaszterezés* a gráfokra kidolgozott eljárások természetes általánosítása, csak itt spektrális felbontás helyett szinguláris felbontást használunk. Tervbe vettem a módszer kiterjesztését irányított gráfokra, akárcsak a konvergencia fogalmát kontingenciatáblákra. Ugyancsak a microarrayek kapcsán merült fel bennem az igény *kis diszkrepanciájú (ún. reguláris) klaszterek és klaszterpárok* keresésére, melyek a kiugró szinguláris értékekkel (nagy abszolút értékű sajátértékek) átvihetők gráfokra is. A minimális és maximális többszempontú vágások speciális esetként adódnak, amennyiben a normált modularitás-mátrix nagy abszolút értékű sajátértékei mind pozitívak (a fizikusok nyelvén „community structure") vagy negatívak („anticommunity structure"). A marginálisok szerint normált kontingenciatáblára és a diszkrepancia vonatkozásában a biklaszterezési problémára nem találtam megoldásokat az irodalomban.

- A téglalap-, ill. szimmetrikus nem-negatív elemű, megfelelően normált mátrixok *diszkrét együttes eloszlások* speciális eseteinek is tekinthetők, a vizsgált normált modularitás-mátrix és kontingenciatábla pedig a *feltételes várható érték vevés* operátorával hozható kapcsolatba. Így általánosan (nem csupán a véges esetben) a Hilbert-terek közti kompakt lineáris operátorok elmélete használható az alacsony rangú reprezentáció hatékonyságát kifejező célfüggvény optimalizálására. Kapcsolatot kerestem a *Rényi-féle maximálkorreláció* [Reny59a, Reny59b] és az általunk vizsgált szinguláris értékek közt. Ezzel választ kívántam kapni arra, hogy adatpontok spektrális klaszterezésére hogyan használható a reprodukáló magú Hilbert-terek elmélete. Ezt a technikát napjainkban a modern képfelismerő eljárások intenzíven használják. A Hilbert-teres megközelítést a spektrális alterek tesztelhetőségének bizonyítására is használni fogom.

- Az ún. *expander mixing lemmát* (pl. [Alon, Ho-Lin-Wid]) szerettem volna kiterjeszteni téglalap elrendezésekre és a csúcsok ill. sorok/oszlopok legalább kettő klaszterére. A lemma élsúlyozott gráfokra, a fenti fogalmakkal elmondva, a csúcsok bármely két részhalmaza közti diszkrepanciát becsli felülről a normált modularitás-mátrix spektrálnormájával. A spektrum, diszkrepancia és egyéb fokszám-jellemzők közti ekvivalenciák képezik a régóta vizsgált ún. kvázirandom tulajdonságok alapjait a $k = 1$ esetben, l. [Thom87, Thom89, Bo] és [Chu-G-W, Chu-G]. Arra gondoltam, hogy ha sikerül bizonyítani az alkalmasan definiált *többrészes diszkrepancia* és spektrum közötti oda–vissza kapcsolatot, akkor definiálhatunk ún. *általánosított kvázirandom tulajdonságokat*, melyek közti implikácók érvényesek determinisztikus gráfsorozatokra is, függetlenül a sztochasztikus modelltől. Ehhez az általánosított kvázirandom gráfokat a gráfkonvergencia fogalmával definiáló [Lov-Sos] cikk is motivációt adott.

- Végső célom tetszőleges k pozitív egészre vizsgálni a megfelelő normált mátrix k legnagyobb abszolút értékű sajátértéke és a hozzájuk tartozó sajáltér kapcsolatát a gráf vagy tábla k-részes diszkrepanciájával, ami egy általános kritérium a klaszterezés homogenitásának mérésére. Ilyen módon a köztes esetet vizsgálom a $k = 1$-nek megfelelő expander mixing lemma és kvázirandomság, továbbá a kis diszkrepanciájú klaszterezést nagyon nagy (de univerzális) k-val általánosan garantáló Szemerédi regularitási lemma [Szem] közt. Spektrális módszerekkel k értékére és a klaszterek mibenlétére is válaszokat adok.

- Ezzel párhuzamosan olyan paraméteres statisztikai keverékmodellek vizsgálatát is célul tűztem ki, amelyekben a részgráfokra és a páros részgráfokra ismert logisztikus modellek alkalmazhatók (α-β modellek és a Rasch modell [Csetal1, Csetal2, Rasch]), a párhuzamos klaszterezésre és paraméterbecslésre pedig az EM (Expectation–Maximization) algoritmus [De-La-Ru].

A fenti kérdésekre a III. részben felsorolt eredmények választ adnak, csak az általánosított kvázirandom tulajdonságok közti implikációkat nem tudom maradéktalanul bizonyítani. Ezért sejtésként fogalmazom meg azokat, kivéve a normált modularitás spektrum és többrészes diszkrepancia közti oda-vissza állításokat. Összefoglalva azt gondolom, hogy a spektrum mindkét végét kell nézni. Amennyiben a normált modularitás-mátrix nagy pozitív (a normált Laplace-mátrix 0-hoz közeli) sajátértékei dominálnak, akkor a spektrális eszközökkel kapott klasztereken belül „szoros", a klaszterpárok közt pedig „laza" a csúcsok közötti kapcsolat. Ellenkező esetben, ha a normált modularitás-mátrix nagy abszolút értékű negatív (a normált Laplace-mátrix 2-höz közeli) sajátértékei dominálnak, akkor a spektrális eszközökkel kapott klasztereken belül „laza", köztük pedig „szoros" a csúcsok közötti kapcsolat. Ezeket Luca Trevisan és társszerzői [Gh-Trev, Le-Gh-Trev, Trev] utóbbi 6-7 évben publikált eredményei is alátámasztják, melyek az ún. magasabb rendű és duális Cheeger-egyenlőtlenséget vezetik be a többosztályos eset és a spekrum másik végének vizsgálatára. Fontos, hogy nem feltétlenül partíciókon optimalizálnak, a ritka vagy sűrű többszempontú vágásokkal kapcsolatba hozható klaszterek ui. náluk nem feltétlenül merítik ki a teljes csúcshalmazt. Miután én a spektrum mindkét végét egyszerre tekintem, módszereimmel mindkét típusú klaszterpárok megjelenhetnek, a konkrét gráftól függ, hogy milyen arányban.

Érdemesnek tartom megemlíteni, hogy fenti céljaimat gyakorlati problémák is motiválták. Az utóbbi évtizedben több olyan projektben (NKFP, OTKA, TÁMOP) vettem részt, mely gráfok vagy kontingenciatáblák formájában megadott nagyméretű hálózatok struktúrájának feltárására irányult. A nagy méretek, folyamatosan változó adatok és az összetett kérdések miatt a klasszikus statisztikai módszerek közvetlenül nem voltak alkalmazhatók; viszont a klasszikus gráfelmélet sem volt alkalmas olyan, a gráfon értelmezett statisztikus mérőszámok becslésére, mint a minimális többszempontú vágások vagy többrészes diszkrepancia, melyek nem érzékenyek az élek vagy élsúlyok kis változásaira. Az általam javasolt algoritmusok beprogramozását azonban az utóbbi időkben diákjaim végezték, akik a dolgozatban található ábrákat is készítették. Jelenleg is dolgozom szakértőként a VTT Technical Research Centre of Finland STOMOGRAPH projektjében, és meghívást kaptam külső tanácsadóként a Ca'Foscari University (Venezia) egy belső projektjébe.

II. Az alkalmazott módszerek

A disszertációban sok helyen használok *lineáris algebra*i tételeket, a sajátértékekre és szinguláris értékekre vonatkozó egyszerű szeparációs tételektől kezdve [Rao] komplikáltabbakat is, pl. Weyl-féle perturbációs elv vagy Davis–Kahan típusú tételek a spektrális alterek eltérésének becslésére [Bhat].

Az első fejezetben egységes jelölésrendszert és módszertant vezetek be azért, hogy hasonló becsléseknél ne kelljen mindig a kezdetekig visszamenni, hanem ugyanarra a technikára, mint reprezetációra tudjak hivatkozni. A *reprezentációs technika* – mely súlyozatlan, súlyozott gráfokra és nem-negatív elemű téglalapmátrixokra is vonatkozik – lehetővé teszi, hogy különböző többszempontú vágásokra alsó becslést adjunk a megfelelő mátrix (szomszédsági, normált Laplace, modularitás, normált kontingenciatábla) spektruma segítségével. A lényeg az, hogy fix pozitív (k) egészre ezen mátrixok k legkisebb (vagy legnagyobb) sajátértékének összege egy k-dimenziós ún. kvadratikus elhelyezési probléma optimumát adja, mely optimum a hozzájuk tartozó sajátvektorok alapján legyártott k-dimenziós reprezentánsokkal valósul meg (a reprezentánsok a gráf csúcsaihoz vagy a kontingenciatábla soraihoz és oszlopaihoz tartoznak). Ezután belátjuk, hogy a kvadratikus célfüggvényt speciális reprezentánsokkal kiértékelve a minimális többszempontú vágást kapjuk, így ez nagyobb, mint az abszolút minimum. Ebből egyszerűen adódik egy alsó becslés. A speciális reprezentánsok a csúcsok k-partícióihoz tartoznak, és az ugyanazon osztálybeli csúcsok reprezentánsai megegyeznek. Azaz egy k-lépcsőn konstans vektorokból álló altér helyettesíti a strukturális sajátértékekhez tartozó sajátalteret (ezt a tényt spektrális relaxációként is szokás emlegetni). Fontos, hogy a két altér eltérése nem más, mint a reprezentánsok ún. *k-varianciája* (az egy osztályba tartozó optimális reprezentánsok belső varianciáinak összege). Ez a többváltozós statisztikából ismert varianciaanalízisbeli tény megkönnyíti a számolásokat.

A reprezentációs technikát kiterjesztem együttes eloszlásokra is, melyhez *Hilbert-terek integráloperátorai*nak elméletét használom és még néhány egyszerű funkcionálanalízisbeli tényt. A kontingenciatáblák és gráfok az együttes eloszlások speciális véges diszkrét esetei, a hozzájuk tartozó mátrixok szinguláris és spektrális felbontása pedig a feltételes várható érték képzés operátorának (mint integráloperátornak) Hilbert spektráltétele által garantált felbontása. Az absztrakció azonban nem önmagáért való. Gyakran ui. adatpontokból indulunk ki, és azokra építünk gráfot. Az adatpontokat (különösen, ha azok lineárisan nem jól szeparálhatók) leképezhetjük egy (sokszor végtelen dimenziós) ún. *reprodukáló magú Hilbert-tér*be. Lényeges, hogy nem kell ezt a leképezést végrehajtani, hanem mivel úgyis csak egy hasonlósági mátrixra van szükségünk, elég az új magfüggvénybe behelyettesíteni azokat. Ez az elmélet (pl. [Ar]), ami a több mint száz éves Riesz–Fréchet-tétel következménye, napjainkban reneszánszát éli, pl. független komponens analízis (ICA) [Bach]. Ebben a kontextusban a Rényi-féle maximálkorreláció, a klasszikus faktoranalízis és az 1970-es években elterjedt korrespondanciaanalízis technikája is egységesen tárgyalható és alkalmazható optimalizálási problémáinkban és a képfelismerésben is.

Az első fejezet másik irányú becsléseihez és a második fejezet perturbációs eredményeihez már szofisztikáltabb módon kell altéreltérési tételeket alkalmaznom. A második fejezetben szintén használom a Wigner-típusú *véletlen mátrixok* elméletét, pl. Füredi–Komlós [Fü-Ko] eredményét a legnagyobb sajátérték nagyságrendjére (1-hez tartó valószínűséggel) és Alon–Krivelevich–Vu [Al-Kr-Vu] *nagy eltérés jellegű tétel*ét a sajátértékek mediánjuktól vagy várható értéküktől való eltérésére. Ennek segítségével és a Borel–Cantelli lemma alkalmazásával majdnem biztos állításokat tudok bizonyítani a sajátértékek nagyságrendjére.

A harmadik fejezetben használom Borgs és társszerzői tesztelhető gráfparaméterekkel kapcsolatos elméletét [Borgsetal1, Borgsetal2] és a Lovász–Sós [Lov-Sos] által bevezetett általánosított kvázirandomság fogalmát. Az ugyancsak itt tárgyalt sztochasztikus blokkmodelleket keverékmodellnek tekintve, azok paramétereinek becslésére és a csúcsok szimultán klaszterezésére a klasszikus EM algoritmust [De-La-Ru] alkalmazom gráfalapú keverékmodellekre.

III. Az elért eredmények

Az eredményeket a disszertáció fejezetei szerint vezetem be. Az első részben az optimalizálási problémákat, a többrészes vágások becslését, és a természetesen adódó gráfalapú mátrixokat tárgyalom a reprezentációs technikákkal együtt (általánosan téglalapmátrixokra és együttes eloszlásokra is). Ehhez egységes,

a [Bol13] könyvbeli jelöléseket használok, és a kandidátusim óta bebizonyított néhány kapcsolódó eredményt is itt sorolok fel. A legtöbb új eredmény a második részben kerül kimondásra: általánosított véletlen gráfok spektrumának és spektrális altereinek jellemzése, spektrum és diszkrepancia közti kapcsolatok. A harmadik rész témája néhány elméleti alkalmazás (tesztelhetőségi kérdések, általánosított kvázirandom tulajdonságok) és paraméterbecslés gráfokra felállított keverékmodellekben. Nem sorolom fel az összes disszertációbeli tételt és nem is feltétlenül ugyanabban a sorrendben, ahogyan ott találhatók, azonban a definíciók és a saját ill. társszerzős tételek számozása és tartalma ugyanaz, mint a disszertációban. Ezeknél a tételeknél zárójelben megjegyzem, hogy hol lettek publikálva és bizonyítva eredetileg. Mivel a disszertációban egységes jelölést használok, ami a hivatkozott cikkeknél általában nincsen így, ezért a kimondott tételek jelölése (esetleg szóhasználata) néha módosul az eredeti cikkekéhez képest.

1. Többszempontú vágások, reprezentáció és spektrum

Először bevezetek néhány jelölést és gráf alapú mátrixot. Legyen $G = (V, \mathbf{W})$ *élsúlyozott gráf*, ahol $V = \{1, \ldots, n\}$ a csúcsok halmaza, az élsúlyokat pedig az $n \times n$-es szimmetrikus $\mathbf{W}$ mátrix tartalmazza, melynek elemeire $w_{ij} = w_{ji} \geq 0$ $(i \neq j)$ és $w_{ii} = 0$ $(i = 1, \ldots, n)$ teljesül. A gyakorlatban w_{ij} az i és j csúcsok közti hasonlóság mérőszáma, és egyszerű gráfok esetén $\mathbf{W}$ a szomszédsági mátrix. $\mathbf{W}$ sorösszegeit, azaz a $d_i = \sum_{j=1}^{n} w_{ij}$ $(i = 1, \ldots, n)$ számokat *általánosított fokszámok*nak nevezzük, melyeket néha a $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ diagonális *fokszám-mátrix*ban vagy a $\mathbf{d} = (d_1, \ldots, d_n)^T$ *fokszám-vektor*ban gyűjtünk össze. A vektorok alapvetően oszlopvektorok.

Ezután adott $1 \leq k \leq n$ egész esetén keressük a csúcsok k-dimenziós $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^k$ reprezentánsait, melyek minimalizálják a

$$Q_k = \sum_{i<j} w_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2 \geq 0 \tag{1}$$

célfüggvényt különböző mellékfeltételek mellett (hasonló elven alapulnak a gráfrajzoló programok is). Speciális reprezentánsokkal Q_k ún. k-részes vágások felírására lesz alkalmas. A feladat megoldását ismertető reprezentációs tételek a mellékfeltételektől függően a következő mátrixok spektrálfelbontását használják.

1. és 4. Definíció: *Az* $\mathbf{L} = \mathbf{D} - \mathbf{W}$ *mátrixot a* $G = (V, \mathbf{W})$ *élsúlyozott gráf Laplace-mátrixának, míg az*

$$\mathbf{L}_D = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{I}_n - \mathbf{W}_D$$

mátrixot a gráf normált Laplace-mátrixának nevezzük.

Megjegyezzük, hogy az $\mathbf{L}_D$ mátrixot [Bol-Tus94]-ban súlyozott Laplace-mátrixnak neveztük, a normált Laplace elnevezés később jelent meg az irodalomban. Mind $\mathbf{L}$ és $\mathbf{L}_D$ pozitív szemidefinit, és a 0 sajátérték multiplicitása megegyezik G összefüggő komponensei (melyeket 0 súlyú élek kötnek össze) számával. G Laplace-spektruma az összefüggő komponensek Laplace-spektrumainak uniója, így a továbbiakban feltesszük, hogy G összefüggő, vagy ami ezzel ekvivalens, $\mathbf{W}$ *irreducibilis*. Mivel $\mathbf{L}_D$ érzéketlen $\mathbf{W}$ skálázására, az általánosság megszorítása nélkül feltehetjük, hogy $\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij} = 1$, ezért a $\sqrt{\mathbf{d}} := (\sqrt{d_1}, \ldots, \sqrt{d_n})^T$ vektor egységnormájú. Ezt a normálást használja a következő definíció.

7. Definíció: *Az* $\mathbf{M} = \mathbf{W} - \mathbf{d}\mathbf{d}^T$ *mátrixot* $G = (V, \mathbf{W})$ *modularitás-mátrixának, az*

$$\mathbf{M}_D = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T = \mathbf{W}_D - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$$

mátrixot pedig G *normált modularitás-mátrixának nevezzük.*

A modularitás-mátrixot fizikusok [New-Gir, New] vezették be, míg a normált modularitás-mátrixot [Bol11c]-ben definiáltam. $\mathbf{M}$ sorainak összege 0, ezért 0 mindig sajátérték $\mathbf{1} := (1, \ldots, 1)^T$ sajátiránnyal. Miután $\mathrm{tr}(\mathbf{M}) < 0$, $\mathbf{M}$-nek mindig vannak negatív sajátértékei, és általában indefinit. Beláttuk a következőt.

8. Tétel ([Boletal15]): *Egy egyszerű, összefüggő gráf modularitás- és normált modularitás-mátrixa pontosan akkor negatív szemidefinit, ha a gráf teljes többrészes.*

Megjegyezzük, hogy $\mathbf{M}$ és $\mathbf{M}_D$ inerciája megegyezik, és a teljes gráf is teljes többrészes (szingleton osztályokkal). Egy másik disszertációbeli tétel (**7. Tétel**) egyik irányban hasonló állítást fogalmaz meg élsúlyozott

gráfokra (l. [Boletal15]). $\mathbf{M}_D$ kapcsolata $\mathbf{L}_D$-vel a következő. Jelölje $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$ az $\mathbf{L}_D$ mátrix sajátértékeit az $\mathbf{u}_0 = \sqrt{\mathbf{d}}, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ ortonormált sajátvektorokkal. Akkor az $\mathbf{M}_D$ mátrix sajátértékei az $1 - \lambda_i$ számok az $\mathbf{u}_i$ sajátvektorokkal ($i = 1, \ldots, n-1$) és még a 0 a $\sqrt{\mathbf{d}}$ sajátvektorral. $\mathbf{M}_D$ spektruma $[-1, 1]$-beli; 1 nem lehet sajátérték, ha G összefüggő, -1 pedig páros G esetén lesz csak sajátérték.

Visszatérve az (1)-beli Q_k célfüggvény minimalizálására, legyen $\mathbf{X}$ a reprezentánsokat soronként tartalmazó $n \times k$-as mátrix (oszlopvektorait jelölje $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$). Ezzel $Q_k = \mathrm{tr}(\mathbf{X}^T \mathbf{L} \mathbf{X})$. Ebből adódik a disszertációbeli **1. Tétel** ([Bol-Tus94], Reprezentációs tétel élsúlyozott gráfokra), melynek értelmében Q_k minimuma a $\sum_{i=1}^n \mathbf{r}_i \mathbf{r}_i^T = \mathbf{X}^T \mathbf{X} = \mathbf{I}_k$ kényszerfeltétel mellett nem más, mint $\mathbf{L}$ legkisebb k sajátértékének összege, és a hozzájuk tartozó sajátvektorok állnak az optimumot elérő $\mathbf{X}^*$ oszlopaiban (a triviális koordinátákat tartalmazó első oszlop el is hagyható).

Ha a csúcsokat is súlyozzuk az $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_n)$ diagonális mátrix pozitív elemeivel, akkor Q_k minimumát a $\sum_{i=1}^n s_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k$ kényszerfeltétel mellett keressük, és az $\mathbf{L}_S = \mathbf{S}^{-1/2} \mathbf{L} \mathbf{S}^{-1/2}$ mátrix spektrálfelbontásával kapjuk. Fontos lesz számunkra az $\mathbf{S} = \mathbf{D}$ speciális eset, melyben a triviális koordinátáktól eleve eltekintünk.

3. Tétel ([Bol-Tus94]) Reprezentációs tétel él- és speciális csúcs-súlyozott gráfokra: *Legyen $G = (V, \mathbf{W})$ összefüggő élsúlyozott gráf $\mathbf{L}_D$ normált Laplace-mátrixszal, melynek sajátértékei $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1}$ az $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ ortonormált sajátvektorokkal. Legyen a $k < n$ pozitív egész olyan, hogy $\lambda_{k-1} < \lambda_k$. Akkor Q_k minimuma a $\sum_{i=1}^n d_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_{k-1}$ és $\sum_{i=1}^n d_i \mathbf{r}_i = \mathbf{0}$ kényszerfeltételek mellett $\sum_{i=1}^{k-1} \lambda_i$. A minimum azokkal az optimális $(k-1)$-dimenziós $\mathbf{r}_1^*, \ldots, \mathbf{r}_n^*$ reprezentánsokkal éretik el, melyek az $\mathbf{X}^* = \mathbf{D}^{-1/2}(\mathbf{u}_1, \ldots, \mathbf{u}_{k-1})$ mátrix sorvektorai.*

Megjegyezzük, hogy mindez megfogalmazható $\mathbf{M}_D$ spektrálfelbontásával is.

A reprezentációs tételek segítségével néhány többszempontú vágásra könnyen alsó becslés adható a spektrummal. Itt csak az ún. normált k-vágást tárgyalom.

5. Definíció: *A $G = (V, \mathbf{W})$ élsúlyozott gráf $U, T \subset V$ csúcshalmazai közti súlyozott vágás $w(U, T) = \sum_{i \in U} \sum_{j \in T} w_{ij}$. Az $U \subset V$ csúcshalmaz térfogata $\mathrm{Vol}(U) = \sum_{i \in U} d_i$ (az előzetes feltételek miatt $\mathrm{Vol}(V) = 1$). Legyen $P_k = (V_1, \ldots, V_k)$ a csúcsok valódi k-partíciója, és jelölje $\mathcal{P}_k$ az összes valódi k-partíció halmazát. G normált k-vágása a $P_k = (V_1, \ldots, V_k)$ partíció tekintetében*

$$f(P_k, G) = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \left(\frac{1}{\mathrm{Vol}(V_a)} + \frac{1}{\mathrm{Vol}(V_b)} \right) w(V_a, V_b) = \sum_{a=1}^{k} \frac{w(V_a, \overline{V}_a)}{\mathrm{Vol}(V_a)} = k - \sum_{a=1}^{k} \frac{w(V_a, V_a)}{\mathrm{Vol}(V_a)},$$

minimális normált k-vágása pedig $f_k(G) = \min_{P_k \in \mathcal{P}_k} f(P_k, G)$.

4. Tétel ([Bol-Mol02]): *A fenti jelölésekkel $f_k(G) \geq \sum_{i=1}^{k-1} \lambda_i$. Tegyük fel, hogy a csúcsok optimális $(k-1)$-dimenziós reprezentánsai a súlyozott k-közép algoritmussal (mely a (3) célfüggvényt minimalizálja) a $V_1, \ldots, V_k$ klaszterekbe sorolhatók úgy, hogy a maximális klaszterátmérőre $\varepsilon \leq \min\{1/\sqrt{2k}, \sqrt{2} \min_i \sqrt{\mathrm{Vol}(V_i)}\}$ teljesül. Akkor $f_k(G) \leq c^2 \sum_{i=1}^{k-1} \lambda_i$, ahol $c = 1 + \varepsilon c' / (\sqrt{2} - \varepsilon c')$ és $c' = 1 / \min_i \sqrt{\mathrm{Vol}(V_i)}$.*

$f_k(G)$ alsó becslése azon alapul, hogy $f(P_k, G)$ speciális kiértékelése Q_k-nak olyan $\mathbf{X}$-el, melynek oszlopvektorai *partícióvektorok* (P_k elemein szakaszonként konstansok a megfelelő kényszerfeltételek mellett), a felső becslés szofisztikáltabb. Könnyen látható, hogy $f_2(G)$ az alábbi $h(G)$ Cheeger-állandó szimmetrikus változata és $f_2(G) \leq 2h(G)$.

6. Definíció: *A fenti jelölésekkel a $G = (V, \mathbf{W})$ élsúlyozott gráf Cheeger-állandója*

$$h(G) = \min_{\substack{U \subset V \\ \mathrm{Vol}(U) \leq \frac{1}{2}}} \frac{w(U, \overline{U})}{\mathrm{Vol}(U)}.$$

A Cheeger-állandóra vonatkozó felső becslés élesítése élsúlyozott gráfra a következő.

6. Tétel ([Bol-Mol04]): *Legyen G összefüggő élsúlyozott gráf. G normált Laplace-mátrixának legkisebb pozitív sajátértékéről tegyük fel, hogy $\lambda_1 \leq 1$. Akkor $\frac{\lambda_1}{2} \leq h(G) \leq \sqrt{\lambda_1(2 - \lambda_1)}$.*

A $\mathcal{P}_k$ feletti minimalizálás NP-nehéz. A spektrális technikák a csúcsok számában polinomiális idejűek. Azonban a spektrális relaxáció pontossága attól függ, milyen közel hozható a Laplace-mátrix k legkisebb sajátértéke által kifeszített altér az ún. partíció-vektorokéhoz. Ezt a közelséget éppen a k-közép algoritmus célfüggvénye fejezi ki. Ennek mérésére bevezetek néhány további jelölést.

Legyen $1 \leq k \leq n$ egész. Az $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^\ell$ pontrendszer *k-varianciája*

$$S_k^2(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \min_{P_k \in \mathcal{P}_k} S_k^2(P_k; \mathbf{r}_1, \ldots, \mathbf{r}_n) = \min_{P_k=(V_1,\ldots,V_k)} \sum_{a=1}^{k} \sum_{j \in V_a} \|\mathbf{r}_j - \mathbf{c}_a\|^2, \tag{2}$$

ahol $\mathbf{c}_a = \frac{1}{|V_a|} \sum_{j \in V_a} \mathbf{r}_j$ az a-adik klaszter súlypontja ($a = 1, \ldots, k$). Most legyenek az $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^\ell$ pontok a $d_1, \ldots, d_n$ pozitív súlyokkal ellátva, ahol $\sum_{i=1}^n d_i = 1$ és $\mathrm{Vol}(U) = \sum_{i \in U} d_i$, $U \subset \{1, \ldots, n\}$. A súlyozott pontrendszer *súlyozott k-varianciáját*

$$\tilde{S}_k^2(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \min_{P_k \in \mathcal{P}_k} \tilde{S}_k^2(P_k; \mathbf{r}_1, \ldots, \mathbf{r}_n) = \min_{P_k=(V_1,\ldots,V_k)} \sum_{a=1}^{k} \sum_{j \in V_a} d_j \|\mathbf{r}_j - \mathbf{c}_a\|^2 \tag{3}$$

definiálja, ahol $\mathbf{c}_a = \frac{1}{\mathrm{Vol}(V_a)} \sum_{j \in V_a} d_j \mathbf{r}_j$ az a-adik pontklaszter súlypontja ($a = 1, \ldots, k$).

5. Tétel ([Bol-Tus94, Bol-Tus00]): *Legyenek $0 = \lambda_0 < \lambda_1 \leq \lambda_2$ az $\mathbf{L}_D$ mátrix legkisebb sajátértékei és $r_1^*, \ldots, r_n^*$ optimális (1-dimenziós) reprezentánsok, melyek a $\mathbf{D}^{-1/2}\mathbf{u}_1$ vektor koordinátái ($\mathbf{u}_1$ a λ_1 sajátértékhez tartozó egységnormájú sajátvektor). Akkor $\tilde{S}_2^2(r_1^*, \ldots, r_n^*) \leq \lambda_1/\lambda_2$.*

Az ún. Newman–Girvan modularitás [New-Gir, New] és annak kiegyensúlyozott és normált változatai [Bol11c] szintén többszempontú vágások, melyek $\mathcal{P}_k$-n a következőt maximalizálják: összegezzük az azonos klaszterbe tartozó i, j csúcspárokra azok tényleges w_{ij} és a függetlenség hipotézise melletti $d_i d_j$ kapcsolatának különbségét. Azaz olyan klasztereket (modulokat) részesítünk előnyben, melyeken belül a csúcsok közti összeköttetések sokkal erősebbek, mint azt véletlen kapcsolódás esetén remélnénk („community structure"). [Bol11c]-ben megmutattam, hogy e mérőszámok maximalizálásához ismét a spektrális relaxáció technikája használható, csak $\mathbf{M}$ és $\mathbf{M}_D$ spektrális felbontása segítségével. Itt a k legnagyobb (pozitív) sajátértéket hasznájuk. Amennyiben a fenti mennyiségeket minimalizáljuk $\mathcal{P}_k$-n, a k legkisebb (negatív) sajátértéket használjuk, és ún. „anticommunity structure"-t kapunk. Persze $\mathbf{M}_D$ spektrumától függ, hogy milyen k-val mely struktúra illeszkedik legjobban az adott gráfra. A 2. részben $\mathbf{M}_D$ legnagyobb abszolút értékű sajátértékeit fogom használni ún. kis diszkrepanciájú klaszterek keresésére.

Bevezettem *kontingenciatáblák* (nem-negatív elemű téglalapmátrixok) sorainak és oszlopainak optimális alacsony-dimenziós reprezentációját és vizsgáltam annak kapcsolatát a normált kontingenciatábla szinguláris felbontásával és a kétszempontú vágások mérőszámával. Legyen $\mathbf{C}$ $m \times n$-es nem-negatív elemű mátrix *Row* és *Col* sor- és oszlop-halmazzal. Adott $k \leq r := \mathrm{rang}(\mathbf{C})$ pozitív egész esetén keressük a sorok és oszlopok $Row = R_1 \cup \cdots \cup R_k$ és $Col = C_1 \cup \cdots \cup C_k$ valódi k-partícióit úgy, hogy az R_a, C_b klaszterpárok közt a mátrixelemek a lehető leghomogénebb mintázatot mutassák ($a, b = 1, \ldots, k$). Jelölje $d_{row,i} = \sum_{j=1}^n c_{ij}$ ($i = 1, \ldots, m$), ill. $d_{col,j} = \sum_{i=1}^m c_{ij}$ ($j = 1, \ldots, n$) a kontingenciatábla sor-, ill. oszlop-összegeit, melyekről feltesszük, hogy pozitívak. Ennél valamivel többet is felteszünk, nevezetesen, hogy a $\mathbf{C}$ mátrix *nem degenerált* ($\mathbf{C}\mathbf{C}^T$ ill. $\mathbf{C}^T\mathbf{C}$ irreducibilis az $m \leq n$ ill. $m > n$ esetekben). A $\mathbf{D}_{row} = \mathrm{diag}(d_{row,1}, \ldots, d_{row,m})$ és $\mathbf{D}_{col} = \mathrm{diag}(d_{col,1}, \ldots, d_{col,n})$ jelölésekkel a $\mathbf{C}$-hez tarozó *normált kontingenciatáblát* a

$$\mathbf{C}_D = \mathbf{D}_{row}^{-1/2} \mathbf{C} \mathbf{D}_{col}^{-1/2} \tag{4}$$

összefüggés definiálja. Nyilvánvalóan $\mathbf{C}_D$ nem érzékeny $\mathbf{C}$ elemeinek skálázására, ezért a továbbiakban feltesszük, hogy $\sum_{i=1}^m \sum_{j=1}^n c_{ij} = 1$.

Szükségünk lesz a $\mathbf{C}_D = \sum_{k=0}^{r-1} s_k \mathbf{v}_k \mathbf{u}_k^T$ szinguláris felbontásra, $r = \mathrm{rang}(\mathbf{C})$. A korrespondencia-analízis elméletéből következik, hogy s_i-k valójában korrelációs együtthatók abszolút értékei, így rájuk $1 = s_0 \geq s_1 \geq \cdots \geq s_{r-1} > 0$ teljesül. Továbbá, ha $\mathbf{C}$ nem degenerált, akkor az 1 szinguláris érték multiplicitása egy, és a hozzá tartozó egységnormájú szinguláris vektorpár: $\mathbf{v}_0 = (\sqrt{d_{row,1}}, \ldots, \sqrt{d_{row,m}})^T$

és $\mathbf{u}_0 = (\sqrt{d_{col,1}}, \ldots, \sqrt{d_{col,n}})^T$ (triviális korrespondancia-faktorok). Adott $1 \leq k \leq r$ egészhez itt is keresheтjük a sorok $\mathbf{r}_1, \ldots, \mathbf{r}_m \in \mathbb{R}^k$ és az oszlopok $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^k$ reprezentánsait, melyekre a

$$Q_k = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \|\mathbf{r}_i - \mathbf{q}_j\|^2 \tag{5}$$

célfüggvény minimális a megfelelő kényszerfeltételekkel.

9. Tétel ([Bol14b] Reprezentációs tétel kontingenciatáblákra: *A fenti jelölésekkel, amennyiben* $k \leq r = \text{rang}(\mathbf{C})$ *és* $s_{k-1} > s_k$*, az (5) célfüggvény minimuma a* $\sum_{i=1}^{m} d_{row,i} \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k$ *és* $\sum_{j=1}^{n} d_{col,j} \mathbf{q}_j \mathbf{q}_j^T = \mathbf{I}_k$ *kényszerfeltételek mellett* $2k - \sum_{i=0}^{k-1} s_i$*, és az ezt elérő optimális sor- ill. oszlop-reprezentánsok a* $\mathbf{D}_{row}^{-1/2}(\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{k-1})$ *ill.* $\mathbf{D}_{col}^{-1/2}(\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1})$ *mátrixok sorvektorai.*

11. Definíció: *A* $\mathbf{C}$ *kontingenciatábla normált kétszempontú* k*-részes vágása a* $P_{row} = (R_1, \ldots, R_k)$*,* $P_{col} = (C_1, \ldots, C_k)$ *partíciók és a* σ *előjelek tekintetében:*

$$\nu_k(P_{row}, P_{col}, \sigma) = \sum_{a=1}^{k} \sum_{b=1}^{k} \left(\frac{1}{\text{Vol}(R_a)} + \frac{1}{\text{Vol}(C_b)} + \frac{2\sigma_{ab}\delta_{ab}}{\sqrt{\text{Vol}(R_a)\text{Vol}(C_b)}} \right) c(R_a, C_b), \tag{6}$$

ahol $c(R_a, C_b) = \sum_{i \in R_a} \sum_{j \in C_b} c_{ij}$ $(a, b = 1, \ldots, k)$*,* $\text{Vol}(R_a) = \sum_{i \in R_a} d_{row,i}$ *és* $\text{Vol}(C_b) = \sum_{j \in C_b} d_{col,j}$ *a klaszterek térfogatai,* δ_{ab} *a Kronecker-delta; továbbá a* σ_{ab} *előjel az 1 vagy* -1 *értéket veheti fel (csak az* $a = b$ *esetben érdekes) és* $\sigma = (\sigma_{11}, \ldots, \sigma_{kk})$ *a releváns előjelek gyűjteménye. A* $\mathbf{C}$ *kontingenciatábla normált kétszempontú* k*-részes vágása* $\nu_k(\mathbf{C}) = \min_{P_{row}, P_{col}, \sigma} \nu_k(P_{row}, P_{col}, \sigma)$*.*

10. Tétel ([Bol14b]): *A fenti jelölések mellett, amennyiben* $k \leq r$ *és* $s_{k-1} > s_k$*, a* $\mathbf{C}$ *kontingenciatábla normált kétszempontú* k*-részes vágására* $\nu_k(\mathbf{C}) \geq 2k - \sum_{i=0}^{k-1} s_i$ *teljesül.*

Megjegyezzük, hogy abban a speciális esetben, melyben $n = m$ és $\mathbf{C}$ szimmetrikus azonosan 0 diagonálissal, egy élsúlyozott gráf súlymátrixát kapjuk, és a normált kontingenciatábla a triviális faktorpártól eltekintve a normált modularitás-mátrixszal azonos, szinguláris értékei pedig a normált modularitás-mátrix sajátértékeinek abszolút értékei. Azonban ν_k nem viszonyítható közvetlenül a minimális normált k-vágáshoz, kivéve a következő speciális eseteket:

- Amennyiben k olyan, hogy a normált modularitás-mátrix $k-1$ legnagyobb abszolút értékű sajátértéke mind pozitív, akkor ν_k a minimális normált k-vágás kétszerese, és az optimális reprezentációban a megegyező indexű sor- és oszlop-reprezentánsok azonosak. Ezért a klasztereken belüli élek súlyai nem játszanak szerepet, és az optimális reprezentáció a klaszterek közti ritka vágásoknak kedvez („community structure").

- Amennyiben k olyan, hogy a normált modularitás-mátrix $k-1$ legnagyobb abszolút értékű sajátértéke mind negatív, akkor az optimális reprezentációban a megegyező indexű sor- és oszlop-reprezentánsok egymás ellentettjei. Így ν_k minimalizálásában a klasztereken belüli élek súlyai játszanak fokozott szerepet, és az optimális reprezentáció a klasztereken belüli ritka élsűrűségnek, míg a klaszterek közti sűrű vágásoknak kedvez („anticommunity structure").

Ezeket a struktúrákat általánosabban vizsgálom a következő részben, ún. reguláris vágások kontextusában.

A reprezentációs problémát általánosítottam együttes eloszlásokra, melyeknek az élsúlyozott gráfok és kontingenciatáblák speciális esetei. Az optimális reprezentánsokat itt általánosabb Hilbert-terek elemeiként definiáltam és beláttam, hogy egyben megoldják a szekvenciális maximálkorreláció keresési feladatot, melynek első lépése a Rényi-féle maximálkorreláció [Reny59a] meghatározása; véges diszkrét esetben pedig a korrespondanciaanalízis feladatát kapjuk. A felsorolandó technikákkal nem csupán egységesen kezelhetők az előző reprezentációs feladatok, de az absztrakció szintén segítségemre lesz a harmadik részben kimondott tesztelhetőségi tételek bizonyításánál.

Legyen (ξ, η) valós értékű valószínűségi változópár, mely az $\mathcal{X} \times \mathcal{Y}$ szorzattér felett van értelmezve. Együttes eloszlásuk $\mathbb{W}$, a $\mathbb{P}$ és $\mathbb{Q}$ marginálisokkal. Tegyük fel, hogy ξ és η függősége reguláris, azaz $\mathbb{W}$

abszolút folytonos a $\mathbb{P}\times\mathbb{Q}$ szorzatmértékre, és jelölje w a Radon–Nikodym deriváltat (Rényi Alfréd [Reny59b] nomenklatúrájával). Breiman és Friedman [Bre-Fri] ACE (Alternating Conditional Expectation) algoritmust leíró cikkének jelöléseivel legyen $H = L^2(\xi)$, ill. $H' = L^2(\eta)$ a ξ, ill. η valószínűségi változók $\mathbb{P}$, ill. $\mathbb{Q}$ eloszlás szerinti 0 várható értékű, véges varianciájú függvényeinek tere, melyek Hilbert-teret alkotnak a kovarianciával, mint belső szorzattal; és melyek természetes módon be vannak ágyazva abba az L^2-térbe, amit hasonlóan a $\mathbb{W}$ együttes eloszlás definiál. A marginálisok közti *feltételes várható érték* képzés operátora valójában integráloperátor, melynek magfüggvénye w, azaz

$$P_{\mathcal{X}} : H' \to H, \quad \psi = P_{\mathcal{X}}\phi = \mathbb{E}(\phi \,|\, \xi), \quad \psi(x) = \int_{\mathcal{Y}} w(x,y)\phi(y)\,\mathbb{Q}(dy)$$

és hasonlóan értelmezhető $P_{\mathcal{Y}} : H \to H'$ is, ami $P_{\mathcal{X}}$ adjungáltja. Tegyük fel, hogy $\int_{\mathcal{X}}\int_{\mathcal{Y}} w^2(x,y)\mathbb{Q}(dy)\mathbb{P}(dx) < \infty$. Ekkor $P_{\mathcal{X}}$ és $P_{\mathcal{Y}}$ kompakt (teljesen folytonos) lineáris operátorok és diszkrét spektrumuk van; a szinguláris felbontásnak megfelelő felbontásuk:

$$P_{\mathcal{X}} = \sum_{i=1}^{\infty} s_i \langle ., \phi_i\rangle_{H'}\psi_i \quad \text{és} \quad P_{\mathcal{Y}} = \sum_{i=1}^{\infty} s_i \langle ., \psi_i\rangle_{H}\phi_i \tag{7}$$

ahol a „szinguláris értékekre" $1 > s_1 \geq s_2 \geq \cdots \geq 0$ teljesül, és ha megszámlálhatóan végtelen sok van belőlük, akkor 0-hoz torlódnak. Megjegyzem, hogy bár $P_{\mathcal{X}}$ és $P_{\mathcal{Y}}$ ortogonális projekciók, nem a teljes teret képezik le, csak egyik marginálist a másikra, azaz a H és a H' terekre vannak megszorítva. Ha ψ_0 és ϕ_0 jelölné a konstans 1 valószínűségi változókat, akkor $\mathbb{E}(\phi_0|\xi) = \psi_0$ és $\mathbb{E}(\psi_0|\eta) = \phi_0$; ezek mégsem alkotnak függvénypárt 1 szinguláris értékkel, mert nem tartoznak a H ill. H' terekhez, ugyanis nem 0 várható értékűek (analóg módon a normált modularitás-mátrixnál mondottakhoz).

Amennyiben speciálisan $\mathbb{W}$ szimmetrikus, akkor $P_{\mathcal{X}} = P_{\mathcal{Y}}$ önadjungált lineáris operátor. Ekkor $P_{\mathcal{X}} : H' \to H$ Hilbert spektráltétele által garantált spektrálfelbontása $P_{\mathcal{X}} = \sum_{i=1}^{\infty}\lambda_i\langle ., \psi_i'\rangle_{H'}\psi_i$, ahol a sajátértékekre $|\lambda_i| \leq 1$ teljesül és a sajátérték–sajátfüggvény egyenlet a $P_{\mathcal{X}}\psi_i' = \lambda_i\psi_i$ alakot ölti (ψ_i és ψ_i' azonos eloszlásúak, de általában nem függetlenek; együttes eloszlásuk $\mathbb{W}$).

A *maximálkorreláció* keresés feladata, melyet Gebelein és Rényi [Reny59a] kezdtek el vizsgálni még a XX. század közepén, a következő. Keressük a $\psi \in H$, $\phi \in H'$ párt, melyek korrelációja a $\mathbb{W}$ együttes eloszlás szerint maximális. A megoldást a $P_{\mathcal{X}}$ operátor szinguláris felbontása adja:

$$\max_{\psi\in H\ \phi\in H'} \mathrm{Corr}_{\mathbb{W}}(\psi,\phi) = \max_{\|\psi\|=\|\phi\|=1} \mathrm{Cov}_{\mathbb{W}}(\psi,\phi) = s_1$$

és a maximum a ψ_1, ϕ_1 páron éretik el. A *korrespondenciaanalízis* feladata a fentinek egyrészt speciális esete, amennyiben véges, diszkrét eloszlásokról van szó; másrészt általánosabb a feladat, amennyiben egymás után keresünk maximálkorrelációkat bizonyos ortogonalitási feltételek mellett. A szorzattér most egy $m \times n$-es kontingenciatábla az $\mathcal{X} = \{1,\ldots,m\}$ sor- és $\mathcal{Y} = \{1,\ldots,n\}$ oszlop-halmazzal és $w_{ij} \geq 0$ elemekkel. $P_{\mathcal{X}}$ és $P_{\mathcal{Y}}$ (7) felbontása pedig a normált kontingenciatábla SVD-jével nyerhető. A korreláció maximalizálása és a megfelelő kvadratikus célfüggvény minimalizálása közti kapcsolat nyilvánvaló a következő, általánosan kimondott reprezentációs tételből.

12. Definíció: *A fenti jelölésekkel legyen $(\mathbf{X},\mathbf{Y})$ k-dimenziós véletlen vektorpár, ahol $\mathbf{X}$ ill. $\mathbf{Y}$ koordinátái H- ill. H'-beliek. Azt mondjuk, hogy az $(\mathbf{X},\mathbf{Y})$ pár a $\mathbb{W}$ együttes eloszlás k-dimenziós reprezentációját valósítja meg, ha $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$, ill. $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_k$ teljesül (azaz $\mathbf{X}$, ill. $\mathbf{Y}$ komponensei korrelálatlanok, egységnyi varianciával), valamint X_i és Y_i együttes eloszlása $\mathbb{W}$. A reprezentáció költsége*

$$Q_k(\mathbf{X},\mathbf{Y}) = \mathbb{E}_{\mathbb{W}}\|\mathbf{X}-\mathbf{Y}\|^2.$$

Az $(\mathbf{X}^*,\mathbf{Y}^*)$ pár optimális reprezentáns, ha a fenti költséget minimalizálja.

11. Tétel ([Bol13] Reprezentációs tétel együttes eloszlásokra): *Legyen $\mathbb{W}$ együttes eloszlás a $\mathbb{P}$ és $\mathbb{Q}$ marginálisokkal. Tegyük fel, hogy a $P_{\mathcal{X}} : H' \to H$ feltételes várható érték vevés operátorának van legalább k pozitív szinguláris értéke, és jelölje $1 > s_1 \geq s_2 \geq \cdots \geq s_k > 0$ a legnagyobbakat. Akkor a fenti k-dimenziós*

reprezentáció minimális költsége $2\sum_{i=1}^{k}(1-s_i)$ *és a minimum az* $\mathbf{X}^*=(\psi_1,\dots,\psi_k)$ *és* $\mathbf{Y}^*=(\phi_1,\dots,\phi_k)$ *optimális reprezentánsokkal érhető el, ahol* ψ_i,ϕ_i *az* s_i *szinguláris értékhez tartozó függvénypár* $(i=1,\dots,k)$.

A szimmetrikus esetben is hasonló állítható sajátértékek segítségével, erről szól a disszertáció **13. Definíció**ja és **12. Tétel**e. Véges esetben, szimmetrikus együttes eloszlásunk ($\mathbb{W}$) tekinthető egy élsúlyozott gráf súlymátrixának. Ekkor a fenti feltételes várható érték vevés operátorának sajátértékei a normált modularitásmátrix sajátértékei; a spektrál- és szinguláris felbontások pedig a marginálisokkal átnormált mátrixokkal kaphatók. Az is igaz, hogy ekkor $\mathbf{M}_D$ legnagyobb sajátértéke az ún. szimmetrikus maximálkorreláció:

$$\mu_1=\max_{\psi,\psi'\text{ i.d.}}\mathrm{Corr}_{\mathbb{W}}(\psi,\psi')=\max_{\substack{\psi,\psi'\text{ i.d.}\\ \mathrm{Var}_{\mathbb{D}}\psi=1}}\mathrm{Cov}_{\mathbb{W}}(\psi,\psi'),$$

ahol $\mathbb{D}$ a szimmetrikus $\mathbb{W}$ eloszlás marginálisa, ψ eloszlása $\mathbb{D}$, továbbá i.d. azonos eloszlásút jelent. Ezzel a **6. Tétel** átfogalmazható a következőképpen ([Bol-Mol02]):

$$\frac{1-\mu_1}{2}\le\min_{\substack{B\subset\mathbb{R}\text{ Borel-halmaz}\\ \psi,\psi'\text{ i.d.}\\ \mathbb{P}_{\mathbb{D}}(\psi\in B)\le 1/2}}\mathbb{P}_{\mathbb{W}}(\psi'\in\overline{B}|\psi\in B)\le\sqrt{1-\mu_1^2},\quad\text{ha}\quad r_1>0.$$

Megjegyzem továbbá, hogy a fenti integráloperátorok magfüggvénye valójában az együttes eloszlás volt, továbbá, hogy a fenti spektrális és szinguláris felbontásokat az alkalmasan normált mátrixok felbontásával kaptuk (ui. a numerikus algoritmusok euklideszi normában egységnyi sajátvektorokat adnak, amiket a marginálisok szerint egységnyi varianciájúvá kell átnormálni). Ennél szofisztikáltabb magokkal is számolhatunk, különösen, ha adatainkban nem-linearitások vannak, vagy olyan metrikus térbeli pontokat akarunk klaszterezni, melyek lineárisan nem jól szeparálhatók. Ilyenkor a rájuk épített gráf hasonlóság-mátrixát olyan módon transzformálhatjuk, hogy pozitív definit magot kapjunk, és az új maggal dolgozunk. Úgy is képzelhetjük, hogy ezzel bizonyos absztrakt térbeli (reprodukáló magú Hilbert-tér) pontok hasonlóságával dolgozunk, de magát a transzformációt nem kell végrehajtanunk, csak az új magfüggvényt megtalálni. Így ahelyett, hogy nem-lineáris módszereket használnánk eredeti pontjaink klaszterezésére, valójában lineáris módszereket használunk az absztrakt térben (feature-space). Ezzel a többdimenziós normális eloszlásra épülő klasszikus statisztikai módszerek átültethetők absztraktabb adatrendszerekre, mely technikákat a független komponens analízis (ICA) [Bach] és képfelismerési eljárások [Shi-Ma] intenzíven használják.

2. Véletlenség kezelése nagy méretű hálózatokban és klaszterezés kis diszkrepanciával

Vizsgáltam ún. felfújt, általános Wigner-zajjal terhelt mátrixok sajátértékeinek és sajátaltereinek aszimptotikus viselkedését (mind négyzetes és téglalap esetben), ha a mátrix mérete tart a végtelenbe a blokkméretekre tett kiegyensúlyozottsági feltételek mellett. Mivel az általánosított véletlen gráfok szomszédsági mátrixa egy speciális zajos mátrixnak felel meg, a felsorolt tételek egyben az ilyen gráfok spektrális karakterizációját is adják (ezeket a 3. részben foglalom össze, ún. általánosított kvázirandom tulajdonságokkal együtt). Megfordítva, egy nagyméretű gráf élsúly-mátrixában vagy egy kontingenciatáblában általános feltételek mellett konstrukciót adtam a blokk-struktúra feltárására a spektrális klaszterezés módszereivel.

16. Definíció: *Legyenek a* w_{ij} $(1\le i\le j\le n)$ *független, valós értékű valószínűségi változók ugyanazon a valószínűségi mezőn értelmezve, továbbá* $w_{ji}=w_{ij}$, $\mathbb{E}(w_{ij})=0$ $(\forall i,j)$, *és* w_{ij}*-k egyenletesen korlátosak (*n*-től függetlenül* $\exists K>0$ *valós szám, hogy* $|w_{ij}|\le K$, $\forall i,j$*). Ekkor az* $n\times n$*-es valós, szimmetrikus* $\mathbf{W}_n=(w_{ij})_{1\le i,j\le n}$ *mátrixot szimmetrikus Wigner-zajnak nevezzük.*

Megjegyzem, hogy az egyenletes korlátosság helyett feltehetnénk, hogy a mátrixelemek normális eloszlásúak vagy ún. sub-Gauss momentumokkal rendelkeznek, a felsorolt eredmények akkor is érvényben maradnának. Az egyenletesen korlátos perturbáció azonban jobban megfelel az élsúly-mátrix perturbációjára, és ugyancsak e mellett a feltétel mellett bizonyította Füredi és Komlós [Fü-Ko], hogy $\|\mathbf{W}_n\|=\max_{1\le i\le n}|\lambda_i(\mathbf{W}_n)|\le 2\sigma\sqrt{n}+\mathcal{O}(n^{1/3}\log n)$ 1-hez tartó valószínűséggel, ha $n\to\infty$, ahol σ a w_{ij} elemek szórásainak közös felső korlátja.

17. Definíció: *Az $n \times n$-es $\mathbf{B}_n$ mátrix szimmetrikus felfújt mátrix, ha van olyan $k < n$ pozitív egész és $\mathbf{P}$ $k \times k$-as, szimmetrikus ún. valószínűség-mátrix $0 < p_{ij} < 1$ elemekkel, továbá $n_1, \ldots, n_k$ pozitív egészek ($\sum_{i=1}^{k} n_i = n$), hogy a $\mathbf{B}_n$ mátrix sorait és oszlopait ugyanúgy permutálva $\mathbf{B}_n$ egy k^2 blokkból álló blokkmátrix alakját ölti, ahol az $n_i \times n_j$-es (i,j) blokkban mindenütt a p_{ij} elemek szerepelnek $(1 \le i, j \le k)$.*

Most k-t rögzítve $\mathbf{P}$-t egyre nagyobb méretű, $n \times n$-es $\mathbf{B}_n$ blokkmátrixszá fújjuk fel, és vizsgáljuk az $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$ zajos mátrixsorozatot, amint $n_1, \ldots, n_k \to \infty$ ($\sum_{i=1}^{k} n_i = n$) körülbelül „azonos sebességgel". Pontosabban feltesszük, hogy

$$\frac{n_i}{n} \ge c \quad \text{valamely} \quad 0 < c \le \frac{1}{k} \quad \text{valós számmal.} \tag{8}$$

Ha $\mathbf{W}_n$ elemeinek egyenletes korlátjáról még azt is feltesszük, hogy

$$K \le \min\{\min_{i,j \in \{1,\ldots,k\}} p_{ij}\ ,\ 1 - \max_{i,j \in \{1,\ldots,k\}} p_{ij}\}, \tag{9}$$

akkor $\mathbf{A}_n$ elemei [0,1]-beliek lesznek, és $G_n = (V, \mathbf{A}_n)$ növekvő véletlen gráfsorozatot alkot. Alkalmas Wigner-zajjal el tudom érni, hogy G_n ún. *általánosított véletlen gráf* legyen.

21. Definíció: *Legyen n természetes szám és $k \le n$ egész. $G_n(\mathbf{P}, \mathcal{P}_k)$ általánosított véletlen gráf a $\mathbf{P}$ valószínűség-mátrixszal a csúcsok $\mathcal{P}_k = (V_1, \ldots, V_k)$ valódi k-partíciójában, ha V_i és V_j csúcsai egymástól függetlenül, p_{ij} valószínűséggel vannak összekötve $(1 \le i \le j \le k)$.*

13. Tétel ([Bol05]): *Legyen $\mathbf{B}_n$ a $k \times k$-as, k rangú szimmetrikus $\mathbf{P}$ valószínűség-mátrix felfújtja $\beta_1, \ldots, \beta_k$ nem-nulla sajátértékekkel, $\mathbf{W}_n$ pedig szimmetrikus Wigner-zaj. Akkor az $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$ zajos mátrixnak vannak $\lambda_1, \ldots, \lambda_k$ **strukturális** sajátértékei, melyekre*

$$|\lambda_i - \beta_i| \le 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n), \quad i = 1, \ldots, k$$

a maradék $n-k$ sajátértékre pedig $|\lambda_j| \le 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n)$, $j = k+1, \ldots, n$ teljesül majdnem biztosan, ha $n \to \infty$ a (8) feltétel mellett.

Mivel $\beta_i = \Theta(n)$ $(i = 1, \ldots, k)$, n növekedésével egyre nagyobb spektrális rés alakul ki $\mathbf{A}_n$ *strukturális* $(\lambda_1, \ldots, \lambda_k)$ és többi sajátértéke közt. Becsülni tudom a távolságot $\mathbf{B}_n$ és $\mathbf{A}_n$ megfelelő sajátalterei közt is, majd alkalmazom az eredményt a $G_n = (V, \mathbf{A}_n)$ élsúlyozott gráfra a (9) feltétel mellett. Tekintem a csúcsok $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^k$ reprezentációját, melyek a strukturális sajátértékekhez tartozó sajátvektorokkal kaphatók a szokásos módon.

14. Tétel ([Bol05]): *A (9) zajfeltétel mellett a $G_n = (V, \mathbf{A}_n)$ zajos véletlen gráf csúcsainak fenti k-dimenziós reprezentációjára*

$$S_k^2(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \mathcal{O}(\frac{1}{n})$$

teljesül majdnem biztosan, ha $n \to \infty$ a (8) feltétellel.

A fenti típusú zajos gráfok Laplace-mátrixa kevésbé kezelhető, viszont normált Laplace-mátrixuk és normált modularitás-mátrixuk spektruma jól karakterizálható.

15. Tétel ([Bol08a]): *Legyen $G_n = (V, \mathbf{A}_n)$ véletlen élsúlyozott gráf, $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$, ahol a $\mathbf{B}_n$ mátrix a k-rangú $\mathbf{P}$ mátrix felfújtja, $\mathbf{W}_n$ pedig a (9) feltételnek eleget tevő szimmetrikus Wigner-zaj. Akkor (n-től függetlenül) létezik $\delta \in (0,1)$ konstans úgy, hogy tetszőleges $0 < \tau < 1/2$ választással G_n normált Laplace-mátrixának van pontosan k darab sajátértéke, melyek a $[0, 1-\delta+n^{-\tau}]$ és $[1+\delta-n^{-\tau}, 2]$ intervallumok uniójában helyezkednek el, míg az összes többi sajátérték $(1-n^{-\tau}, 1+n^{-\tau})$-beli majdnem biztosan, ha $n \to \infty$ a (8) feltétel mellett. Ekvivalens módon, a zajos gráf normált modularitás-mátrixának van $k-1$ sajátértéke legalább $\delta - n^{-\tau}$ abszolút értékkel, míg a többiek legfeljebb $n^{-\tau}$ abszolút értékűek, $\forall\, 0 < \tau < 1/2$.*

Amennyiben a normált modularitás-mátrix $k-1$ strukturális sajátértékéhez tartozó (fokszám-mátrixszal) transzformált sajátvektorai segítségével reprezentálunk, akkor a **16. Tétel** ([Bol08a]) azt állítja, hogy a reprezentánsok súlyozott k-varianciája majdnem biztosan $\mathcal{O}(n^{-2\tau})$ a 15. Tételbeli feltételekkel ($\forall\, 0 < \tau <$

1/2). A **17. Tétel** ([Bol04, Bol08b]) és a disszertáció 2.1 Táblázata egyéb blokkos struktúrákról nyújt áttekintést, míg [Bol11a] a $k = 2$ speciális esetet vizsgálja részletesen.

Megfordítva, szeretnénk felfedezni blokkstruktúrát egy nagyméretű mátrixban, melynek elemeit esetleg hibával tudjuk megfigyelni.

18. Tétel ([Bol05]): *Legyen $(\mathbf{A}_n)$ $n \times n$-es szimmetrikus mátrixok sorozata, nem-negatív, egyenletesen korlátos elemekkel, $n \to \infty$. Tegyük fel, hogy $\mathbf{A}_n$-nek van legalább k darab, $\sqrt{n}$-nél nagyobb rendű sajátértéke (k rögzített), és a $G_n = (V, \mathbf{A}_n)$ gráf csúcsainak van olyan k-partíciója, melyben a (strukturális sajátértékekhez tartozó sajátvektorokkal legyártott) reprezentánsok k-varianciája $\mathcal{O}(1/n)$. Akkor explicit konstrukció adható olyan k^2 blokkból álló szimmetrikus felfújt $\mathbf{B}_n$ mátrixra, mellyel $\|\mathbf{A}_n - \mathbf{B}_n\| = \mathcal{O}(\sqrt{n})$.*

A konstrukció spektrális klaszterezéssel és a klasztercentrumok alkalmas forgatásával történik (szinguláris felbontásokon keresztül). A **9. Állítás**ban ([Bol05]) megmutattam, hogy az elemekre tett egyenletes korlátossági feltételek mellett egy $n \times n$-es, nem-negatív elemű véletlen mátrixnak nagyon általános feltételek mellett van legalább egy $\sqrt{n}$-nél nagyobb rendű sajátértéke.

A fenti eredmények kiterjesztők téglalapmátrixok perturbációira is azzal a különbséggel, hogy a normált mátrix esetében a sorok és oszlopok számának végtelenbe tartását enyhén szinkronizálni kell.

22. Definíció: *Az $m \times n$-es valós, véletlen $\mathbf{W}_{m\times n}$ mátrixot Wigner-zajnak nevezzük, ha elemei független, egyenletesen korlátos, 0 várható értékű valószínűségi változók.*

Az egyenletes korlátosság azért fontos, mert e mellett a feltétel mellett terjesztette ki Achlioptas és McSherry [Ac-Mc] Füredi és Komlós [Fü-Ko] eredményét téglalapmátrixokra. Ennek értelmében a 22. Definícióban szereplő $\mathbf{W}_{m\times n}$ Wigner-zaj spektrálnormája (legnagyobb szinguláris értéke) $\sqrt{m+n}$ rendű 1-hez tartó valószínűséggel, ha $m, n \to \infty$.

23. Definíció: *Az $m \times n$-es valós $\mathbf{B}$ mátrix felfújt mátrix, ha van olyan $a \times b$-es $\mathbf{P}$ valószínűség-mátrix $0 < p_{ij} < 1$ elemekkel, továbbá $m_1, \dots, m_a$ ($\sum_{i=1}^{a} m_i = m$) ill. $n_1, \dots, n_b$ ($\sum_{i=1}^{b} n_i = n$) pozitív egészek, hogy sorainak és oszlopainak alkalmas permutálásával $\mathbf{B}$ egy $a\times b$-es blokkmátrix alakját ölti, ahol az $m_i \times n_j$-es (i, j) blokkon belül az összes elem p_{ij}-vel egyenlő* ($1 \le i \le a$, $1 \le j \le b$).

Az a, b egészeket és $\mathbf{P}$ elemeit rögzítve, a valószínűség-mátrixot egyre nagyobb $\mathbf{B}_{m\times n}$ mátrixszá fújjuk fel, majd azt $m \times n$-es Wigner-zajjal terheljük. Az $\mathbf{A}_{m\times n} = \mathbf{B}_{m\times n} + \mathbf{W}_{m\times n}$ zajos mátrix és a belőle nyert normált mátrix szinguláris felbontásának aszimptotikus viselkedését vizsgáljuk, ha $m, n \to \infty$ az alábbi feltételek mellett (a második csak a normált mátrixhoz kell):

F1 Van olyan $0 < c \le \frac{1}{a}$ konstans, hogy $\frac{m_i}{m} \ge c$ $(i = 1, \dots, a)$ és olyan $0 < d \le \frac{1}{b}$ konstans, hogy $\frac{n_j}{n} \ge d$ $(j = 1, \dots, b)$.

F2 Vannak olyan $C \ge 1$, $D \ge 1$ és $C_0 > 0$, $D_0 > 0$ konstansok és m_0, n_0 küszöbindexek, hogy $m \le C_0 n^C$ és $n \le D_0 m^D$, ha $m \ge m_0$ és $n \ge n_0$.

Amennyiben $\mathbf{W}_{m\times n}$ elemeinek K egyenletes korlátjáról még azt is feltesszük, hogy

$$K \le \min\Big\{ \min_{\substack{i\in\{1,\dots,a\}\\ j\in\{1,\dots,b\}}} p_{ij} \,,\, 1 - \max_{\substack{i\in\{1,\dots,a\}\\ j\in\{1,\dots,b\}}} p_{ij} \Big\}, \tag{10}$$

akkor az $\mathbf{A}_{m\times n}$ mátrix elemei [0,1]-beliek. Alkalmas Wigner-zajjal itt is el tudom érni, hogy $\mathbf{A}_{m\times n}$ véletlen bináris mátrix legyen: elemei az (i, j) blokkban független Bernoulli eloszlásúak p_{ij} paraméterrel $(i = 1, \dots, a;\ j = 1, \dots, b)$, és a különböző blokkok elemei is függetlenek. (Ilyen modelleket gyakran használnak microarray analízisben.)

A perturbációs vizsgálatokban használtam, hogy a 23. Definícióban szereplő $\mathbf{B}_{m\times n}$ mátrixnak van k pozitív $s_1, \dots, s_k$ szinguláris értéke, melyek $\Theta(\sqrt{mn})$ rendűek, ahol $k = \operatorname{rang}(\mathbf{B}_{m\times n}) = \operatorname{rang}(\mathbf{P})$.

19. Tétel ([Bol-Fr-Kr10]): *A fenti jelölésekkel az $\mathbf{A}_{m\times n} = \mathbf{B}_{m\times n} + \mathbf{W}_{m\times n}$ mátrixnak vannak $z_1, \dots, z_k$ **strukturális** szinguláris értékei, melyekre*

$$|z_i - s_i| = \mathcal{O}(\sqrt{m+n}), \quad i = 1, \dots, k$$

többi szinguláris értékére pedig $z_j = \mathcal{O}(\sqrt{m+n})$, $j = k+1, \ldots, \min\{m,n\}$ *teljesül majdnem biztosan, ha* $m, n \to \infty$ *a felfújt mátrix blokkméreteire tett* ***F1*** *feltétel mellett.*

Legyenek az $\mathbf{A}_{m\times n}$ mátrix sorainak és oszlopainak reprezentánsai az $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)$ és az $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ mátrix sorvektorai, ahol $\mathbf{y}_i, \mathbf{x}_i$ a z_i strukturális szinguláris értékhez tartozó szinguláris vektorpár.

20. Tétel ([Bol-Fr-Kr10]): A fenti jelölések és a Wigner-zaj elemeire tett (10) feltételek mellett

$$S_a^2(\mathbf{Y}) = \mathcal{O}\left(\frac{m+n}{mn}\right) \quad \text{és} \quad S_b^2(\mathbf{X}) = \mathcal{O}\left(\frac{m+n}{mn}\right)$$

majdnem biztosan, ha $m, n \to \infty$ *a felfújt mátrix blokkméreteire tett* ***F1*** *feltétel mellett.*

Az $\mathbf{A}_{m\times n}$-ből nyert normált kontingenciatábla szinguláris értékeiről is beláttuk, hogy van közöttük k strukturális a [0,1] intervallum miniatűr világában, ha **F2** is teljesül.

21. Tétel ([Bol-Fr-Kr10]): *A fenti jelölésekkel, van olyan* $\delta > 0$ *konstans (*m*-től és* n*-től függetlenül), hogy tetszőleges* $0 < \tau < 1/2$ *választással: az* $\mathbf{A}_{m\times n}$ *mátrixból nyert normált mátrix* k *legnagyobb szinguláris értéke a* $[\delta - \max\{n^{-\tau}, m^{-\tau}\}, 1 + \max\{n^{-\tau}, m^{-\tau}\}]$ *intervallumba esik, míg a többi szinguláris értéke legfeljebb* $\max\{n^{-\tau}, m^{-\tau}\}$ *majdnem biztosan, ha* $m, n \to \infty$ *a felfújt mátrix blokkméreteire tett* **F1**, *és az* m, n *viszonyára tett* **F2** *feltétel mellett.*

Megjegyzem, hogy a Wigner-zaj elemeinek egyenletes korlátjára tett (10) feltétel mellett a zajos mátrix nem=negatív elemű (így kontingenciatábla), és a normált kontingenciatábla strukturális sajátértékei $[\delta - \max\{n^{-\tau}, m^{-\tau}\}, 1]$-beliek lesznek (1 szükségképpen szingláris érték). Azt is beláttuk (**22. Tétel**, [Bol-Fr-Kr10]), hogy a korrespondencia-faktorokkal nyert $(k-1)$-dimenziós sor- ill. oszlop-reprezentánsok (az 1 szinguláris értékhez tartozó triviális faktorpártól eltekintünk) súlyozott a- ill. b-varianciája majdnem biztosan 0-hoz tart a fenti feltételek mellett.

Egy általános $m \times n$-es véletlen mátrixnak (az elemekre tett egyenletes korlátossági feltételek mellett) tipikusan szokott lenni $\sqrt{m+n}$-nél nagyobb rendű szinguláris értéke (hacsak nem egy Wigner-zaj, de még annál sem zárja ki ezt a lehetőséget a Füredi és Komlós ill. Achlioptas és McSherry tételek gyenge, 1-hez tartó valószínűséggel teljesülő állítása). Ez esetben az alábbi tétel bizonyításában konstrukciót adtunk a blokkstruktúra feltárására.

23. Tétel ([Bol-Fr-Kr10]): *Legyen* $(\mathbf{A}_{m\times n})$ *nem-negatív elemű mátrixsorozat egyenletesen korlátos elemekkel,* $m, n \to \infty$*. Tegyük fel, hogy* $\mathbf{A}_{m\times n}$*-nek van pontosan* k *darab* $\sqrt{m+n}$*-nél nagyobb rendű szinguláris értéke (*k *rögzített). Ha vannak olyan* $a \geq k$ *és* $b \geq k$ *egészek, hogy az optimális sor- és oszlop-reprezentánsok* a*- és* b*-varianciája* $\mathcal{O}(\frac{m+n}{mn})$ *nagyságrendű, akkor explicit konstrukció adható olyan* $\mathbf{B}_{m\times n}$ *felfújt mátrixra (*$a \times b$ *blokkal), melyre* $\|\mathbf{A}_{m\times n} - \mathbf{B}_{m\times n}\| = \mathcal{O}(\sqrt{m+n})$.

Fontos, hogy a különbség spektrál-normájának nagyságrendje annyi, mint egy Wigner-zajé. Ezzel kvázi zajtalanítottuk a mátrixot. Ez azért is lényeges, mert a konstrukció alapját képző szinguláris felbontások nagy mátrixokra csak véletlenített algoritmusokkal, közelítően határozhatók meg. A véletlenítés általában egy alkalmas zaj-mátrix hozzáadását jelenti (ami ritkítja vagy digitalizálja a felbontandó mátrixot), viszont ez a perturbáció – mivel a tételben leválasztott hibával azonos nagyságrendű – nem befolyásolja a konstrukció eredményét.

Ezután még általánosabban, *kis diszkrepanciájú klaszterek*et és klaszterpárokat keresek gráfokban és kontingenciatáblákban. Az alábbi tételek ötletet adnak adott élsúlyozott gráf vagy kontingenciatábla esetén az optimális klaszterszám és a klaszterek választásához úgy, hogy a klaszterpárok közti élsűrűség a lehető leghomogénebb legyen. Az eredményeket alkalmazom irányított gráfok kimeneti és bemeneti klasztereinek keresésére is, melyek közti információáramlás a lehető leghomogénebb. A téglalapokkal kezdtem. Először is bevezetem a többrészes diszkrepancia fogalmát.

25. Definíció: *A* $\mathbf{C}$ *nem-negatív elemű téglalapmátrix többrészes diszkrepanciája a sorok* $R_1, \ldots, R_k$ *és oszlopok* $C_1, \ldots, C_k$ *valódi* k*-partíciójában*

$$\mathrm{md}(\mathbf{C}; R_1, \ldots, R_k, C_1, \ldots, C_k) = \max_{\substack{1\leq a,b\leq k \\ X\subset R_a,\, Y\subset C_b}} \frac{|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)|}{\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}},$$

ahol $c(X,Y)$, $\mathrm{Vol}(X)$ *és* $\mathrm{Vol}(Y)$ *definíciója a 1. részbeli, míg* $\rho(R_a,C_b)=\frac{c(R_a,C_b)}{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$ *jelöli az* R_a *és* C_b *közti relatív sűrűséget. A* $\mathbf{C}$ *mátrix* k*-részes diszkrepanciája*

$$\mathrm{md}_k(\mathbf{C})=\min_{\substack{(R_1,\dots,R_k)\\(C_1,\dots,C_k)}}\mathrm{md}(\mathbf{C};R_1,\dots,R_k,C_1,\dots,C_k).$$

Vegyük észre, hogy a többrészes diszkrepancia érzéketlen $\mathbf{C}$ elemeinek skálázására, így $\sum_{i=1}^n\sum_{j=1}^m c_{ij}=1$ nyugodtan feltehető. Megjegyezzük, hogy $\mathrm{md}(\mathbf{C};R_1,\dots,R_k,C_1,\dots,C_k)$ az a legkisebb α, melyre minden R_a,C_b pár és minden $X\subset R_a$, $Y\subset C_b$ esetén

$$|c(X,Y)-\rho(R_a,C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)|\le\alpha\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)} \tag{11}$$

teljesül. Ez hasonlít a Szemerédi lemma ϵ-reguláris párjaihoz, de térfogattal számosság helyett, és a partíció elemei nem azonos méretűek. Az ötletet a expander mixing lemma ($k=1$ eset) és Alon et al. [Aletal] térfogat-reguláris klaszterpár fogalma adta. Téglalapmátrixokra ez úgy hangzik, hogy az R_a,C_b klaszterpár α*-térfogat reguláris*, ha azokra tetszőleges $X\subset R_a$, $Y\subset C_b$ esetén (11) teljesül a jobb oldalon $\alpha\sqrt{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$-vel. Az alábbi tétel az expander mixing lemma megfordítását jelenti a k-osztályos esetben; a $k=1$ esetben l. [Bil-Lin, Bo-Nik, But].

24. Tétel ([Bol16]): *Legyen* $\mathbf{C}$ *nem-negatív elemű nem degenerált valós mátrix és* $1\le k\le\mathrm{rang}(\mathbf{C})$ *egész. Akkor a (4) -beli* $\mathbf{C}_D$ *mátrix* k*-adik legnagyobb szinguláris értékére (a triviális 1-től eltekintve)*

$$s_k\le 9\mathrm{md}_k(\mathbf{C})(k+2-9k\ln\mathrm{md}_k(\mathbf{C})), \tag{12}$$

teljesül, ha $0<\mathrm{md}_k(\mathbf{C})<1$.

Megjegyezzük, hogy $\mathrm{md}_k(\mathbf{C})=0$, ha $\mathbf{C}$ blokkmátrix k sor- és oszlop-blokkal, amikoris $s_k=0$; továbbá $\mathrm{md}_k(\mathbf{C})<1$ nem szigorú megszorítás, hiszen $s_k<1$ miatt a felső becslés csak 1-nél jóval kisebb $\mathrm{md}_k(\mathbf{C})$ esetén érdekes. Pl. $\mathrm{md}_1(\mathbf{C})\le 1.866\times10^{-3}$, $\mathrm{md}_2(\mathbf{C})\le 8.459\times10^{-4}$, $\mathrm{md}_3(\mathbf{C})\le 5.329\times10^{-4}$, stb. A direkt irányban a következő mondható el.

25. Tétel ([Bol14b]): *Legyen* $\mathbf{C}$ *nem degenerált* $m\times n$*-es kontingenciatábla* $d_{row,1},\dots,d_{row,m}$ *és* $d_{col,1},\dots,d_{col,n}$ *sor- és oszlopösszegekkel. Tegyük fel, hogy* $\sum_{i=1}^n\sum_{j=1}^m c_{ij}=1$ *és nincsenek dominált sorok és oszlopok:* $d_{row,i}=\Theta(\frac{1}{m})$, $i=1,\dots,m$ *és* $d_{col,j}=\Theta(\frac{1}{n})$, $j=1,\dots,n$, *ha* $m,n\to\infty$. $\mathbf{C}_D$ *szinguláris értékei:* $1=s_0>s_1\ge\dots\ge s_{k-1}>\varepsilon\ge s_i$, $i\ge k$. *Az* $(R_1,\dots,R_k)$ *sor- és* $(C_1,\dots,C_k)$ *oszlop-partíció olyan, hogy minimalizálja az optimális sor- és oszlopreprezentánsok (3)-beli* $\tilde S_k^2(\mathbf{X})$ *és* $\tilde S_k^2(\mathbf{Y})$ *súlyozott* k*-varianciáját. Tegyük fel, hogy léteznek* $0<K_1,K_2\le\frac{1}{k}$ *konstansok, melyekkel* $|R_i|\ge K_1n$ *és* $|C_i|\ge K_2m$ $(i=1,\dots,k)$. *Akkor az* R_i,C_j *párok* $\mathcal{O}(\sqrt{2k}(\tilde S_k(\mathbf{X})\tilde S_k(\mathbf{Y}))+\varepsilon)$*-térfogat regulárisak* $(i,j=1,\dots,k)$.

Ilyen módon a k-részes diszkrepanciára felső becslést kapunk s_k és a súlyozott k-varianciák segítségével, melyek „kicsik", ha nagy a rés s_{k-1} és s_k közt. A 24. és 25. Tétel együttes üzenete az, hogy a spektrális klaszterezés módszereivel a gyakorlati alkalmazók igényeinek megfelelő kis diszkrepanciájú klaszterek és klaszterpárok kaphatók, ahol k választásához a $\mathbf{C}_D$ mátrix spektrumában való tájékozódás segít. Az elmélet átmegy irányított és irányítatlan gráfokra is.

28. Definíció: *Az irányítatlan élsúlyozott* $G=(V,\mathbf{W})$ *gráf diszkrepanciája a csúcsok* $V_1,\dots,V_k$ *valódi* k*-partíciójában*

$$\mathrm{md}(G;V_1,\dots,V_k)=\max_{\substack{1\le a\le b\le k\\X\subset V_a,\,Y\subset V_b}}\frac{|w(X,Y)-\rho(V_a,V_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)|}{\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}}.$$

$G=(V,\mathbf{W})$ *minimális* k*-részes diszkrepanciája pedig*

$$\mathrm{md}_k(G)=\min_{(V_1,\dots,V_k)}\mathrm{md}(G;V_1,\dots,V_k).$$

26. Tétel ([Bol16]): *Legyen* $G=(V,\mathbf{W})$ *irányítatlan élsúlyozott gráf és* $1\le k\le\mathrm{rang}(\mathbf{W})$ *egész. Akkor*

$$|\mu_k|\le 9\mathrm{md}_k(G)(k+2-9k\ln\mathrm{md}_k(G)), \tag{13}$$

ha $0 < \mathrm{md}_k(\mathbf{C}) < 1$*, ahol* μ_k *a* G *gráf normált modularitás-mátrixának* k*-adik legnagyobb abszolút értékű sajátértéke.*
Létezik a 25. Tétellel analóg állítás is (**27. Tétel**), l. [Bol11b, Bol14a].

Legyen most $G = (V, \mathbf{W})$ *irányított, élsúlyozott gráf*, melynek $n \times n$-es $\mathbf{W}$ élsúly mátrixa (zéró diagonálissal) a következő: w_{ij} az $i \to j$ irányított él súlya ($i \neq j$). Ekkor az általánosított *be- és kifokok*:

$$d_{in,j} = \sum_{i=1}^{n} w_{ij} \quad (j = 1, \ldots, n) \quad \text{és} \quad d_{out,i} = \sum_{j=1}^{n} w_{ij} \quad (i = 1, \ldots, n);$$

továbbá $\mathbf{D}_{in} = \mathrm{diag}(d_{in,1}, \ldots, \mathbf{d}_{in,n})$ és $\mathbf{D}_{out} = \mathrm{diag}(d_{out,1}, \ldots, \mathbf{d}_{out,n})$ a *be- és kifok-mátrixok.*

29. Definíció: *A* $G = (V, \mathbf{W})$ *irányított, élsúlyozott gráf diszkrepanciája a csúcsok* $V_{in,1}, \ldots, V_{in,k}$ *bemeneti és* $V_{out,1}, \ldots, V_{out,k}$ *kimeneti klaszterezésében*

$$\begin{aligned}&\mathrm{md}(G; V_{in,1}, \ldots, V_{in,k}, V_{out,1}, \ldots, V_{out,k}) \\ &= \max_{\substack{1 \le a,b \le k \\ X \subset V_{out,a},\, Y \subset V_{in,b}}} \frac{|w(X,Y) - \rho(V_{out,a}, V_{in,b})\mathrm{Vol}_{out}(X)\mathrm{Vol}_{in}(Y)|}{\sqrt{\mathrm{Vol}_{out}(X)\mathrm{Vol}_{in}(Y)}},\end{aligned}$$

ahol $w(X,Y)$ *az* $X \to Y$ *élek összsúlya, míg* $\mathrm{Vol}_{out}(X) = \sum_{i \in X} d_{out,i}$ *és* $\mathrm{Vol}_{in}(Y) = \sum_{j \in Y} d_{in,j}$*. G minimális* k*-részes diszkrepanciája pedig*

$$\mathrm{md}_k(G) = \min_{\substack{(V_{in,1}, \ldots, V_{in,k}) \\ (V_{out,1}, \ldots, V_{out,k})}} \mathrm{md}(G; V_{in,1}, \ldots, V_{in,k}, V_{out,1}, \ldots, V_{out,k}).$$

28. Tétel ([Bol16]): *Legyen* $G = (V, \mathbf{W})$ *irányított, élsúlyozott gráf és* $1 \le k \le \mathrm{rang}(\mathbf{W})$ *egész. Akkor*

$$s_k \le 9\mathrm{md}_k(G)(k + 2 - 9k \ln \mathrm{md}_k(G)),$$

ha $0 < \mathrm{md}_k(\mathbf{C}) < 1$*, ahol* s_k *a* $\mathbf{W}_D = \mathbf{D}_{out}^{-1/2}\mathbf{W}\mathbf{D}_{in}^{-1/2}$ *mátrix* k*-adik legnagyobb szinguláris értéke (a triviális 1-től eltekintve).*

3. Elméleti alkalmazások és további elképzelések

A Lovász László és társszerzői által részletesen tárgyalt [Borgsetal1, Borgsetal2] gráfkonvergencia és gráfparaméter tesztelhetőségi elméletet alkalmaztam csúcs- és élsúlyozott gráfokra, továbbá kiterjesztettem azt kontingenciatáblákra. A vágás-normában vett távolságfogalom egyben alkalmas különböző méretű hálózatok (gráfok vagy kontingenciatáblák) összehasonlítására.

A tesztelhető gráfparaméterek a gráfon, mint statisztikai mintán értelmezett olyan nemparaméteres statisztikák, melyek konzisztensen becsülhetők a nagy méretű gráfból történő alkalmas mintavételezéssel (más megközelítésben l. [Bi-Ch]). A minimális vágások általában nem tesztelhetők, de [Bol-Ko-Kr12]-ben beláttuk, hogy a klasztermeretekre tett különböző kiegyensúlyozottsági feltételek mellett bizonyos minimális többszempontú vágássűrűségek tesztelhetők. Mivel ezek rutin alkalmazások, a disszertációban nem foglalkozom velük. Beláttam azt is, hogy konvergens gráfsorozatokra a normált modularitás-spektrum konvergens, és amennyiben a strukturális sajátértékek száma (k) rögzített, az azokhoz tartozó sajátvektorok altere is konvergens, ezért a csúcsreprezentánsok k-varianciája tesztelhető. Megmutattuk továbbá, hogy a 2. részben vizsgált zajos gráf- és kontingenciatábla-sorozatok a homomorfizmus-sűrűségekkel definiált értelemben konvergálnak. A következőkben $\mathbf{W}_n$ a G_n gráf élsúly-mátrixát jelöli, nem pedig Wigner-zajt.

29. Tétel ([Bol14a]): *Legyen* $G_n = (V_n, \mathbf{W}_n)$ *egy konvergens súlyozott gráfsorozat általános tagja [0,1]-beli élsúlyokkal, a csúcsok súlyai pedig legyenek az általánosított fokszámok. Tegyük fel, hogy nincsenek domináns csúcsok. Jelölje* W *a* (G_n) *sorozat limit-grafonját,*

$$1 \ge |\mu_{n,1}| \ge |\mu_{n,2}| \ge \cdots \ge |\mu_{n,n}| = 0$$

pedig G_n normált modularitás-mátrixának spektrumát. Jelölje továbbá $\mu_i(P_{\mathbb{W}})$ annak a $P_{\mathbb{W}} : L^2(\xi') \to L^2(\xi)$ integráloperátornak az i-edik legnagyobb abszolút értékű sajátértékét, mely a W grafonnak megfelelő $\mathbb{W}$ együttes eloszlás szerint vesz feltételes várható értéket, ahol ξ és ξ' azonos eloszlású valószínűségi változók a $\mathbb{W}$ szimmetrikus együttes eloszlás marginálisaival (l. 1. rész). Akkor

$$\forall i \geq 1 : \qquad \mu_{n,i} \to \mu_i(P_{\mathbb{W}}), \quad ha \quad n \to \infty.$$

A 29. Tétel implikálja, hogy tetszőleges pozitív k egészre, a normált modularitás-mátrix $k-1$ legnagyobb abszolút értékű sajátértékeinek együttese tesztelhető. Hasonlót bizonyítottam a megfelelő sajátalterekről is.

30. Tétel ([Bol14a]): *A fenti jelölésekkel tegyük fel még azt is, hogy valamely $0 < \varepsilon < \delta \leq 1$ konstansokkal G_n normált modularitás-spektruma teljesíti a következőt:*

$$1 \geq |\mu_{n,1}| \geq \cdots \geq |\mu_{n,k-1}| \geq \delta > \varepsilon \geq |\mu_{n,k}| \geq \cdots \geq |\mu_{n,n}| = 0,$$

és jelölje $\mathbf{u}_{n,1}, \ldots, \mathbf{u}_{n,n}$ a fenti strukturális sajátértékekhez tartozó ortonormált sajátvektorokat. Akkor feltéve, hogy G_n-nek nincsenek dominált csúcsai, a transzformált $\mathbf{D}_n^{-1/2}\mathbf{u}_{n,1}$, …,$\mathbf{D}_n^{-1/2}\mathbf{u}_{n,k-1}$ vektorok által kifeszített $(k-1)$-dimenziós altér konvergál a $P_{\mathbb{W}}$ operátor analóg alteréhez. Pontosabban, ha $\mathbf{P}_{n,k-1}$ jelöli a G_n gráf normált modularitás-mátrixának $k-1$ legnagyobb abszolút értékű sajátértékéhez tartozó transzformált sajátvektorai által kifeszített altérre való vetítést, $\mathbf{P}_{k-1}$ pedig a $P_{\mathbb{W}}$ analóg alterére való vetítést, akkor $\|\mathbf{P}_{n,k-1} - \mathbf{P}_{k-1}\| \to 0$, ha $n \to \infty$.

Mivel az alkalmas csúcs-reprezentánsok k-varianciája ezen alterek folytonos függvénye, a k-variancia is tesztelhető. A bizonyításban használtam, hogy mind a modularitás-mátrix, mind a limit-grafon Hilbert-Schmidt típusú integráloporátor magfüggvényének tekinthető (l. 1. rész), továbbá használtam a vágás-norma és a Schatten-4 norma közti összefüggéseket. A fentiek üzenete az, hogy domináns csúcsok hiányában a gráf kisebb, alkalmasan randomizált része is alkalmas a klaszterstruktúra felderítésére. Ennek különösen jelentősége van nagyméretű hálózatoknál, mikor annak egy kisebb részére kell csak polinomiális idejű spektrális klaszterezési algoritmusokat alkalmazni. Természetesen a spektrum önmagában nem informatív, ezért használjuk a sajátvektorokon alapuló reprezentánsokat is. A konvergencia és tesztelhetőség elméletét [Bol10]-ben kiterjesztettem kontingenciatáblákra is.

A $\mathbf{W}_n$ Wigner-zajról beláttuk, hogy az annak megfelelő grafon-sorozat vágás-normában majdnem biztosan 0-hoz tart, ha $n \to \infty$. Ennek felhasználásával a 2. rész zajos gráfsorozatai vágás-normában is konvergálnak.

17. Állítás ([Bol-Ko-Kr12]): *Legyen $\mathbf{A}_n := \mathbf{B}_n + \mathbf{W}_n$ a zajos $(G_{\mathbf{A}_n})$ gráfsorozat általános tagjának élsúly mátrixa, ahol $\mathbf{B}_n$ a $k \times k$-as szimmetrikus $\mathbf{P} = (p_{ij})$ valószínűség-mátrix felfújtja az $n_1, \ldots, n_k \to \infty$ blokkméretekkel, melyekre $\lim_{n\to\infty} \frac{n_i}{n} = r_i$ $(i = 1, \ldots, k)$, $n = \sum_{i=1}^k n_i$ teljesül; továbbá a $\mathbf{W}_n$ Wigner-zaj elemeinek egyenletes korlátjáról (9)-et is feltesszük. Akkor a $(G_{\mathbf{A}_n})$ gráfsorozat majdnem biztosan konvergál vágás-normában. A határérték a W_H szakaszonként konstans grafon, mely a H-val jelölt csúcs- és élsúlyozott gráfhoz van rendelve, ahol H csúcs- és élsúlyai a következők:*

$$\alpha_i(H) = r_i \quad (i = 1, \ldots, k), \qquad \beta_{ij}(H) = p_{ij} \quad (i, j = 1, \ldots, k).$$

A 2. részben láttuk, hogy speciális Wigner-zajjal a fenti $(G_{\mathbf{A}_n})$ általánosított véletlen gráfsorozatot definiál (21. Definíció). Az általánosított kvázirandom gráfokat Lovász és Sós [Lov-Sos] éppen a gráfkonvergenciával definiálták.

31. Definíció: *Legyen adva egy H modell-gráf k csúccsal; a csúcsok az $r_1, \ldots, r_k$ pozitív számokkal vannak súlyozva, míg az élsúlyok a $\mathbf{P} = (p_{uv})$ szimmetrikus valószínűség-mátrix elemei. A (G_n) gráfsorozat H-kvázirandom, ha $n \to \infty$ esetén $G_n \to W_H$ (a homomorfizmus sűrűségek értelmében).*

A [Lov-Sos] cikkben a szerzők bizonyították, hogy a fenti általánosított kvázirandom gráf csúcshalmaza a $C_1, \ldots, C_k$ osztályokra particionálható úgy, hogy $\frac{|C_u|}{n} \to r_u$ $(u = 1, \ldots, k)$, ha $n \to \infty$; továbbá a C_u által indukált részgráf egy kvázirandom gráfsorozat általános tagja p_{uu}-hoz tartó élsűrűséggel $(u = 1, \ldots, k)$, míg

a C_u és C_v közti páros gráf egy kvázirandom páros gráfsorozat általános tagja p_{uv}-hez tartó élsűrűséggel $(u \neq v)$, ha $n \to \infty$.

Az általánosított véletlen gráfokra a 2. részben bizonyított tulajdonságok, továbbá a diszkrepancia és spektrum közti összefüggések és a normált modularitás-mátrix spektrumának és spektrális altereinek konvergenciája lehetővé teszi, hogy olyan tulajdonságokat definiáljunk, melyek a sztochasztikus modelltől függetlenül is teljesülnek növekvő gráfsorozatokra. A $k = 1$ esetben a kvázirandomságnak kiterjedt irodalma van, pl. [Chu-G-W, Chu-G, Lov08, Sim-Sos].

Sejtés (általánosított kvázirandom tulajdonságok) ([Bol15, Bol-El16]): *Legyen (G_n) gráfsorozat, G_n csúcshalmaza V_n, szomszédsági mátrixa $\mathbf{A}_n = (a_{ij}^{(n)})$ és normált modularitás-mátrixa $\mathbf{M}_{D,n}$. Legyen k rögzített pozitív egész és $|V_n| = n \to \infty$ úgy, hogy nincsenek domináns csúcsok. Akkor a következő tulajdonságok ekvivalensek:*

P0. *van egy csúcs- és élsúlyozott k csúcsú H gráf, hogy $n \to \infty$ esetén $G_n \to W_H$ a homomorfizmus sűrűségek értelmében, ahol W_H a H-hoz tartozó lépcsős grafon.*

PI. *$\mathbf{A}_n$-nek van k strukturális $\lambda_{1,n}, \dots, \lambda_{k,n}$ sajátértéke, melyek normálva konvergálnak: $\frac{1}{n}|\lambda_{i,n}| \to q_i$, ha $n \to \infty$ $(i = 1, \dots, k)$ valamely pozitív $q_1, \dots, q_k$ valós számokkal, míg a többi sajátérték $o(n)$ abszolút értékű; továbbá az $\mathbf{A}_n$ strukturális sajátértékeihez tartozó sajátvektorokkal definiált k-dimenziós reprezentánsok (2)-beli $S_{k,n}^2$ k-varianciája $o(1)$.*

PII. *Van olyan (n-től független) $0 < \delta < 1$ konstans, hogy $\mathbf{M}_{D,n}$-nek van $k-1$ strukturális sajátértéke legalább δ abszolút értékkel, míg a többi sajétérték $o(1)$ rendű; továbbá az $\mathbf{M}_{D,n}$ mátrix strukturális sajátértékeihez tartozó transzformált sajátvektorokkal definiált $(k-1)$-dimenziós reprezentánsok (3)-beli $\tilde{S}_{k,n}^2$ súlyozott k-varianciája $o(1)$.*

PIII. *V_n-nek létezik olyan $\mathcal{P}_k = (C_1, \dots, C_k)$ partíciója és van olyan (n-től független) $0 < \theta < 1$ konstans, hogy $\mathrm{md}_1(G_n) > \theta, \dots, \mathrm{md}_{k-1}(G_n) > \theta$ és $\mathrm{md}(G_n; C_1, \dots, C_k) = o(1)$.*

PIV. *V_n-nek létezik olyan $\mathcal{P}_k = (C_1, \dots, C_k)$ partíciója és van olyan (n-től független) $k \times k$-as szimmetrikus $\mathbf{P} = (p_{uv})$ valószínűség-mátrix, hogy tetszőleges $1 \leq u \leq v \leq k$ és $i, j \in C_u$ választással*

$$\sum_{t \in C_v} a_{it}^{(n)} = p_{uv} n_v + o(n) \qquad \text{és} \qquad \sum_{t \in C_v} a_{it}^{(n)} a_{jt}^{(n)} = p_{uv}^2 n_v + o(n)$$

teljesül, ahol $n_v = |C_v|$, $v = 1, \dots, k$.

A P0<–>PIV ekvivalencia következik abból, hogy a Lovász-Sós [Lov-Sos] cikkben konstruált csúcspartícióhoz tartozó részgráfokra és páros részgráfokra Thomason és Chung, Graham, Wilson [Thom87, Thom89, Chu-G-W] eredményei alkalmazhatók. A normált szomszédsági mátrix spektrumának konvergenciája következik a [Borgsetal1] cikk eredményeiből. Megfordítva, a spektrum konvergenciája általában nem implikálja a gráfsorozat konvergenciáját, de az általánosított kvázirandom esetben igen, ha a strukturális sajátértékekhez tartozó sajátalterek konvergenciáját is figyelembe vesszük (a k-variancia formájában), l. a 13-18. Tétel és a 29-30. Tétel a normált modularitás mátrix spektrumának és spektrális altereinek konvergenciájára. (Lehet, hogy a spektrális rés önmagában is lenyomja a k-varianciát és fontos szerepe van annak, hogy kiindulási mátrixunk nem-negatív elemű.) A PII<–>PIII ekvivalenciát a diszkrepancia és spektrum közti összefüggések garantálják, l. 26-27. Tétel. Speciálisan az általánosított véletlen gráfokra a sejtésbeli tulajdonságokat bizonyítani tudom: ezek a disszertćió 2. fejezetének fent említett tételeiből következnek, és a disszertáció **18. Állítás**ában vannak összefoglalva. Ebben az esetben valamivel több is igaz: $\mathbf{A}_n$ ill. $\mathbf{M}_{D,n}$ nem strukturális sajátértékeinek nagyságrendje $\sqrt{n}$ ill. $n^{-\tau}$ $(0 < \tau < 1/2)$. Ennek oka az, hogy egy általánosított véletlen gráf szomszédsági mátrixa előállítható egy blokkos mátrix és egy Wigner-zaj összegeként. A grafonra normált Wigner-zaj azonban nagy sebességgel konvergál az azonosan nulla grafonhoz. Az általánosított kvázirandom esetben a konvergencia lehet sokkal lassabb, így a spektrumbeli rés is kevésbé gyorsan tágul. Mindenesetre, a PII<–>PIII ekvivalencia azt mutatja, hogy spektrális eszközökkel közelítőleg kis diszkrepanciájú klasztereket ill. klaszterpárokat találhatunk.

Valószínű, hogy a P0–>PIV–>PIII–>PII–>PI–>P0 irányban körbe lehet menni. Fentiek alapján a PIV–>PIII implikáció hiányzik csak, de itt látok esélyt arra, hogy [Thom87, Thom89] eredményei a részgráfokra és a páros részgráfokra átmennek a diszkrepancia és közös szomszédok számának viszonylatában.

Az eddigi módszerek nemparaméteresek voltak. Ezen kívül félparaméteres modelleket is vizsgáltam és a Dempster, Laird és Rubin által a [De-La-Ru] cikkben bevezetett EM algoritmust alkalmaztam egy homogén és inhomogén sztochasztikus blokkmodellben a paraméterek becslésére és ezzel együtt a gráf csúcsainak klaszterezésére. Ezt az alkalmazást gráfra, mint statisztikai mintára nem találtam az irodalomban.

Legyen a statisztikai minta egy n csúcson értelmezett egyszerű gráf $n \times n$-es, szimmetrikus szomszédsági mátrixa, jelölje ezt $\mathbf{A} = (a_{ij})$. (Látszólag egyetlen mintánk van, azonban a diagonális feletti elemeket, mint független valószínűségi változókat tekintjük statisztikai mintának.)

Homogén sztochasztikus blokkmodell:

- Adott k egészre $(1 < k < n)$ a csúcsok függetlenül tartoznak a C_u klaszterekbe π_u valószínűséggel, $u = 1, \ldots, k$; $\sum_{u=1}^{k} \pi_u = 1$.
- C_u és C_v csúcsai egymástól függetlenül,

$$\mathbb{P}(i \sim j | i \in C_u, j \in C_v) = p_{uv}, \quad 1 \leq a, b \leq k$$

  valószínűséggel vannak összekötve.

A modell paramétereit a $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ vektorban és a $k \times k$-as, szimmetrikus $\mathbf{P} = (p_{uv})$ mátrixban foglaljuk össze.

Inhomogén sztochasztikus blokkmodell:

- Adott k egészre $(1 < k < n)$ a csúcsok függetlenül tartoznak a C_u klaszterekbe π_u valószínűséggel, $u = 1, \ldots, k$; $\sum_{u=1}^{k} \pi_u = 1$.
- Az $i \in C_u$ és $j \in C_v$ csúcsok egymástól függetlenül, p_{ij} valószínűséggel vannak összekötve, ahol

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_{iv} + \beta_{ju}, \quad 1 \leq u \leq v \leq k.$$

A modell paraméterei most a $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ vektorban és az $n \times k$-as $\mathbf{B} = (\beta_{iv})$ mátrixban vannak összegyűjtve. Ezt a logit típusú modellt V. Csiszár és társszerzői [Csetal2] javasolták, ahol a modell építőelemei a [Ch-Dia-Sl, Csetal1, Rin-Pe-Fi] cikkekben leírt $\alpha - \beta$ modellek ill. a páros részgráfokra adaptált Rasch-modell [Rasch].

Itt $\mathbf{A}$ hiányos adatrendszer, mivel a csúcsok klaszterbe tartozását (tagságát) nem ismerjük. Ezért az $\mathbf{A}$ adatmátrixot a csúcsok $\Delta_1, \ldots, \Delta_n$ ún. tagsági vektoraival egészítjük ki, melyek független, azonos eloszlású k-dimenziós $Poly(1, \underline{\pi})$ véletlen vektorok. A keverékeloszlás alakú likelihood függvényt maximalizáljuk az EM algoritmus alternáló E és M lépéseiben. Kezdeti paraméterértékekből és klaszterezésből kiindulva ez a következő:

E -lépés: kiszámoljuk Δ_i feltételes várható értékét az előző lépésbeli modell-paraméterek és tagságok alapján (feltételes várható érték képzés és Bayes-tétel). Ezzel a csúcsok ún. fuzzy klaszterezését kapjuk, de az is lehet, hogy abba a klaszterbe soroljuk a csúcsot, amelybe a legnagyobb valószínűséggel tartozna. Ezzel egy új klaszterbesorolást kapunk.

M -lépés: az összes u, v klaszterpárrapárra $(1 \leq u \leq v \leq k)$ külön-külön maximalizáljuk a likelihoodot a paraméterekben, mellyel a paraméterek új becslését kapjuk. Ezekkel térünk vissza a E-lépésbe.

Mivel exponenciális eloszláscsaládban vagyunk, az algoritmus konvergenciája (bizonyos feltételek mellett, pl. a részgráfok nem ún. threshold-gráfok) a likelihood-függvény egy lokális maximumához garantált, l. B–Elbanna [Bol-El15].

Összefoglalás

Röviden áttekintjük, hogy milyen eredményekre jutottunk és hogy jelen vizsgálatok mennyire járulnak hozzá a spektrális klaszterezés elméleti és gyakorlati kérdéseinek tisztázásához.

- A Laplace- és modularitás-mátrix fogalmát kiterjesztettük hipergráfokra és élsúlyozott gráfokra. Egységesen tárgyaltuk a gráf és kontingencia tábla alapú probémákat az együttes eloszlások és reprezentációs technika nyelvén. Ez egyrészt lehetőséget ad a különféle többszempontú vágások becslésére, másrészt motiválja a modern (nem feltétlenül Gauss-eloszlás alapú) többváltozós statisztikai módszerek használatát (ICA, reprodukáló magú Hilbert-terek). Rögzített k pozitív egész esetén bevezettük és kezeltük egy gráf csúcsainak (ill. kontingenciatábla sorainak és oszlopainak) k-partícióin értelmezett mennyiségeket, pl. a k-részes diszkrepanciát.

- Jellemeztük az általánosított véletlen gráfok spektrumát, spektrális altereit, és a csúcsklasztereken belüli és köztes diszkrepanciákat, amennyiben a csúcsok száma végtelenbe tart (bizonyos kiegyensúlyozottsági feltételek mellett). Bevezettük a többrészes diszkrepancia fogalmát és oda-vissza állításokat bizonyítottunk közte és a gráf spektrális jellemzői közt. Sejtjük, hogy ezek a tulajdonságok nagyméretű növekvő gráfsorozatokra a sztochasztikus modelltől függetlenül is ekvivalensek. Így bevezettünk ún. általánosított kvázirandom tulajdonságokat. Ezzel a köztes helyzetet térképeztük fel a $k = 1$ osztályos eset (expander mixing lemma és megfordításai) és a Szemerédi regularitási lemma (a csúcsok számától függetlenül maximált nagy számú klaszterrel tetszőlegesen kis diszkrepancia elérhető) esete közt. Azt a szerencsés köztes esetet vizsgáltuk, amelyben valamely k-ra a spektrum belsejében van rés (k strukturális sajátérték és a többi közt), továbbá a strukturális sajátértékekhez tartozó sajátvektorokkal képzett reprezentánsok jól klaszteresednek, mely esetben konstrukciót adtunk klaszterezésre kis k-részes diszkrepanciával.

- Bizonyos feltételek mellett beláttuk a normált modularitás-mátrix strukturális sajátértékeinek, sajáltereinek és a reprezentánsok k-varianciáinak tesztelhetőségét a homomorfizmus-sűrűségek értelmében a Hilbert-teres megközelítés segítségével. Ez azt jelenti, hogy a spektrális klaszterek a nagyméretű gráfból alkalmas mintavételezéssel kiválasztott kisebb részgráf alapján konzisztensen becsülhetők; magát a tesztelhetőséget pedig használni lehet az általánosított kvázirandom ekvivalenciák bizonyításánál.

- Megmutattuk, hogyan használható az EM-algoritmus gráfokra felállított keverékmodellek paraméterbecslésére. Valós életbeli adatokon akkor kaptunk jól értelmezhető eredményeket, mikor a kezdő klaszterezést spektrális módszerekkel választottuk. Ilyen módon a spektrális klaszterezés finomhangolását hajtjuk végre, miközben paramétereket is becslünk. Ez példa arra, hogy a spektrális és keverékfelbontási módszerek sikeresen összeházasíthatók a felhasználók igényeivel (l. Ravi Kannan felvetése a Tézisek I. bekezdésében).

Köszönetnyilvánítás

Köszönettel tartozom aspiráns témavezetőmnek, Tusnády Gábornak a vele megkezdett hipergráf klaszterezési közös munkáért és azért, hogy bevont ezzel kapcsolatos gyakorlati alkalmazásokba. Köszönöm társszerzőimnek (elsősorban Friedl Katalinnak, Krámli Andrásnak és Tusnády Gábornak) a közös munkát, továbbá Lovász Lászlónak, Simonovits Miklósnak és T. Sós Verának a tőlük kapott hasznos tanácsokat. Köszönöm tanszék- és intézetvezetőimnek (Simon Károly és Horváth Miklós), hogy számomra fél év sabbaticalt és később egy év jelentősen csökkentett oktatási terhelést biztosítottak, amikor a legfontosabb eredményeket (24-30. Tételek) sikerült bebizonyítanom. Sok más kollégának és diáknak mondok még köszönetet a disszertációban.

Az értekezéshez kapcsolódó szerzői dolgozatok

[Bol93] Bolla, M., Spectra and Euclidean representation of hypergraphs, *Discret. Math.* **117** (1993), 19–39.

[Bol-Tus94] Bolla, M. and Tusnády, G., Spectra and Optimal Partitions of Weighted Graphs, *Discret. Math.* **128** (1994), 1–20.

[Boletal98] Bolla, M., Michaletzky, Gy., Tusnády, G., Ziermann, M., Extrema of sums of heterogeneous quadratic forms, *Linear Algebra Appl.* **269** (1998), 331-365.

[Bol-Tus00] Bolla, M. and Tusnády, G., Hipergráfok összefüggőségének vizsgálata a spektrumon keresztül (Investigating connectivity of hypergraphs by spectra), *Mat. Lapok* 95/1-2 (2000), 1–27.

[Bol-Mol02] Bolla, M. and Molnár-Sáska, G., Isoperimetric Properties of Weighted Graphs Related to the Laplacian Spectrum and Canonical Correlations, *Studia Sci. Math. Hung.* **39** (2002), 425–441.

[Bol-Mol04] Bolla, M. and M.-Sáska, G., Optimization problems for weighted graphs and related correlation estimates, *Discret. Math.* **282** (2004), 23–33.

[Bol04] Bolla, M., Distribution of the Eigenvalues of Random Block-Matrices *Linear Algebra Appl.* **377** (2004), 219–240.

[Bol05] Bolla, M., Recognizing linear structure in noisy matrices, *Linear Algebra Appl.* **402** (2005), 228-244.

[Bol08a] Bolla, M., Noisy random graphs and their Laplacians, *Discret. Math.* **308** (2008), 4221–4230.

[Bol08b] Bolla, M., On the Spectra of Weighted Random Graphs Related to Social Networks. In *Social Networksd: Development, Evaluation and Influence*. Hannah L. Schneider and Lilli M. Huber eds, Nova Science Publishers (2008), New York, pp. 131–158.

[Bol-Fr-Kr10] Bolla, M., Friedl, K. and Krámli, A., Singular value decomposition of large random matrices (for two-way classification of microarrays), *J. Multivariate Anal.* **101** (2010), 434–446.

[Bol10] Bolla, M., Statistical inference on large contingency tables: convergence, testability, stability. In: Proc. of the COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris, Physica-Verlag, Springer (2010), pp. 817-824.

[Bol11a] Bolla, M., Beyond the expanders, *Int. J. Comb.* (2011), Paper 787596.

[Bol11b] Bolla, M., Spectra and structure of weighted graphs, *Electronic Notes in Discrete Mathematics* **38** (2011), 149-154.

[Bol11c] Bolla, M., Penalized versions of the Newman–Girvan modularity and their relation to normalized cuts and k-means clustering, *Phys. Rev. E* **84** (2011),016108.

[Bol-Ko-Kr12] Bolla, M., Kói, T. and Krámli, A., Testability of minimum balanced multiway cut densities, *Discret. Appl. Math.* **160** (2012), 1019–1027.

[Bol13] Bolla, M., *Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables*, Wiley (2013).

[Bol14a] Bolla, M., Modularity spectra, eigen-subspaces and structure of weighted graphs, *European J. Combin.* **35** (2014), 105–116.

[Bol14b] Bolla, M., SVD, discrepancy, and regular structure of contingency tables, *Discret. Appl. Math.* **176** (2014), 3–11.

[Boletal15] Bolla, M., Bullins, B., Chaturapruek, S., Chen, S., Friedl, K., Spectral properties of modularity matrices, *Linear Algebra and Its Applications* **73** (2015), 359-376.

[Bol-El15] Bolla, M., Elbanna, A., Estimating parameters of a probabilistic heterogeneous block model via the EM algorithm, *Journal of Probability and Statistics* (2015), Article ID 657965.

[Bol15] Bolla, M., Generalized quasirandom properties of graphs, arXiv:1508.04369v3 (2015).

[Bol-El16] Bolla, M., Elbanna, A., Matrix and discrepancy view of generalized random and quasirandom graphs, *Special Matrices* **4**, Issue 1 (2016), 31-45.

[Bol16] Bolla, M., Relating multiway discrepancy and singular values of nonnegative rectangular matrices, *Discret. Appl. Math.* **203** (2016), 26-34.

Egyéb hivatkozások

[Ac-Mc] Achlioptas, D. and McSherry, F., Fast computation of low-rank matrix approximations, *J. ACM* **54** (2007), Article 9.

[Alon] Alon, N., Eigenvalues and expanders. *Combinatorica* **6** (1986), 83–96.

[Al-Kr-Vu] Alon, N., Krivelevich, M. and Vu, V. H., On the concentration of eigenvalues of random symmetric matrices, *Isr. J. Math.* **131** (2002), 259–267.

[Aletal] Alon, N., Coja-Oghlan, A., Han, H., Kang, M., Rödl, V. and Schacht, M., Quasi-randomness and algorithmic regularity for graphs with general degree distributions, *Siam J. Comput.* **39** (2010), 2336–2362.

[Ar] Aronszajn, N., Theory of Reproducing Kernels, *Trans. Am. Math. Soc.* **68** (1950), 337–404.

[Bach] Bach, F. R. and Jordan, M. I., Kernel Independent Component Analysis, *J. Mach. Learn. Res.* **3** (2002), 1–48.

[Bhat] Bhatia, R., Matrix Analysis. Springer (1997).

[Bi-Ch] Bickel, P. J. and Chen, A., A nonparametric view of network models and Newman-Girvan and other modularities, *Proc. Natl. Acad. Sci. USA* **106** (2009), 21068–21073.

[Bil-Lin] Bilu, Y. and Linial, N., Lifts, discrepancy and nearly optimal spectral gap, *Combinatorica* **26** (2006), 495–519.

[Bo] Bollobás, B., *Random graphs*. Academic Press, New York (1987).

[Bo-Nik] Bollobás, B. and Nikiforov, V.,Hermitian matrices and graphs: singular values and discrepancy, *Discret. Math.* **285** (2004), 17–32.

[Borgsetal1] Borgs, C, Chayes, J. T., Lovász, L., T.-Sós, V. and Vesztergombi, K., Convergent graph sequences I: Subgraph Frequencies, metric properties, and testing, *Advances in Math.* **219** (2008), 1801–1851.

[Borgsetal2] Borgs, C, Chayes, J. T., Lovász, L., T.-Sós, V. and Vesztergombi, K., Convergent sequences of dense graphs II: Multiway cuts and statistical physics, *Ann. Math.* **176** (2012), 151–219.

[Bre-Fri] Breiman, L. and Friedman, J. H., Estimating optimal transformations for multiple regression and correlation, *J. Am. Stat. Assoc.* **80** (1985), 580–619.

[But] Butler, S., Using discrepancy to control singular values for nonnegative matrices, *Linear Algebra Appl.* **419** (2006), 486–493.

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

[Chu] Chung, F., *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics **92**. American Mathematical Society, Providence RI (1997).

[Chu-G-W] Chung, F., Graham, R. and Wilson, R. K., Quasi-random graphs, *Combinatorica* **9** (1989), 345–362.

[Chu-G] Chung, F. and Graham, R., Quasi-random graphs with given degree sequences, *Random Struct. Algorithms* **12** (2008), 1–19.

[Csetal1] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejtő, L. and Tusnády, G., When the degree sequence is a sufficient statistic, *Acta Math. Hung.* **134** (2011), 45–53.

[Csetal2] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejto, L. and Tusnady, G., Testing goodness of fit of random graph models, *Algorithms* **5** (2012), 629–635.

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

[Erd-Reny] Erdős, P. and Rényi, A., On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960), 17–61.

[Fid72] Fiedler, M., Bounds for eigenvalues of doubly stochastic matrices, *Linear Algebra Appl.* **5** (1972), 299–310.

[Fid73] Fiedler, M., 1973 Algebraic connectivity of graphs, *Czech. Math. J.* **23** (1973), 298–305.

[Fü-Ko] Füredi, Z. and Komlós, J., The eigenvalues of random symmetric matrices, *Combinatorica* **1** (1981), 233–241.

[Gh-Trev] Gharan, S. H., Trevisan, L., A new regularity lemma and faster approximation algorithms for low threshold rank graphs. In *Proc. APPROX-RANDOM'13* (2013), pp. 303–316.

[Hof72] Hoffman, A. J., Eigenvalues and partitionings of the edges of a graph, *Linear Algebra Appl.* **5** (1972), 137–146.

[Hol-Las-Lei] Holland, P., Laskey, K. B. and Leinhardt, S., Stochastic blockmodels: some first steps. *Social Networks* **5** (1983), 109–137.

[Ho-Lin-Wid] Hoory, S., Linial, N. and Widgerson, A., Expander graphs and their applications, *Bull. Amer. Math. Soc. (N. S.)* **43** (2006), 439–561.

[Juh-Mály] Juhász, F. and Mályusz, K., Problems of cluster analysis from the viewpoint of numerical analysis. In *Numerical Methods, Coll. Math. Soc. J. Bolyai* (ed. Rózsa P), Vol 22, pp. 405–415. North-Holland, Amsterdam (1980).

[Klugetal] Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M., Spectral biclustering of microarray data: coclustering genes and conditions, *Genome Res.* **13** (2003), 703–716.

[Le-Gh-Trev] Lee, J. R., Gharan, S. O. and Trevisan, L., Multi-way spectral partitioning and higher-order Cheeger inequalities. In *Proc. 44th Annual ACM Symposium on the Theory of Computing (STOC 2012)*, pp. 1117–1130. New York NY (2012).

[Lov93] Lovász, L., Random walks on graphs: a survey. In *Combinatorics, Paul Erdős is Eighty. János Bolyai Society, Mathematical Studies* Vol. 2, pp. 1–46. Keszthely, Hungary (1993).

[Lov08] Lovász, L., Very large graphs. In In: *Current Developments in Mathematics* (Jerison D, Mazur B, Mrowka T, Schmid W, Stanley R and Yan ST eds), pp. 67-128. International Press, Somerville, MA (2008).

[Lov-Sos] Lovász, L. and T.-Sós V., Generalized quasirandom graphs, *J. Comb. Theory B* **98** (2008), 146–163.

[Lux] von Luxburg, U., A tutorial on spectral clustering. *Stat. Comput.* **17** (2007), 395–416.

[McSh] McSherry, F., Spectral partitioning of random graphs. In *Proc. 42nd Annual Symposium on Foundations of Computer Science (FOCS 2001)*, pp. 529–537. Las Vegas, Nevada (2001).

[New-Gir] Newman, M. E. J. and Girvan, M., Finding and evaluating community structure in networks, *Phys. Rev. E* **69** (2004), 026113.

[New] Newman, M. E. J., *Networks, An Introduction*. Oxford University Press (2010).

[Ng-Jo-We] Ng, A. Y., Jordan, M. I. and Weiss, Y., On spectral clustering: analysis and an algorithm. In *Proc. 14th Neural Information Processing Systems Conference (NIPS 2001)* (Dietterich TG, Becker S and Ghahramani Z eds), pp. 849–856. MIT Press, Cambridge, USA (2001).

[Rao] Rao, C. R., Separation theorems for singular values of matrices and their applications in multivariate analysis, *J. Multivariate Anal.* **9** (1979), 362–377.

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

[Reny59a] Rényi, A., On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10** (1959), 441–451.

[Reny59b] Rényi, A., New version of the probabilistic generalization of the large sieve. *Acta Math. Acad. Sci. Hung.* **10** (1959), 218–226.

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

[Shi-Ma] Shi, J. and Malik, J., Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (2000), 888–905.

[Sim-Sos] Simonovits, M. and T.-Sós, V., Szemerédi's partition and quasi-randomness, *Random Struct. Algorithms* **2** (1991), 1–10.

[Szem] Szemerédi, E., Regular partitions of graphs. In *Colloque Inter. CNRS. No. 260, Problémes Combinatoires et Théorie Graphes* (Bermond J-C, Fournier J-C, Las Vergnas M and Sotteau D eds), 1976, pp. 399–401.

[Thom87] Thomason, A., Pseudo-random graphs, *Ann. Discret. Math.* **33**, 307–331 (1987).

[Thom89] Thomason, A., Dense expanders and pseudo-random bipartite graphs, *Discret. Math.* **75**, 381–386 (1989).

[Trev] Trevisan, L., Max cut and the smallest eigenvalue. In *Proc. 41th Annual ACM Symposium on the Theory of Computing (STOC 2009)*, pp. 1117–1130. Bethesda, Maryland USA (2009).