Gráfok és kontingenciatáblák klaszterezése spektrális módszerekkel

Bolla Marianna
BME Matematika Intézet
Sztochasztika Tanszék
marib@math.bme.hu

2018. június 21.
MTA

A kitűzött cél

Ravi Kannan (Microsoft Research, India).
Talk at Simons Institute, Berkeley (2013. December 9.):
Clustering – Does Theory Help?

„*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?*"

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Előzmények

- Czeizel Endre adatai, Tusnády, Prékopa, Juhász F., Mályusz;
- A. J. Hoffman (DIMACS), C. R. Rao (Penn State Univ.);
- 1990-2005: Spectral clustering, partition cuts, normalized cuts, Cheeger, dual Cheeger ineq. (F. Chung, L. Trevisan);
- Millenium:
  - Erdős–Rényi modell általánosításai, SBM;
  - Newman–Girvan modularitás, social networks;
  - human genom, microarray, téglalapok (Friedl K., Krámli A.);
  - gráfkonvergencia, tesztelhető gráfparaméterek (Lovász L., Szegedy B. et al.);
  - általános kvázirandom gráfok (T. Sós V., Lovász L.);
  - Szemerédi Regularitási Lemma gyenge és spektrális változatai;
  - reprodukáló magú Hilbert-terek reneszánsza a többváltozós statisztikában, Rényi-féle maximálkorreláció.

Jelölések

$G = (V, \mathbf{A})$ él-súlyozott gráf, $|V| = n$

$\mathbf{A} = (a_{ij})$: $n \times n$-es súlyozott szomszédsági mátrix
$a_{ij} = a_{ji} \geq 0$ $(i \neq j)$ és általában $a_{ii} = 0$ (i=1,. . . ,n)
(a_{ij} =0/1: egyszerű gráf)

$d_i := \sum_{j=1}^{n} a_{ij}$ $(i = 1, \ldots, n)$ általánosított fokok
$\mathbf{d} := (d_1, \ldots, d_n)^T$: fokszám-vektor, $\sqrt{\mathbf{d}} := (\sqrt{d_1}, \ldots, \sqrt{d_n})^T$

$\mathbf{D} := \mathtt{diag}\,(d_1, \ldots, d_n)$: fokszám-mátrix

Néha feltesszük, hogy $\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} = 1$
(normált mátrixokat és diszkrepanciát nem befolyásolja)

Laplace- és modularitás-mátrixok

$\mathbf{L} = \mathbf{D} - \mathbf{A}$: Laplace-mátrix
$\mathbf{L}_D = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$: normált Laplace-mátrix

$\mathbf{M} = \mathbf{A} - \mathbf{d}\mathbf{d}^T$: modularitás-mátrix $(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} = 1)$
(M.E.J. Newman: Networks. An Introduction (2010))
$m_{ij} = a_{ij} - d_i d_j$: diszkrepancia

$\mathbf{M}$ általában indefinit és egyszerű gráfokra (B, BSM diákok, Friedl, LAA (2015)): $\mathbf{M}$ negatív szemidefinit $\Leftrightarrow G = K_{n_1,\ldots,n_k}$

$\mathbf{M}_D = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$: normált modularitás-mátrix (B, Phys. Rev. E (2011))
$\mathrm{Spec}\,(\mathbf{M}_D) \in [-1, 1]$ (kompakt operátor)
1 nem sajátérték, ha G összefüggő ($\mathbf{A}$ irreducibilis)
0 mindig sajátérték $\sqrt{\mathbf{d}}$ sajátvektorral

G spektrális rése $= 1 - \|\mathbf{M}_D\|$

Sajátvektorok és reprezentáció

$\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^n$: $\mathbf{A}$ k legnagyobb abszolút értékű sajátértékéhez tartozó ortonormált sajátvektorokkal

$$(\mathbf{u}_1, \ldots, \mathbf{u}_k) = \begin{pmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \vdots \\ \mathbf{r}_n^T \end{pmatrix},$$

ahol $\mathbf{r}_1, \ldots, \mathbf{r}_n$ k-dimenziós csúcs-reprezentánsok. k-varianciájuk

$$S_k^2 = \min_{(U_1,\ldots,U_k)\in\mathcal{P}_k} \sum_{i=1}^{k} \sum_{v\in U_i} \|\mathbf{r}_v - \mathbf{c}_i\|^2, \quad \mathbf{c}_i = \frac{1}{|U_i|} \sum_{v\in U_i} \mathbf{r}_v,$$

ahol $(U_1, \ldots, U_k) \in \mathcal{P}_k$: V valódi k-partíciója.
Minimalizálás: k-közép algoritmus. Ostrovsky et. al., J. ACM (2012): ha $S_k^2 \leq \epsilon^2 S_{k-1}^2$, akkor létezik PTAS.

Súlyozott k-variancia, altér-távolságok

$\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$: $\mathbf{M}_D$ $k-1$ legnagyobb abszolút értékű sajátértékéhez tartozó ortonormált sajátvektorokkal

$$(\mathbf{D}^{-1/2}\mathbf{u}_1, \ldots, \mathbf{D}^{-1/2}\mathbf{u}_{k-1}) = \begin{pmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \vdots \\ \mathbf{r}_n^T \end{pmatrix},$$

$$\tilde{S}_k^2 = \min_{(U_1,\ldots,U_k)\in\mathcal{P}_k} \sum_{i=1}^{k} \sum_{v\in U_i} d_v \|\mathbf{r}_v - \mathbf{c}_i\|^2$$

ahol $\mathbf{c}_i = \frac{1}{\mathtt{Vol}(U_i)} \sum_{v\in U_i} d_v \mathbf{r}_v$, $\quad \mathtt{Vol}(U_i) = \sum_{v\in U_i} d_v$.
Súlyozott k-variancia minimalizálása: súlyozott k-közép algoritmussal. Fontos, hogy S_k^2 és $\tilde{S}_k^2$: négyzetes eltérés a k ($k-1$) vezető sajátértékhez tartozó sajátaltér és a $\mathcal{P}_k$-n szakaszonként konstans vektorok (partíció-vektorok) altere közt (ANOVA).

Vágások

TÉNYLEGES mínusz VÁRT kapcsolat $X, Y \subset V$ közt:

$$\sum_{i \in X} \sum_{j \in Y} (a_{ij} - d_i d_j) = a(X, Y) - \mathtt{Vol}(X)\mathtt{Vol}(Y),$$

ahol $a(X, Y) = \sum_{i \in X} \sum_{j \in Y} a_{ij}$ az X és Y közti súlyozott vágás, $\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} = 1$.

Ha $a_{ij} = 0/1$, akkor $a(X, Y)$ az X és Y közt átmetsző élek számával $(e(X, Y))$ arányos.

Ha $X \cap Y \neq \emptyset$, akkor az $X \cap Y$-beli éleket duplán számítjuk be.

$\rho(X, Y) := \frac{a(X,Y)}{\mathtt{Vol}(X)\mathtt{Vol}(Y)}$ térfogat-sűrűség X és Y közt.

Többrészes diszkrepancia

Definíció

$G = (V, \mathbf{A})$ diszkrepanciája az $(U_1, \ldots, U_k)$ k-partícióban

$$\mathtt{md}(G; U_1, \ldots, U_k) = \max_{\substack{1 \le i \le j \le k \\ X \subset U_i,\, Y \subset U_j}} \mathtt{md}(X, Y; U_i, U_j),$$

ahol

$$\mathtt{md}(X, Y; U_i, U_j) = \frac{|a(X, Y) - \rho(U_i, U_j)\mathtt{Vol}(X)\mathtt{Vol}(Y)|}{\sqrt{\mathtt{Vol}(X)\mathtt{Vol}(Y)}}$$
$$= |\rho(X, Y) - \rho(U_i, U_j)|\sqrt{\mathtt{Vol}(X)\mathtt{Vol}(Y)}$$

G minimális k-részes diszkrepanciája

$$\mathtt{md}_k(G) = \min_{(U_1, \ldots, U_k) \in \mathcal{P}_k} \mathtt{md}(G; U_1, \ldots, U_k).$$

Megjegyzések

B, DAM (2016)

$\mathtt{md}(G; U_1, \ldots, U_k)$ az a legkisebb α, mellyel minden U_i, U_j pár és minden $X \subset U_i$, $Y \subset U_j$ teljesíti:

$$|a(X,Y) - \rho(U_i,U_j)\mathtt{Vol}(X)\mathtt{Vol}(Y)| \leq \alpha\sqrt{\mathtt{Vol}(X)\mathtt{Vol}(Y)}$$

$\mathtt{md}_k(G)$ az a legkisebb α, ami elérhető a csúcsok valódi k-partícióin

Alon et al. (2010): térfogat-reguláris klaszter-párok

Szemerédi Regularitási Lemma (1976): ε-reguláris klaszter-párok, de egyszerű gráf, majdnem azonos méretű klaszterek (equitable partition), kivételes „kis” klaszter, számosságok vannak a sűrűség nevezőjében (nem térfogatok), „túl kicsi” $X \subset U_i$, $Y \subset U_j$ kizárva:

$$|X| > \varepsilon|U_i|, \quad |Y| > \varepsilon|U_j|$$

Kezdeti eredmények a reprezentációs technikával

Tétel (4. Tétel)

$G = (V, \mathbf{A})$ gráf, $\mathbf{L}_D$ s.értékei: $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$. Tegyük fel, hogy a csúcsok $(k-1)$-dimenziós reprezentánsai a súlyozott k-közép algoritmussal $V_1, \ldots, V_k$ klaszterekbe sorolhatók úgy, hogy a maximális klaszterátmérőre $\varepsilon \leq \min\{1/\sqrt{2k}, \sqrt{2}\min_i \sqrt{\mathtt{Vol}(V_i)}\}$ teljesül. Akkor

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

ahol $c = 1 + \varepsilon c'/(\sqrt{2} - \varepsilon c')$, $c' = 1/\min_i \sqrt{\mathtt{Vol}(V_i)}$ és

$$f_k(G) = \min_{(U_1,\ldots,U_k)\in\mathcal{P}_k} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left(\frac{1}{\mathtt{Vol}(U_i)} + \frac{1}{\mathtt{Vol}(U_j)} \right) a(U_i, U_j).$$

Normált k-vágás, izoperimetria, $k = 2$

Biz: B, Molnár-S. G., Stud. Sci. Math. Hung.(2002)
$f_k(G)$: normált k-vágás; $f_2(G) \leq 2\cdot$Cheeger-állandó

Tétel (6. Tétel; B, Molnár-S. G., DM (2004))

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

$\lambda_1 \leq 1$ biztosan teljesül, ha $\exists\, a_{ij} = 0$ $(i \neq j)$

Tétel (5. Tétel; B, Tusnády, DM (1994))

Legyenek $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$ $\mathbf{L}_D$ *s.értékei és* $\mathbf{u}_1$ *a* λ_1*-hez tartozó normált s.vektor. Akkor*

$$\tilde{S}_2^2(\mathbf{D}^{-1/2}\mathbf{u}_1) \leq \lambda_1/\lambda_2$$

($k > 2$-re nem általánosítható, véletlen ellenpélda)
L. Trevisan et al.: higher-order and dual Cheeger inequalities

Normált kontingenciatábla, korrespondencia-analízis

$\mathbf{C}_{m\times n}$: kontingenciatábla $(c_{ij} \geq 0)$

$d_{row,i} = \sum_{j=1}^{n} c_{ij}$, $\mathbf{D}_{row} = \mathtt{diag}\,(d_{row,1}, \ldots, d_{row,m})$

$d_{col,j} = \sum_{i=1}^{m} c_{ij}$, $\mathbf{D}_{col} = \mathtt{diag}\,(d_{col,1}, \ldots, d_{col,n})$

$\mathbf{C}_D = \mathbf{D}_{row}^{-1/2}\mathbf{C}\mathbf{D}_{col}^{-1/2}$: normált kontingenciatábla

Tegyük fel, hogy $\mathbf{C}\mathbf{C}^T$ irreducibilis; ekkor $\mathbf{C}_D$ szinguláris értékei $1 = s_0 > s_1 \geq \cdots \geq s_{r-1} > 0$, ahol $r = \mathtt{rang}\,(\mathbf{C})$.

Normált kétszempontú k-vágás:

$$\nu_k(P_{row}, P_{col}, \sigma) =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k}\left(\frac{1}{\mathtt{Vol}(R_i)} + \frac{1}{\mathtt{Vol}(C_j)} + \frac{2\sigma_{ij}\delta_{ij}}{\sqrt{\mathtt{Vol}(R_i)\mathtt{Vol}(C_j)}}\right) c(R_i, C_j)$$

$$\nu_k(\mathbf{C}) = \min_{P_{row}, P_{col}, \sigma} \nu_k(P_{row}, P_{col}, \sigma),$$

ahol $\sigma_{ij} = \pm 1$ és P_{row}, P_{col} a sorok ill. oszlopok k-partíciói.

Téglalap becslések

Tétel (10. Tétel; B, DAM (2014))

$$\nu_k(\mathbf{C}) \geq 2k - \sum_{i=0}^{k-1} s_i$$

$s_0 = 1$, s_1 =Rényi-féle max. korreláció, korrespondencia-analízis
Biz: reprezentációs technika (spektrális relaxáció)

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

minimuma

- a megfelelően normált sor- és oszlop-reprezentánsokkal (korrespondencia-faktorok): $2k - \sum_{i=0}^{k-1} s_i$
- partíció-vektorokkal: $\nu_k(\mathbf{C})$

$m = n$ szimmetrikus eset $\Longrightarrow$ klaszterek típusai

Legyen $\mathbf{M}_D$ spektruma: $1 \geq |\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_n| = 0$.
Válasszunk egy k egészet úgy, hogy $|\mu_{k-1}|$ és $|\mu_k|$ közt „rés” van.
Akkor

- ha $\mu_1, \ldots, \mu_{k-1}$ mind pozitív, akkor a hozzájuk tartozó transzformált sajátvektorokkal reprezentálva $f_k(G)$-re kapunk jó közelítést: „community structure”;
- ha $\mu_1, \ldots, \mu_{k-1}$ mind negatív, akkor a hozzájuk tartozó transzformált sajátvektorokkal reprezentálva és klaszteresítve a a gráf maximális normált k-vágására kapunk jó közelítést: „anti-community structure”;
- egyébként „kis” diszkrepanciájú klaszterpárokat kapunk: „regular structure”.

Véletlenség, nagy méretek, diszkrepancia

Definíció (Általánosított véletlen gráf)

Adott egy H modellgráf k csúccsal, $r_1, \ldots, r_k$ csúcs-súlyokkal ($r_i > 0$, $\sum_{i=1}^{k} r_i = 1$) és p_{ij} él-súlyokkal, melyek a $k \times k$-as, szimmetrikus $\mathbf{P}$ valószínűség-mátrix elemei ($0 < p_{ij} < 1$, $\mathtt{rang}(\mathbf{P}) = k$). A G_n egyszerű gráf egy H-n értelmezett általánosított véletlen gráfsorozat n-edik tagja, ha

- *n csúcsa van;*
- *minden v csúcshoz egy klaszter-tagság $c_v \in \{1, \ldots, k\}$ van rendelve az $r_1, \ldots, r_k$ eloszlás szerint;*
- *ismerve tagságukat, a $v \neq u$ csúcs-pár $p_{c_v c_u}$ valószínűséggel van összekötve;*
- *mindezek a döntések függetlenek.*

Szimmetrikus Wigner-zaj

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* $|w_{ij}| \le K$, $\forall i, j$*). Akkor az* $n \times n$*-es valós, szimmetrikus* $\mathbf{W}_n = (w_{ij})_{1 \le i,j \le n}$ *mátrixot szimmetrikus Wigner-zajnak nevezzük.*

Füredi és Komlós Combinatorica (1981):
$\|\mathbf{W}_n\| \le 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3} \log n)$ 1-hez tartó val.séggel $(n \to \infty)$,
ahol $\mathtt{Var}\,(w_{ij}) \le \sigma^2$ (előtte Bernoullira Juhász F.)

Alon–Krivelevich–Vu tétel + Borel–Cantelli lemma $\Longrightarrow$
$\|\mathbf{W}_n\|$ majdnem biztosan (m.b.) $\sqrt{n}$ nagyságrendű $(n \to \infty)$.

Felfújás

P-t egyre nagyobb $n \times n$-es $\mathbf{B}_n$ blokk-mátrixszá fújjuk fel $n_1, \ldots, n_k$ blokkméretekkel ($\sum_{i=1}^{k} n_i = n$) úgy, hogy

$$\frac{n_i}{n} \to r_i, \quad i = 1, \ldots k \quad \text{erős kiegyensúlyozottság.}$$

Sokszor a gyenge kiegyensúlyozottság is elég:

$$\frac{n_i}{n} \geq c \quad \text{valamely} \quad 0 < c \leq \frac{1}{k} \quad \text{valós számmal.}$$

$G_n = (V_n, \mathbf{A}_n)$ (ált. véletlen, egyszerű gráf) előállítható:

$$\mathbf{A}_n = \mathbf{B}_n \text{ (súlyozott determinisztikus) } + \mathbf{W}_n \text{ (zaj)}$$

ha

$$K \leq \min\{ \min_{i,j \in \{1,\ldots,k\}} p_{ij} \, , \, 1 - \max_{i,j \in \{1,\ldots,k\}} p_{ij} \}.$$

Általánosított véletlen gráfok m.b. tulajdonságai

Ha $n \to \infty$ az erős kiegyensúlyozottsági feltétellel, akkor m.b.

- $G_n \to W_H$ a homomorfizmus-sűrűségek értelmében; B, Kói, Krámli, DAM (2012)
- $\mathbf{A}_n$-nek van k $\Theta(n)$ nagyságrendű strukturális sajátértéke, a többi $\mathcal{O}(\sqrt{n})$; $S^2_{k,n} = \mathcal{O}(\frac{1}{n})$; B, LAA (2005)
- $\exists\, 0 < \delta < 1$ (n-től független, csak k-tól függ), hogy $\mathbf{M}_{D,n}$-nek van $k-1$ structurális s.értéke (0-tól δ-val elszeparálva), a többi $o(1)$; $\tilde{S}^2_{k,n} = o(1)$; B, DM (2008); B, Friedl, Krámli, JMVA (2010): téglalapokra;
- $\exists\, 0 < \theta < 1$ (n-től független, csak k-tól függ), hogy $\mathtt{md}_1(G_n) > \theta$, $\ldots$, $\mathtt{md}_{k-1}(G_n) > \theta$ és $\mathtt{md}(G_n; U_1, \ldots, U_k) = o(1)$ a tagságok szerinti klaszterezésben;
- Az U_i-k által generált részgráfok m.b. regulárisak, az U_i és U_j közti páros gráfok m.b. biregulárisak ($i \neq j$) (nagy eltérések).

Diszkrepancia versus spektrum becslésekhez nem is kell véletlenség

Tétel (26. Tétel, B, DAM (2016))

$G = (V, \mathbf{A})$ élsűlyozott gráf, $\mathbf{A}$ irreducibilis, $1 \leq k < \mathtt{rang}(\mathbf{A})$ egész. Feltéve, hogy $0 < \mathtt{md}_k(G) < 1$,

$$|\mu_k| \leq 9\mathtt{md}_k(G)(k + 2 - 9k \ln \mathtt{md}_k(G)),$$

ahol μ_k: $\mathbf{M}_D$ k-adik legnagyobb abszolút értékű s.értéke.

- Mivel $|\mu_k| \leq 1$, csak „nagyon kis" $\mathtt{md}_k(G)$)-re hasznos a becslés.
- Kiterjeszthető irányított gráfokra és téglalapokra is (szinguláris értékekkel), ilyenekra lett eredetileg bebizonyítva.
- Ez az „expander mixing lemma" bizonyos megfordítása irreguláris gráfokra.

Expander mixing lemma és megfordításai

$k = 1$ esetben „expander mixing lemma" irreguláris G gráfra:

Tétel (Chung-Graham, RSA (2008))

$$\mathtt{md}_1(G) \leq \|\mathbf{M}_D\| = |\mu_1|$$

$k = 1$ esetben reguláris gráfokra megfordítás:

Tétel (Bilu and Linial, Combinatorica (2006))

$G = (V, \mathbf{A})$ egyszerű, d-reguláris gráf, $|V| = n$. Tegyük fel, hogy $|e(S,T) - \frac{d|S||T|}{n}| \leq \alpha\sqrt{|S||T|}$, $\forall\, S, T \subset V$, $S \cap T = \emptyset$. Akkor $|\lambda_i(\mathbf{A})| \leq \mathcal{O}(\alpha(1 + \log \frac{d}{\alpha}))$, $i > 1$.

Mivel reguláris gráfokra $\mathbf{A}$ és $\mathbf{M}_D$ s.értékei, továbbá $\mathtt{Vol}(S)$ és $|S|$ arányíthatók, a fenti Tétel az α 1-részes diszkrapanciával:
$|\lambda_2(\mathbf{A})| \leq C\alpha(1 - A\log\alpha)$, $A, C \in \mathbb{R}$ (l. 26. Tétel, $k = 1$).

k-osztályos „expander mixing lemma”

Tétel (25. Tétel, B, EJC (2014))

$G = (V, \mathbf{A})$ egyszerű gráf ($\mathbf{A}$ irreducibilis, $|V| = n$). Tegyük fel, hogy nincs dominált csúcs ($d_v = \Theta(n)$, $v \in V$, $o(n)$ csúcs kivételével). Legyenek $\mathbf{M}_D$ sajátértékei $|\mu_1| \geq \cdots \geq |\mu_{k-1}| > \varepsilon \geq |\mu_k| \geq \cdots \geq |\mu_n| = 0$ és tegyük fel, hogy a $(\mu_1, \ldots, \mu_{k-1}$-alapú) spektrális klaszterezással nyert $(V_1, \ldots, V_k)$ partíció (a minimalizált k-variancia $s^2 = \tilde{S}_k^2$) teljesíti az erős kiegyensúlyozottsági feltételeket. Akkor

$$\mathtt{md}_k(G) \leq \mathtt{md}(G; V_1, \ldots, V_k) = \mathcal{O}(\sqrt{2ks} + \varepsilon).$$

Azóta többet tudok a konstansokról (B, Elbanna, A., DAM (2018))

Speciális esetekben $s = 0 \Longrightarrow \mathtt{md}_k(G) \leq B|\mu_k|$

$\tilde{S}_k^2 = 0$, ha a reprezentánsokat adó $\mathbf{D}^{-1/2}\mathbf{u}_1, \ldots, \mathbf{D}^{-1/2}\mathbf{u}_{k-1}$ vektorok partíció-vektorok (szimmetriák a gráfban). Pl.

- $k = 2$, G páros gráf: $\mu_1 = -1$;
  $\tilde{S}_2^2$ „kicsi", ha $|\mu_2|$ el van szeparálódva $|\mu_1| = 1$-től (l. 5. Tétel és Alon, Combinatorica (1986) „bipartite expanders").
- $k = 2$, G páros, bireguláris gráf: $\mu_1 = -1$, $\tilde{S}_2^2 = 0 \Longrightarrow$ $\mathtt{md}_2(G) \leq B|\mu_2|$, valamely B abszolút konstanssal.
  Konstans erejéig ez Evra et al., arXiv (2014) „expander mixing lemma for bipartite graphs" tételének állítása, ahol a csúcsklaszterek mérete és térfogata konstans szorzóban térnek el egymástól a biregularitás miatt.

Általánosított kvázirandom gráfok

Definíció (Lovász, Sós, J. Comb. Theory B (2008))

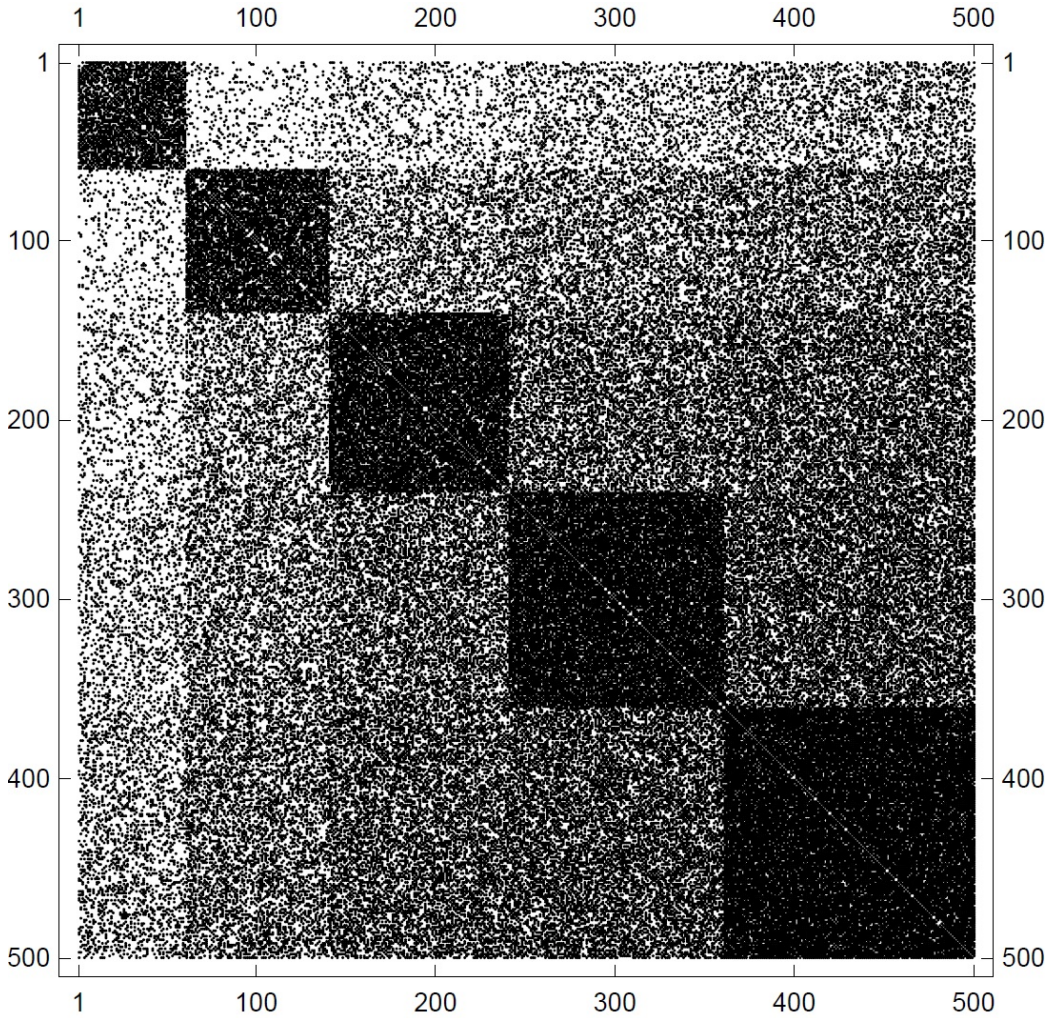
Adott egy H modell-gráf k csúccsal, $r_1, \dots, r_k > 0$ csúcs-súlyokkal és $0 \le p_{ij} = p_{ji} \le 1$ $(1 \le i \le j \le k)$ él-súlyokkal (ezek a $\mathbf{P}$ valószínűség-mátrix elemei). Azt mondjuk, hogy (G_n) H-kvázirandom, ha $G_n \to W_H$ $(n \to \infty)$, ahol W_H lépcsős grafon (konvergencia a homomorfizmus-sűrűségek értelmében).

Szerzők belátták, hogy G_n csúcshalmaza (V) az $U_1, \dots, U_k$ részekre particionálható úgy, hogy

- $\frac{|U_i|}{|V|} \to r_i$, $i = 1, \dots, k$ (erős kiegyensúlyozottság)
- az U_i által indukált részgráf egy p_{ii} élsűrűségű kvázirandom gráfsorozat általános tagja $(i = 1, \dots, k)$.
- az U_i és U_j által indukált páros gráf egy p_{ij} élsűrűségű páros kvázirandom gráfsorozat általános tagja $(i, j = 1, \dots, k;$ $i \neq j)$.

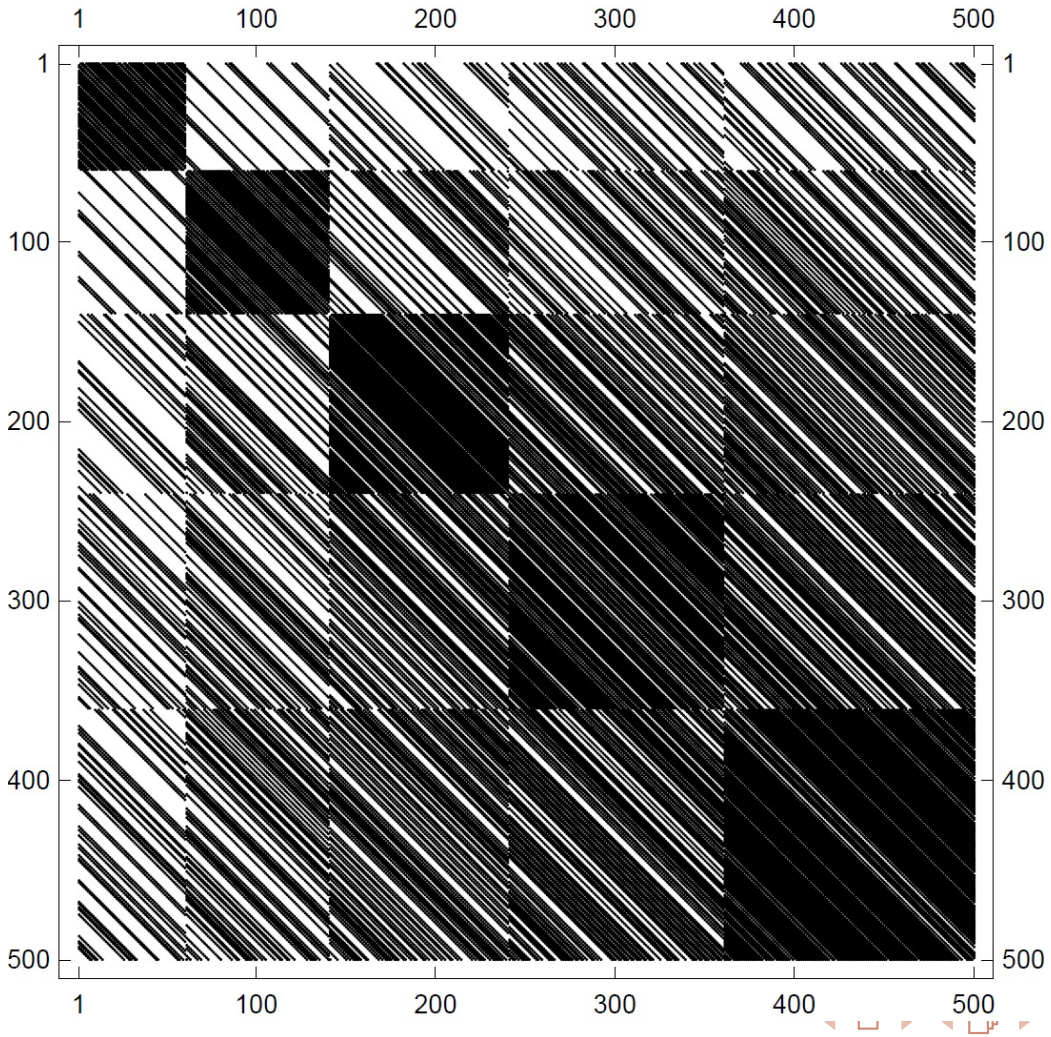
Általánosított véletlen gráf, $n = 500$, $k = 5$

$\mathbf{M}_D$ s.értékei: 0.304, 0.214, 0.17, 0.153, -0.097, -0.094, -0.093, . . .

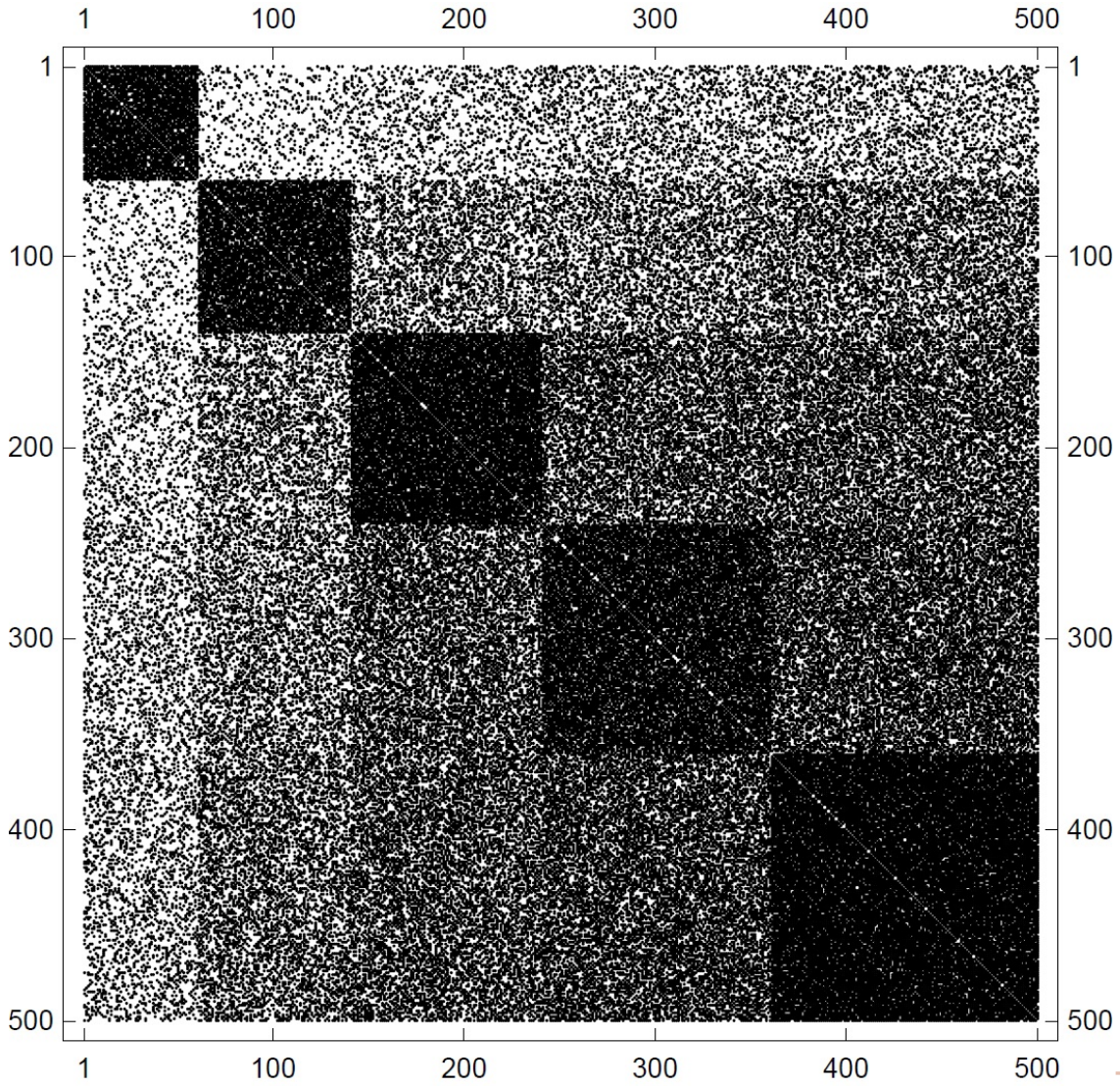


Általánosított kvázirandom gráf, $n = 500$, $k = 5$

$\mathbf{M}_D$ s.értékei: 0.318, 0.207, 0.154, 0.115, -0.100, -0.099, -0.091,. . .



Ugyanaz a csúcsok átcímkézésével



Általanosított kvázirandom tulajdonságok

$G_n = (V_n, \mathbf{A}_n)$, $|V_n| = n \to \infty$ növekvő egyszerű gráfsorozat, k pedig rögzített pozitív egész. Tekintjük a következő tulajdonságokat (nincs semmiféle sztochasztikus modell).

P0. Van oly an H modell-gráf k csúccsal, $r_1, \ldots, r_k > 0$ csúcs-súlyokkal és $0 \le p_{ij} = p_{ji} \le 1$ $(1 \le i \le j \le k)$ él-súlyokkal ($\mathbf{P} = (p_{ij})$, $\mathtt{rang}\,(\mathbf{P}) = k$), hogy $G_n \to W_H$ $(n \to \infty)$, a homomorfizmus-sűrűségek értelmében.

PI. $\mathbf{A}_n$-nek van k *strukturális* $\lambda_{1,n}, \ldots, \lambda_{k,n}$ s.értéke, hogy $\frac{1}{n}|\lambda_{i,n}| \to q_i$, $n \to \infty$ $(i = 1, \ldots, k)$ valamely $q_1, \ldots, q_k$ pozitív valósakkal; a többi s.érték pedig $o(n)$. A strukturális s.értékekhez tarozó s.vektorokkal reprezentálva $S^2_{k,n} = o(1)$, ahol a k-varianciát minimalizáló $(U_1, \ldots, U_k)$ partíció teljesíti az erős kiegyensúlyozottsági feltételt.

folytatás

PII. G_n-nek nincsenek domináns csúcsai és $\exists\, 0 < \delta < 1$ (független n-től, csak k-tól függhet), hogy $\mathbf{M}_{D,n}$-nek van $k-1$ *strukturális* s.értéke, melyekre $|\mu_{i,n}| \geq \delta$ $(i = 1, \ldots, k-1)$, míg $|\mu_{i,n}| = o(1)$ $(i \geq k$. A strukturális s.értékekhez tarozó transzformált s.vektorokkal reprezentálva $\tilde{S}^2_{k,n} = o(1)$, ahol a súlyozott k-varianciát minimalizáló $(U_1, \ldots, U_k)$ partíció teljesíti az erős kiegyensúlyozottsági feltételt.

PIII. Vannak $(U_1, \ldots, U_k)$ csúcs-klaszterek, melyek teljesítik az erős kiegyensúlyozottsági feltételt, és $\exists\, 0 < \theta < 1$ (független n-től, csak k-tól függhet), hogy $\mathtt{md}_1(G_n), \ldots, \mathtt{md}_{k-1}(G_n) \geq \theta$ és $\mathtt{md}_k(G_n; U_1, \ldots, U_k) = o(1)$.

folytatás

PIV. Vannak $(U_1, \ldots, U_k)$ csúcs-klaszterek $n_1, \ldots, n_k$ méretekkel $(\sum_{i=1}^{k} n_i = n)$, melyek teljesítik az erős kiegyensúlyozottsági feltételt, és van olyan $k \times k$-as, k rangú szimmetrikus $\mathbf{P} = (p_{ij})$ valószínűség-mátrix, hogy

$$\sum_{u,v \in U_i} |N_2(u, v; U_j) - p_{ij}^2 n_j| = o(p_{ij}^2 n_i^2 n_j) = o(n^3),\ \forall i, j = 1, \ldots, k$$

ahol $N_2(u, v; U_j)$ jelöli u, v közös szomszédainak számát U_j-ben. (Általánosított véletlen gráfnál tagonként is igaz.)

Tétel (B, arXiv:1508.04369v6)

$$PIV \iff P0 \Longrightarrow PI, PII; \quad PII \Longrightarrow PIII$$

PII és PIII megerősítése az ekvivalenciához

Legyen PII+ és PIII+ az eredeti PII és PIII megtoldva a következővel:

- Van olyan $k \times k$-as, k rangú szimmetrikus $\mathbf{P} = (p_{ij})$ valószínűség-mátrix, hogy

$$d(U_i, U_j) = \frac{e(U_i, U_j)}{|U_i| \cdot |U_j|} = p_{ij} + o(1) \quad (1 \leq i \leq j \leq k), \quad n \to \infty$$

- $\forall\, 1 \leq i \leq j \leq k$ és $u \in U_i$:

$$N_1(u; U_j) = (1 + o(1)) p_{ij} n_j.$$

Sejtés:

$$P0 \Longrightarrow PII+ \Longrightarrow PIII+ \Longrightarrow PIV \Longrightarrow P0$$

így P0, PII+, PIII+, PIV mind ekvivalensek.

A bizonyításokhoz használtam

- a Chung-Graham-Wilson, Combinatorica (1989) (egyrészes) és Thomason, A., Dense expanders and pseudo-random biparite graphs, DM (1989) (páros) kvázirandom becsléseket
- a 25. és 26. Tételt
- Borgs et al., Advances in Math. (2008) és Ann. Math. (2012) gráfkonvergencia és tesztelhető gráfparaméterek eredményeit
- $\mathbf{M}_D$ spektrumának és spektrális altereinek tesztelhetőségét

Jelölés: $P_{\mathbb{W}} : L^2(\xi') \to L^2(\xi)$ integrál-operátor, mely feltételes várható értéket vesz a W grafonhoz rendelt $\mathbb{W}$ (1-re normált) együttes eloszlás szerint, ahol ξ, ξ' azonos $\mathbb{D}$ eloszlásúak ($\mathbb{W}$ marginálisa), együttes eloszlásuk $\mathbb{W}$, és

$$L^2(\xi) = \{\xi : V \to \mathbb{R},\ \mathbb{E}_{\mathbb{D}}(\xi) = 0,\ \mathtt{Var}_{\mathbb{D}}(\xi) < \infty\}$$

Tétel (29. Tétel, B, EJC (2014))

$G_n = (V_n, \mathbf{A}_n) \to W$, G_n összefüggő súlyozott gráf [0,1]-beli él-súlyokkal, a csúcs-súlyok pedig az általános fokok. $\mathbf{M}_{D,n}$ spektruma: $|\mu_{n,1}| \geq |\mu_{n,2}| \geq \cdots \geq |\mu_{n,n}| = 0$. Akkor $\forall\, i \geq 1$: $\mu_{n,i} \to \mu_i(P_{\mathbb{W}})$ $(n \to \infty)$, ahol $\mu_i(P_{\mathbb{W}})$ a $P_{\mathbb{W}}$ kompakt, önadjungált operátor i-edik legnagyobb abszolút értékű s.értéke.

Tétel (30. Tétel, B, EJC (2014))

Tegyük fel, hogy van olyan $0 < \varepsilon < \delta \leq 1$, hogy $|\mu_{n,1}| \geq \cdots \geq |\mu_{n,k-1}| \geq \delta > \varepsilon \geq |\mu_{n,k}| \geq \cdots \geq |\mu_{n,n}| = 0$. Akkor $\mathrm{Span}\{\mathbf{D}_n^{-1/2}\mathbf{u}_{n,1}, \ldots, \mathbf{D}_n^{-1/2}\mathbf{u}_{n,k-1}\}$ konvergál $P_{\mathbb{W}}$ analóg alteréhez, azaz ha $\mathbf{P}_{n,k-1}$ jelöli a fenti altérre való vetítést, $\mathbf{P}_{k-1}$ pedig az analóg vetítést ($P_{\mathbb{W}}$ s.függvényeivel), akkor $\|\mathbf{P}_{n,k-1} - \mathbf{P}_{k-1}\| \to 0$, $n \to \infty$.

Következmány: $\tilde{S}_k^2$ szintén tesztelhető.

PI megerősítése, $P0 \Longrightarrow PI$, $PI+ \Longrightarrow P0$

Tétel (18. Tétel, B, LAA (2005))

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 klaszterezással történik.
Azt is 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.

EM algoritmus

Az eddigi módszerek nemparaméteresek voltak.
Félparaméteres modellek: klaszterezés + a csúcsok paraméterzezése tagság szerint.
Dempster, Laird és Rubin, J. Royal Statist. Soc. B (1977): Maximum likelihood from incomplete data via the EM algorithm.
Statisztikai minta egy n csúcson értelmezett egyszerű gráf $n \times n$-es, szimmetrikus szomszédsági mátrixa, $\mathbf{A} = (a_{ij})$. Látszólag egyetlen mintánk van, azonban a diagonális feletti elemeket, mint független valószínűségi változókat tekintjük statisztikai mintának.
A hiányos adatrendszer, mivel a csúcsok klaszterbe tartozását (tagságát) nem ismerjük. Ezért az **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.

Homogén sztochasztikus blokkmodell

- Adott k egészre $(1 < k < n)$ a csúcsok függetlenül tartoznak a C_i klaszterekbe π_i valószínűséggel, $i = 1, \ldots, k$; $\sum_{i=1}^{k} \pi_i = 1$.
- C_i és C_j csúcsai egymástól függetlenül,

$$\mathbb{P}(u \sim v | u \in C_i, v \in C_j) = p_{ij}, \quad 1 \leq i \leq j \leq k$$

valószínűséggel vannak összekötve.

A modell paraméterei:

- $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ *tagsági vektor*
- $k \times k$-as, szimmetrikus $\mathbf{P} = (p_{ij})$ *valószínűség-mátrix*

Inhomogén sztochasztikus blokkmodell

- Adott k egészre $(1 < k < n)$ a csúcsok függetlenül tartoznak a C_i klaszterekbe π_i valószínűséggel, $i = 1, \ldots, k$; $\sum_{i=1}^{k} \pi_i = 1$.
- Az $u \in C_i$ és $v \in C_j$ csúcsok egymástól függetlenül, p_{uv} valószínűséggel vannak összekötve, ahol

$$\ln \frac{p_{uv}}{1 - p_{uv}} = \beta_{uj} + \beta_{vi}, \quad 1 \leq i \leq j \leq k.$$

Csiszár V. et al., Algorithms (2012) és Rasch-modell (1961)
A modell paraméterei:

- $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ *tagsági vektor*
- $n \times k$-as $\mathbf{B} = (\beta_{uj})$ mátrix

A likelihood függvény a következő keverék

$$\sum_{1\le i,j\le k} \pi_i \pi_j \prod_{u\in C_i, v\in C_j} p_{uv}^{a_{uv}} (1-p_{uv})^{(1-a_{uv})}$$

$\mathbf{A} = (a_{uv})$: hiányos 0/1 adatmátrix
A homogán modellben: p_{uv} modell-paraméter
Az inhomogén modellben:

$$p_{uv} = \frac{e^{\beta_{uj}} e^{\beta_{vi}}}{1 + e^{\beta_{uj}} e^{\beta_{vi}}}, \qquad 1 - p_{uv} = \frac{1}{1 + e^{\beta_{uj}} e^{\beta_{vi}}}$$

ahol $u \in C_i$, $v \in C_j$, $\quad 1 \le i \le j \le k$

A fokszámsorozat elégséges statisztika, ez lép fel az ML becslésben.

Iteráció: kezdeti paraméterek és klaszterezés

E : kiszámoljuk Δ_u feltételes várható értékét az előző lépésbeli modell-paraméterek és tagságok alapján (feltételes várható érték képzés és Bayes-tétel) $\Longrightarrow$ csúcsok fuzzy klaszterezése (vagy hogy abba a klaszterbe soroljuk a csúcsot, amelybe a legnagyobb valószínűséggel tartozna)

M : az összes új i, j klaszterpárrapárra $(1 \leq i \leq j \leq k)$ külön-külön maximalizáljuk a likelihoodot a paraméterekben $\Longrightarrow$ a paraméterek új becslése $\Longrightarrow$ **E**, ...

Exponenciális eloszláscsaládban az algoritmus konvergenciája (bizonyos feltételek mellett, pl. a részgráfok nem esnek az Erdős–Gallai politóp határára) a likelihood-függvény egy lokális maximumához garantált, B, Elbanna, JPS (2015).
Spektrális klaszterezés jó kezdés (Ravi Kannan)

Köszönöm mindazoknak

- akik a témához közel hoztak és a bizonyításokban segítettek: Tusnády Gábor, Friedl Katalin, Krámli András, Bojan Mohar
- akiktől hasznos tanácsokat kaptam: T. Sós Vera, Lovász László, Simonovits Miklós
- akik pozitív időt biztosítottak mindehhez (2010: fél év sabbatical, 2014: egy év csökkentett oktatási terhelés): Horváth Miklós (intézet igazgató), Simon Károly (tanszékvezető)
- akik nélkül nem tudtam volna összegyűjteni és leírni mindezeket: Csizmadia Ákos (könyvtár) és egyéb technikai segítség
- kollégáknak, akik ötleteket adtak és diákok, akik adatokon alkalmazták az algoritmusokat: Kiss Gergő, BME és BSM diákok

A hallgatóságnak köszönöm a figyelmet

Revisiting the notion of graph convergence

Lovász, Szegedi, J. Comb. Theory B (2006)
Borgs et al., Ann. Math. (2012)
$G_n \to W_H$ means that for any simple graph F:

$$\frac{\hom(F, G_n)}{|V(G_n)|^{|V(F)|}} \to \hom(F, H) = \sum_{\psi: V(F) \to V(H)} \prod_{i \in V(F)} r_{\psi(i)} \prod_{ij \in E(F)} p_{\psi(i)\psi(j)}.$$

If $|V(F)| = m$, then

$$\hom(F, H) = \hom(F, W_H) = \int_{[0,1]^m} \prod_{\{i,j\} \in E(F)} W_H(x_i, x_j)\, dx_1 \dots dx_m.$$

Construction of a generalized quasirandom graph

Given k, $\mathbb{P}$, and vertex-weights of the model graph H: consider the instance when there are k sets $U_1, \dots, U_k \subset V$ of sizes $n_1, \dots, n_k$ such that $\frac{n_i}{n} = r_i$ $(i = 1, \dots, k)$. Let us choose the independent irrational numbers α_{ij} $(1 \le i \le j \le k)$.
Then the subgraph on the vertex-set U_i is constructed as follows:

$$u \sim v \Leftrightarrow \{(u-v)^2 \alpha_{ii}\} \le p_{ii}, \quad i = 1, \dots, k.$$

The bipartite subgraph between U_i and U_j: $v \in U_i$ and $u \in U_j$

$$u \sim v \Leftrightarrow \{(u-v)^2 \alpha_{ij}\} \le p_{ij}, \quad 1 \le i < j \le k.$$

Analytical number theoretical considerations guarantee that the above fractional parts are symmetrically well-distributed over $[0,1]^2$ if $n \to \infty$ and $\frac{n_i}{n} \to r_i$ $(i = 1, \dots, k)$. V. T. Sós, Pinch, G. Kiss

Sharp concentration theorem

Theorem

W *is an* $n \times n$ *real symmetric matrix, its entries in and above the main diagonal are independent random variables with absolute value at most 1.* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$*: eigenvalues of* **W**.
For any $t > 0$*:*

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

and the same estimate holds for the probability

$$\mathbb{P}\left(|\lambda_{n-i+1} - \mathbb{E}(\lambda_{n-i+1})| > t\right).$$

Alon, Krivelevich, Vu, Israel J. Math. (2002)

Consequence

Lemma

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

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

with probability tending to 1 as $n \to \infty$.

Borel–Cantelli lemma$\Rightarrow$

The spectral norm of $\mathbf{W}_n$ is $\mathcal{O}(\sqrt{n})$ almost surely.

P0$\Longrightarrow$PIV

We use the results of Chung–Graham–Wilson, Combinatorica (1989); Lovász–T. Sós, JCTB (2008) in view of which:
The vertex set of the generalized quasirandom graph G_n (defined by P0) can be partitioned into classes $U_1, \ldots, U_k$ in such a way that $\frac{|U_i|}{n} \to r_i$ $(i = 1, \ldots, k)$, that gives the strong balancing; the subgraph $G_{ii,n}$ is the general term of a quasirandom graph sequence with edge-density tending to p_{ii} $(i = 1, \ldots, k)$, whereas $G_{ij,n}$ is the general term of a bipartite quasirandom graph sequence with edge-density tending to p_{ij} $(i \neq j)$ as $n \to \infty$. Therefore, for the subgraphs, the equivalent statements of Chung–Graham–Wilson of the usual (1-class) quasirandomness are applicable, and similar considerations can be made for the bipartite subgraphs as well.

Chung–Graham–Wilson: Quasi-random graphs, $k = 1$, $p = \frac{1}{2}$

- $P_1(s)$: for all graphs $M(s)$ on s vertices,

$$N^*_{G_n}(M(s)) = (1+o(1))n^s(\frac{1}{2})^{\binom{s}{2}} \quad \text{labelled induced subgraphs.}$$

$$N^*_{G_n}(M(s)) = (1+o(1))n^s p^{|E(M(s))|}(1-p)^{\binom{s}{2}-|E(M(s))|}.$$

- $P_2(t)$: $e(G_n) \geq (1+o(1))\frac{n^2}{4}$, $N_{G_n}(C_t)) \leq (1+o(1))n^t(\frac{1}{2})^t$.

$$2e(G_n) \geq (1+o(1))pn^2, \quad \hom(C_t, G_n) \leq (1+o(1))n^t p^t.$$

Chung–Graham–Wilson continued

- P_3: $e(G_n) \geq (1+o(1))\frac{n^2}{4}$, $\lambda_1 = (1+o(1))\frac{n}{2}$, $\lambda_2 = o(n)$.

  $2e(G_n) \geq (1+o(1))pn^2, \quad \lambda_1 = (1+o(1))pn, \quad \lambda_2 = o(n).$

- P_4 : $\forall S \subset V, e(S) = \frac{1}{4}|S|^2 + o(n^2)$.
  $\forall X \subset V$: $\quad e(X,X) = p|X|^2 + o(n^2)$.
- P_7: $\sum_{u,v} |N_2(u,v) - \frac{n}{4}| = o(n^3)$,
  $\sum_{u,v} |N_2(u,v) - p^2 n| = o(n^3)$.

Then for $s \geq 4$ and $t \geq 4$ even,

$$P_2(4) \Rightarrow P_2(t) \Rightarrow P_1(s) \Rightarrow P_3 \Rightarrow \cdots \Rightarrow P_7 \Rightarrow P_2(4).$$

Quasirandom graph: satisfies any (all) of the above properties.

Lemma 1.

Lemma

If $(G_{ii,n})$ is quasirandom, then

$$\sum_{u,v\in U_i} N_2(u,v;U_i) \geq (1+o(1))p_{ii}^2 n_i^3, \quad i=1,\ldots,k.$$

Proof: We drop the index n of the adjacency entries.

$$\sum_{u,v\in U_i} N_2(u,v;U_i) = \sum_{u,v\in U_i}\sum_{t\in U_i} a_{ut}a_{vt}$$

$$= \sum_{t\in U_i}\sum_{u\in U_i} a_{ut}\sum_{v\in U_i} a_{vt} = \sum_{t\in U_i}[N_1(t;U_i)]^2 \geq \frac{1}{n_i}[\sum_{t\in U_i} N_1(t;U_i)]^2$$

$$= \frac{1}{n_i}[2e(U_i)]^2 \geq \frac{1}{n_i}[(1+o(1))p_{ii}n_i^2]^2 = (1+o(1))p_{ii}^2 n_i^3,$$

Proof of Lemma 1, continued

where $N_1(t; U_i)$ denotes the number of neighbors of t in U_i, and $e(U_i)$ is the number of edges within the induced subgraph $G_{ii,n}$ of G_n, induced by U_i. In the first inequality we used the Cauchy–Schwarz, and in the second one, the first part of the equivalent quasirandom property P_2 of Chung–Graham–Wilson.

Lemma 2.

Lemma

If $(G_{ij,n})$ is bipartite quasirandom, then

$$\sum_{u,v\in U_i} N_2(u,v;U_j) = (1+o(1))p_{ij}^2 n_i^2 n_j, \quad i \neq j.$$

Proof:

$$\sum_{u,v\in U_i} N_2(u,v;U_j) = \sum_{u,v\in U_i}\sum_{t\in U_j} a_{ut}a_{vt}$$

$$= \sum_{t\in U_j}\sum_{u\in U_i} a_{ut}\sum_{v\in U_i} a_{vt} = \sum_{t\in U_j}[N_1(t;U_i)]^2 \geq \frac{1}{n_j}[\sum_{t\in U_j} N_1(t;U_i)]^2$$

$$= \frac{1}{n_j}[e(U_i,U_j)]^2 \geq \frac{1}{n_j}[(1+o(1))p_{ij}n_i n_j]^2 = (1+o(1))p_{ij}^2 n_i^2 n_j,$$

Proof of Lemma 2, continued

where $e(U_i, U_j)$ is the number of cut-edges between U_i and U_j, i.e., the number of edges in the induced bipartite subgraph $G_{ij,n}$ of G_n, induced by the U_i, U_j pair. Here, in the first inequality we used the Cauchy–Schwarz, and in the second one, the equivalent quasirandom property of bipartite quasirandom graphs.

P0$\Longrightarrow$PIV

In view of the lemmas and the the Cauchy–Schwarz inequality:

$$\left[\sum_{u,v\in U_i}|N_2(u,v;U_j)-p_{ij}^2 n_j|\right]^2 \le n_i^2 \sum_{u,v\in U_i}|N_2(u,v;U_j)-p_{ij}^2 n_j|^2$$

$$= n_i^2\left\{\sum_{u,v\in U_i}[N_2(u,v;U_j)]^2 - 2p_{ij}^2 n_j \sum_{u,v\in U_i} N_2(u,v;U_j) + n_i^2(p_{ij}^2 n_j)^2\right\}$$

$$\le n_i^2\left\{(1+o(1))p_{ij}^4 n_i^2 n_j^2 - 2(1+o(1))p_{ij}^4 n_i^2 n_j^2 + p_{ij}^4 n_i^2 n_j^2\right\}$$

$$= n_i^2 o(1) p_{ij}^4 n_i^2 n_j^2 = o(p_{ij}^4 n_i^4 n_j^2),$$

Proof continued

$\sum_{u,v\in U_i}[N_2(u,v;U_j)]^2 \sim \hom(C_4, G_{ij,n})$

- $i=j$: by $P_2(4)$,

$$\hom(C_4, G_{ii,n}) \le (1+o(1))p_{ii}^4 n_i^4.$$

- $i\neq j$: by Lovász–Sós (bipartite quasirandom graphs),

$$\frac{\hom(C_4, G_{ij,n})}{n_i^2 n_j^2} = (1+o(1))p_{ij}^4.$$

Only 4-cycles in the above bipartition have to be considered; these 4-cycles have 2 vertices from U_i and 2 from U_j, and any 2 of the common neighbors of $u, v \in U_i$ in U_j are possible candidates to close a (labelled) 4-cycle with them.

Proof of PIV$\Longrightarrow$P0

By C-G-W $P_7 \Rightarrow P_1(s)$, the subgraphs $G_{ii,n}$ are quasirandom. Likewise, if $i \neq j$, the bipartite subgraphs $G_{ij,n}$ are bipartite quasirandom.
Therefore, G_n is built of quasirandom and bipartite quasirandom blocks, so under the strong balancing condition, they together form a generalized quasirandom graph sequence on k classes and model graph H, the vertex-weights of which are $r_1, \ldots, r_k$ of the strong balancing condition, and the edge-weights are entries of the probability matrix $\mathbb{P} = (p_{ij})$.

For P0$\Longrightarrow$PII we use the following

Theorem (B, EJC (2014))

$G_n = (V_n, \mathbf{A}_n) \to W$, G_n connected with edge-weights in [0,1] and the vertex-weights are the generalized degrees. Assume that there are no dominant vertices. $|\mu_{n,1}| \geq |\mu_{n,2}| \geq \cdots \geq |\mu_{n,n}| = 0$ is the spectrum of $\mathbf{M}_{D,n}$.
Let $\mu_i(P_{\mathbb{W}})$ be the i-th largest absolute value eigenvalue of the integral operator $P_{\mathbb{W}} : L^2(\xi') \to L^2(\xi)$ taking conditional expectation with respect to the joint measure $\mathbb{W}$ embodied by the normalized limit graphon W, and ξ, ξ' are identically distributed random variables with the marginal distribution of their symmetric joint distribution $\mathbb{W}$.
Then for every $i \geq 1$: $\mu_{n,i} \to \mu_i(P_{\mathbb{W}})$ as $n \to \infty$.

P0$\Longrightarrow$PI

We use Theorem 6.7 of Borgs et al., Ann. Math. (2012), where the authors prove that if the sequence (W_{G_n}) of graphons converges to the limit graphon W, then both ends of the spectra of the integral operators, induced by W_{G_n}'s as kernels (these are the numbers $\frac{1}{n}\lambda_{i,n}$), converge to the ends of the spectrum of the integral operator induced by W as kernel. We apply this argument for the limit graphon W_H of (G_n). The same argument as in P0 $=>$ PII can be applied to the convergence of the spectral subspaces, so the convergence of the k-variances is also obtained. The steps are proportional to r_i's $\Longrightarrow$ strong balancing.
PI does not necessarily implies P0!

Strengthening of PI

PI+: $\mathbf{A}_n$ has k *structural* eigenvalues $\lambda_{1,n}, \ldots, \lambda_{k,n}$ such that the normalized eigenvalues converge: $\frac{1}{n}\lambda_{i,n} \to q_i$ as $n \to \infty$ $(i = 1, \ldots, k)$ with some non-zero reals $q_1, \ldots, q_k$, and the remaining eigenvalues are $o(\sqrt{n})$. Further, the k-variance $S^2_{k,n}$ of the k-dimensional vertex representatives, based on the eigenvectors corresponding to the structural eigenvalues of $\mathbf{A}_n$, is $o(\frac{1}{n})$. The k-partition $P_{k,n} = (U_{1n}, \ldots, U_{kn})$ of the vertices of G_n minimizing this k-variance satisfies: $\frac{|U_{in}|}{n} \to r_i$ with some r_i $(i = 1, \ldots, k)$. Also assume that there is a $k \times k$ symmetric probability matrix $\mathbb{P} = (p_{ij})$ of rank k such that

$$d(U_{in}, U_{jn}) := \frac{e(U_{in}, U_{jn})}{|U_{in}||U_{jn}|} \to p_{ij} \quad (1 \le i \le j \le k), \quad n \to \infty. \quad (1)$$

(I.e., the within- and between-cluster edge densities converge to the entries of $\mathbb{P}$.)

PI+$\Longrightarrow$P0

By B, LAA (2005) we are able to find a blown-up matrix $\mathbf{B}_n$ of rank k and an error-matrix $\mathbf{E}_n$ with $\|\mathbf{E}_n\| = o(\sqrt{n})$ such that $\mathbf{A}_n = \mathbf{B}_n + \mathbf{E}_n$ $(n = k, k+1, \dots)$. Say $\mathbf{B}_n$ is the blown-up matrix of the $k \times k$ pattern matrix $\mathbb{P}_n$, the ij entry $p_{ij}^{(n)}$ of which is the common entry of the $U_{in} \times U_{jn}$ block of $\mathbf{B}_n$.
Then using the relation between the cut-norm of a graphon and a matrix, further, between the cut-norm and the spectral norm of a matrix, and the transformation of a graph into graphon, we get that

$$\|W_{\mathbf{E}_n}\|_\square \le \frac{1}{n^2}\|\mathbf{E}_n\|_\square \le \frac{1}{n^2} n \|\mathbf{E}_n\| = \frac{1}{n} o(\sqrt{n}) = o(n^{-1/2}),$$

where $\|\mathbf{E}_n\|$ is the spectral-norm, $\|\mathbf{E}_n\|_\square$ is the matrix cut-norm of $\mathbf{E}_n$, and $W_{\mathbf{E}_n}$ denotes the graphon corresponding to the symmetric matrix $\mathbf{E}_n$ of uniformly bounded entries.

Using the Steiner equality, we get that the squared Frobenius norm of $\mathbf{A}_n - \mathbf{B}_n$, restricted to the ij block, is

$$\|(\mathbf{A}_n - \mathbf{B}_n)_{ij}\|_F^2 = \sum_{u \in U_{in}} \sum_{v \in U_{jn}} (a_{uv}^{(n)} - p_{ij}^{(n)})^2$$
$$= \sum_{u \in U_{in}} \sum_{v \in U_{jn}} (a_{uv}^{(n)} - d(U_{in}, U_{jn}))^2 + |U_{in}||U_{jn}|(d(U_{in}, U_{jn})$$

where the edge-density $d(U_{in}, U_{jn})$ is now viewed as the average of the entries of $\mathbf{A}_n$ in the $U_{in} \times U_{jn}$ block. Then by the inequality between the Frobenius and spectral norms,

$$\|(\mathbf{A}_n - \mathbf{B}_n)_{ij}\|_F^2 \leq n\|\mathbf{A}_n - \mathbf{B}_n\|^2 = n\|\mathbf{E}_n\|^2 = no^2(\sqrt{n}).$$

Therefore, for every $1 \leq i \leq j \leq k$ pair: $(d(U_{in}, U_{jn}) - p_{ij}^{(n)})^2 \leq \frac{1}{|U_i||U_j|} no^2(\sqrt{n}) = \frac{1}{\frac{|U_{in}|}{n}\frac{|U_{jn}|}{n}} n(\frac{o(\sqrt{n})}{n})^2 = no^2(n^{-1/2})$ as $\frac{|U_{in}|}{n} \to r_i$ when $n \to \infty$ $(i = 1, \ldots, k)$.

Eventually, we prove the $G_n \to W_H$ convergence by proving that the cut-distance between the corresponding graphons tends to 0.

H is a model graph with vertex-weights r_i's and edge-weights p_{ij}'s in the PI+ conditions.

Using the triangle inequality, we get

$$\|W_{G_n} - W_H\|_\square \leq \|W_{G_n} - W_{\mathbf{B}_n}\|_\square + \|W_{\mathbf{B}_n} - W_{G_n/P_{k,n}}\|_\square + \|W_{G_n/P_{k,n}} - W_H\|_\square$$

were $G_n/P_{k,n}$ is the factor graph of G_n with respect to the k-partition $P_{k,n}$. This is an edge- and vertex-weighted graph on k vertices, with vertex-weights $\frac{|U_{in}|}{n}$ and edge-weights $d(U_{in}, U_{jn})$, $i, j = 1, \ldots, k$.

The first term is $\|W_{\mathbf{E}_n}\|_\square = o(n^{-1/2})$. To estimate the second term, observe that because $\mathbf{B}_n$ is the blown-up matrix of $\mathbb{P}_n$ with respect to the k-partition $P_{k,n}$, after conveniently permuting its rows (and columns, accordingly). The gaphon $W_{\mathbf{B}_n}$ is also stepwise constant over the unit square, where the sides are divided into k parts: the interval I_j has lengths $\frac{|U_{jn}|}{n}$ $(j = 1, \ldots, k)$, and over $I_i \times I_j$ the stepfunction takes on the value $p_{ij}^{(n)}$. By its nature, the graphon $W_{G_n/P_{k,n}}$ is stepwise constant with the same subdivision of the unit square, and over $I_i \times I_j$ it takes on the value $d(U_{in}, U_{jn})$, $i, j = 1, \ldots, k$. But in view of the above, $\|W_{\mathbf{B}_n} - W_{G_n/P_{k,n}}\|_\square = \sqrt{n} o(n^{-1/2}) = o(1)$. The third term is $o(1)$, because of the assumptions $\frac{|U_{in}|}{n} \to r_i$ $(i = 1, \ldots, k)$ and $d(U_{in}, U_{jn}) \to p_{ij}$, $i, j = 1, \ldots, k$. Therefore, $\|W_{G_n} - W_H\|_\square = o(1)$ and so, $G_n \to H$, which finishes the proof.

Lemma

Under P0, the following holds for except $o(n_i)$ vertices $u \in U_i$, and for every $i = 1, \ldots, k$: $N_1(u; U_i) = (1 + o(1))p_{ii}n_i$.
Under P0, the following holds for except $o(n_i)$ vertices $u \in U_i$, and for every $1 \leq i < j \leq k$: $N_1(u; U_j) = (1 + o(1))p_{ij}n_j$.

The statement follows from the $P_1(s)$ $(\forall s) => P_0'$ implication of C-G-W and its bipartite analogue.

The subgraphs ara almost-regular, the bipartite subgraphs are almost-biregular: weaker than quasirandomness.

Proof in the $i = j$ case

Let $(U_1, \dots, U_k)$ be the k-partition, guaranteed by PIII+, such that $\mathtt{md}_k(G_n; U_1, \dots U_k) = o(1)$. Then by the extra conditions of PIII+, for $X \subset U_i$, $\mathtt{Vol}(X) = |X|(1 + o(1)) \sum_{\ell=1}^{k} p_{i\ell} n_\ell$, and so,

$$
\begin{aligned}
& e(X,X) - p_{ii}|X|^2 = e(X,X) - [d(U_i,U_i) + o(1)]|X|^2 \\
& = e(X,X) - \frac{e(U_i,U_i)}{\frac{\mathtt{Vol}^2(U_i)}{(1+o(1))^2(\sum_{\ell=1}^{k} p_{i\ell} n_\ell)^2}} \frac{\mathtt{Vol}^2(X)}{(1+o(1))^2(\sum_{\ell=1}^{k} p_{i\ell} n_\ell)^2} - o(1)|X|^2 \\
& = [e(X,X) - \rho(U_i,U_i)\mathtt{Vol}^2(X)] - o(1)\rho(U_i,U_i)\mathtt{Vol}^2(X) - o(1)|X|^2 \\
& \le \mathtt{md}_k(G_n; U_1, \dots U_k)\sqrt{\mathtt{Vol}^2(X)} - o(1)e(U_i,U_i)\left(\frac{\mathtt{Vol}(X)}{\mathtt{Vol}(U_i)}\right)^2 - o(n^2) \\
& = o(n^2).
\end{aligned}
$$

Then P_4 implies P_7 of Chung–Graham–Wilson, that is our PIV.

Proof of Theorem 25.

By an easy analysis of variance argument it follows that

$$s^2 = \sum_{i=1}^{k} \mathtt{dist}^2(\mathbf{u}_i, F),$$

where $F = \mathtt{Span}\{\mathbf{D}^{1/2}\mathbf{z}_1, \ldots, \mathbf{D}^{1/2}\mathbf{z}_k\}$ with the so-called normalized partition vectors $\mathbf{z}_1, \ldots, \mathbf{z}_k$ of coordinates

$z_{ji} = \frac{1}{\sqrt{\mathtt{Vol}(V_i)}}$ if $j \in V_i$ and 0, otherwise $(i = 1, \ldots, k)$.

The vectors $\mathbf{D}^{1/2}\mathbf{z}_1, \ldots, \mathbf{D}^{1/2}\mathbf{z}_k$ form an orthonormal system.

By B, Tusnády, Discrete Math (1994), we can find another orthonormal system $\mathbf{v}_1, \ldots, \mathbf{v}_k \in F$ such that

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

We approximate the matrix $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ ($\lambda_1 = 1$, $\mathbf{u}_1 = \sqrt{\mathbf{d}}$)
by the rank k matrix $\sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ with the following accuracy (in spectral norm):

$$\left\| \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T - \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \right\| \leq \sum_{i=1}^{k} |\lambda_i| \cdot \left\| \mathbf{u}_i \mathbf{u}_i^T - \mathbf{v}_i \mathbf{v}_i^T \right\| + \left\| \sum_{i=k+1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \right\|$$

This is further estimated from above with

$$\sum_{i=1}^{k} \sin \alpha_i + \varepsilon \leq \sum_{i=1}^{k} \|\mathbf{u}_i - \mathbf{v}_i\| + \varepsilon \leq \sqrt{2ks} + \varepsilon$$

where α_i is the angle between $\mathbf{u}_i$ and $\mathbf{v}_i$, and for it,

$$\sin \frac{\alpha_i}{2} = \frac{1}{2}\|\mathbf{u}_i - \mathbf{v}_i\|$$

holds, $i = 1, \ldots, k$.

Based on these considerations and relation between the cut norm and the spectral norm, the densities to be estimated in the defining formula of volume regularity can be written in terms of stepwise constant vectors in the following way.
$\mathbf{y}_i := \mathbf{D}^{-1/2}\mathbf{v}_i$ is stepwise constants on the partition $(V_1, \ldots, V_k)$, $i = 1, \ldots, k$

$\Longrightarrow \sum_{i=1}^{k} \lambda_i \mathbf{y}_i \mathbf{y}_i^T$ is a symmetric block-matrix on $k \times k$ blocks belonging to the above partition of the vertices.

Let $\widetilde{w}_{ab}$ denote its entries in the (a, b) block $(a, b = 1, \ldots, k)$.

The rank k approximation of the matrix $\mathbf{W}$ is performed with the following accuracy of the perturbation $\mathbb{E}$:

$$\|\mathbb{E}\| = \left\| \mathbf{W} - \mathbf{D}(\sum_{i=1}^{k} \lambda_i \mathbf{y}_i \mathbf{y}_i^T)\mathbf{D} \right\| =$$

$$= \left\| \mathbf{D}^{1/2}(\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} - \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T)\mathbf{D}^{1/2} \right\|$$

Consequently, the entries of $\mathbf{W}$ – for $i \in V_a$, $j \in V_b$ – can be decomposed as

$$w_{ij} = d_i d_j \widetilde{w}_{ab} + \eta_{ij}$$

where the cut norm of the $n \times n$ symmetric error matrix $\mathbf{E} = (\eta_{ij})$ restricted to $V_a \times V_b$ (otherwise it contains entries all zeroes) and denoted by $\mathbb{E}_{ab}$, is estimated as follows:

$$\|\mathbf{E}_{ab}\|_{\square} \leq C\sqrt{\mathtt{Vol}(V_a)}\sqrt{\mathtt{Vol}(V_b)}(\sqrt{2k}s + \varepsilon),$$

where the constant C does not depend on n (due to the balancing conditions on the vertex degrees and cluster sizes).

Summarizing, for $a, b = 1, \ldots, k$ and $X \subset V_a$, $Y \subset V_b$:

$$|w(X,Y) - \rho(V_a,V_b)\mathtt{Vol}(X)\mathtt{Vol}(Y)| =$$

$$\left|\sum_{i\in X}\sum_{j\in Y}(d_i d_j \widetilde{w}_{ab} + \eta_{ij}^{ab}) - \frac{\mathtt{Vol}(X)\mathtt{Vol}(Y)}{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)}\sum_{i\in V_a}\sum_{j\in V_b}(d_i d_j \widetilde{w}_{ab} + \eta_{ij}^{ab})\right|$$

$$= \left|\sum_{i\in X}\sum_{j\in Y}\eta_{ij}^{ab} - \frac{\mathtt{Vol}(X)\mathtt{Vol}(Y)}{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)}\sum_{i\in V_a}\sum_{j\in V_b}\eta_{ij}^{ab}\right|$$

$$\leq 2C(\sqrt{2ks} + \varepsilon)\sqrt{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)}$$

that gives the required statement both in the $a \neq b$ and $a = b$ case.

Proof of Theorem 26.

$$\mathbf{F} := \mathbf{W} - \mathbf{DRD},$$

where $\mathbf{R} = (\rho(C_a, C_b))$ is the $n \times n$ block-matrix of $k \times k$ blocks with entries equal to $\rho(C_a, C_b)$ over the block $C_a \times C_b$. With the indicator vectors $\mathbf{1}_X$ and $\mathbf{1}_Y$ of $X \subset C_a$ and $Y \subset C_b$,

$$|\langle \mathbf{1}_X, \mathbf{F}\mathbf{1}_Y \rangle| \leq \alpha^* \sqrt{\langle \mathbf{1}_X, \mathbf{W}\mathbf{1}_n \rangle \langle \mathbf{1}_m, \mathbf{W}\mathbf{1}_Y \rangle},$$

where $\mathbf{1}_n$ denotes the all 1's vector of size n.

$$\mathbf{D}^{-1/2}\mathbf{F}\mathbf{D}^{-1/2} = \mathbf{W}_D - \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}.$$

Since the rank of the matrix $\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$ is at most k, by the estimate of Thompson, we obtain the following upper estimate for $|\mu_k|$:

$$|\mu_k| \leq s_{max}(\mathbf{D}^{-1/2}\mathbf{F}\mathbf{D}^{-1/2}) = \|\mathbf{D}^{-1/2}\mathbf{F}\mathbf{D}^{-1/2}\|,$$

where $\|.\|$ denotes the spectral norm.

Let $\mathbf{v} \in \mathbb{R}^n$ be the left and $\mathbf{u} \in \mathbb{R}^n$ be the right unit-norm singular vector corresponding to the maximal singular value of $\mathbf{D}^{-1/2}\mathbf{F}\mathbf{D}^{-1/2}$ ($\mathbf{u} = \pm\mathbf{v}$).
In view of Lemma (Bollobás–Nikiforov), there is a step-vector $\mathbf{x} \in \mathbb{C}^n$ such that $\|\mathbf{v} - \mathbf{D}^{1/2}\mathbf{x}\| \leq \frac{1}{3}$ and $\|\mathbf{D}^{1/2}\mathbf{x}\| \leq 1$.
With them, we estimate the discrepancies.

Hilbert-terek, együttes eloszlások

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 maximálkorrelácó-keresési feladatot, melynek első lépése a Rényi-féle maximálkorreláció meghatározása; véges diszkrét esetben pedig a korrespondenciaanalízis feladatát kapjuk.

Az itt felsorolt technikákkal nem csupán egységesen kezelhetők az előző tézisekben kitűzött feladatok, de az absztrakció szintén segítségemre lesz a 9. Tézisben kimondott tesztelhetőségi tételek bizonyításánál (végtelen élsúlyozott gráf- vagy kontingenciatábla sorozatokat tekintünk, melyeknek határértékei az együttes eloszlás szerinti feltételes várható értéket vevő integráloperátor magfüggvényei lesznek).

Valószínűségi változók Hilbert-terei

(ξ, η) valós értékű valószínűségi változópár az $\mathcal{X} \times \mathcal{Y}$ szorzatér felett.
Együttes eloszlásuk $\mathbb{W}$, a $\mathbb{P}$ és $\mathbb{Q}$ marginálisokkal.
Tfh. 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, 1959).
$H = L^2(\xi)$ ill. $H' = L^2(\eta)$: a ξ, ill. η valószínűségi változók $\mathbb{P}$, ill. $\mathbb{Q}$ mérték 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 (Breiman és Friedman, ACE algoritmus, 1985).

Feltételes várható érték képzés operátor

Integráloperátor w magfüggvénnyel:

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

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

$P_{\mathcal{X}}$ és $P_{\mathcal{Y}}$ geometriailag vetítések, és egymás adjungáltjai a

$$\langle P_{\mathcal{X}}\phi, \psi\rangle_H = \langle P_{\mathcal{Y}}\psi, \phi\rangle_{H'} = \mathtt{Cov}\,_{\mathbb{W}}(\psi, \phi)$$

reláció miatt, ahol a $\mathtt{Cov}\,_{\mathbb{W}}$ kovariancia-függvény:

$$\begin{aligned}\mathtt{Cov}\,_{\mathbb{W}}(\psi, \phi) &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x)\phi(y)\mathbb{W}(dx, dy) \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(x)\phi(y)w(x,y)\mathbb{Q}(dy)\mathbb{P}(dx)\end{aligned}$$

Szőkefalvi–Nagy és Riesz (1952), Rényi (1959)

Tegyük fel, hogy

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} w^2(x,y)\mathbb{Q}(dy)\mathbb{P}(dx) < \infty.$$

Diszkrét $\{w_{ij}\}$ együttes és $\{p_i\}$ $(p_i = \sum_j w_{ij})$, $\{q_j\}$ $(q_j = \sum_i w_{ij})$ marginális eloszlások esetén ez azt jelenti, hogy

$$\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} \left(\frac{w_{ij}}{p_i q_j}\right)^2 p_i q_j = \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} \frac{w_{ij}^2}{p_i q_j} < \infty,$$

míg abszolút folytonos eloaszlás esetén, $f(x,y)$ együttes és $f_1(x)$ $(f_1(x) = \int f(x,y)\,dy)$, $f_2(y)$ $(f_2(y) = \int f(x,y)\,dx)$ marginális sűrűségekkel pedig, hogy

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \left(\frac{f(x,y)}{f_1(x)f_2(y)}\right)^2 f_1(x)f_2(y)\,dx\,dy = \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{f^2(x,y)}{f_1(x)f_2(y)}\,dx\,dy < \infty.$$

Spektrális és szinguláris felbontások

Ekkor $P_{\mathcal{X}}$ és $P_{\mathcal{Y}}$ Hilbert–Schmidt operátorok $\Longrightarrow$ kompaktak (teljesen folytonosak): diszkrét spektrumuk van.

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

ahol $1 > s_1 \geq s_2 \geq \cdots \geq 0$ (ha megszámlálhatóan végtelen sok van belőlük, akkor 0-hoz torlódnak). Amennyiben $\mathbb{W}$ szimmetrikus (H és H' izomorfak olyan értelemben is, hogy azonos eloszlású elemeik egy-egyértelműen egymáshoz rendelhetők), $P_{\mathcal{X}} = P_{\mathcal{Y}}$ önadjungált lineáris operátor. Ekkor $P_{\mathcal{X}} : H' \to H$ 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

$$P_{\mathcal{X}} \psi_i' = \lambda_i \psi_i$$

Maximálkorrelácó (Gebelein és Rényi)

Keresendő ξ és η függvényei körében az a $\psi \in H$, $\phi \in H'$ pár, melyek korrelációja a $\mathbb{W}$ együttes eloszlás szerint maximális:

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

és a maximum a ψ_1, ϕ_1 páron éretik el.
$s_1 = 0$ pontosan akkor, ha ξ és η függetlenek.
Ekvivalens minimumkeresési feladat:

$$\min_{\|\psi\|=\|\phi\|=1} \|\psi-\phi\|^2 = \min_{\|\psi\|=\|\phi\|=1} (\|\psi\|^2+\|\phi\|^2-2\mathrm{Cov}_{\mathbb{W}}(\psi, \phi)) = 2(1-s_1).$$

Korrespondenciaanalízis

Szorzattér: kontingenciatábla $w_{ij} \geq 0$ elemekkel
$(\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} = 1)$.
$\mathcal{X} = \{1, \ldots, m\}$: sorok, $\mathcal{Y} = \{1, \ldots, n\}$: oszlopok.
Marginálisok: $p_1, \ldots, p_m$ és $q_1, \ldots, q_n$.
$P_{\mathcal{X}} : H' \to H$, $P_{\mathcal{X}}\phi = \psi$ operátor hatása:

$$\psi(i) = \frac{1}{p_i} \sum_{j=1}^{n} w_{ij}\phi(j) = \sum_{j=1}^{n} \frac{w_{ij}}{p_i q_j}\phi(j)q_j, \quad i = 1, \ldots, m.$$

$P_{\mathcal{X}}$ integráloperátor a $\frac{w_{ij}}{p_i q_j}$ magfüggvénnyel és SVD-je a

$$\sqrt{p_i}\psi(i) = \sum_{j=1}^{n} \frac{w_{ij}}{\sqrt{p_i}\sqrt{q_j}}(\sqrt{q_j}\phi(j)), \quad i = 1, \ldots, m.$$

miatt a $\mathbf{W}_{corr} = \mathbf{P}^{-1/2}\mathbf{W}\mathbf{Q}^{-1/2}$ mátrix $\sum_{k=0}^{r-1} s_k \mathbf{v}_k \mathbf{u}_k^T$ SVD-jéből adódik. Az s_i-hez tartozó ψ_i, ϕ_i függvénypár lehetséges felvett értékei a $\mathbf{P}^{-1/2}\mathbf{v}_i$, $\mathbf{Q}^{-1/2}\mathbf{u}_i$ vektor koordinátái $(i = 1, \ldots, r-1)$

Reprezentációs tétel együttes eloszlásokra

Definíció

Az $(\mathbf{X}, \mathbf{Y})$ *k-dimenziós véletlen vektorpár – ahol* $\mathbf{X}$ *ill.* $\mathbf{Y}$ *koordinátái H- ill. H′-beliek,* X_i *és* Y_j *korrelálatlanok, ha* $i \neq j$*, különben pedig* X_i *és* Y_i *együttes eloszlása* $\mathbb{W}$ *– a* $\mathbb{W}$ *együttes eloszlás k-dimenziós reprezentációját valósítja meg, ha* $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$, $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_k$, *és reprezentáció költsége*

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

Tétel

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 k legnagyobb szinguláris értéke pozitív:* $1 > s_1 \geq s_2 \geq \cdots \geq s_k > 0$*. Akkor a fenti k-dimenziós reprezentáció minimális költsége* $2\sum_{i=1}^{k}(1 - s_i)$ *és a minimum a* $(\psi_1, \ldots, \psi_k)$ *és* $(\phi_1, \ldots, \phi_k)$

Reprezentációs tétel szimmetrikus együttes eloszlásokra

Definíció

Az $\mathbf{X}$ *k-dimenziós véletlen vektor – koordinátái H-beliek – a* $\mathbb{W}$ *együttes eloszlás k-dimenziós reprezentációja, ha* $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$. *A reprezentáció költsége* $Q_k(\mathbf{X}) = \mathbb{E}_{\mathbb{W}}\|\mathbf{X} - \mathbf{X}'\|^2$, *ahol* $\mathbf{X}$ *és* $\mathbf{X}'$ *azonos eloszlásúak;* X_i *and* X_i' *együttes eloszlása* $\mathbb{W}$, *míg* X_i *és* X_j' *korrelálatlanok* $(i \neq j)$.

Tétel

Legyen $\mathbb{W}$ *szimmetrikus együttes eloszlás a* $\mathbb{P}$ *marginálissal. Tegyük fel, hogy a* $P_{\mathcal{X}} : H' \to H$ *feltételes várható érték vevés operátorának (H és H′ izomorfak) k legnagyobb sajátértéke pozitív:* $1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$. *Akkor a fenti k-dimenziós reprezentáció minimális költsége* $2\sum_{i=1}^{k}(1 - \lambda_i)$ *és a minimum a* $(y_1, \ldots, y_k)$ *optimális k-dimenziós reprezentánssal érhető el.*

A szimmetrikus maximálkorreláció és RKHS

Szimmetrikus $\mathbb{W}$ esetén is a Rényi-féle maximálkorreláció a feltételes várható érték vevés operátorának legnagyobb szinguláris értéke, vagy ami ezzel ekvivalens, sajátértékéi abszolút értékének a maximuma, azaz mindig pozitív. A legnagyobb sajátérték az ún. szimmetrikus maximálkorrelációt adja, ami azonos eloszlású függvénypáron vétetik fel (és nem feltétlenül pozitív):

$$r_1 = \max_{\psi,\psi' \text{ i.d.}} \mathtt{Corr}_{\mathbb{W}}(\psi, \psi').$$

A Cheeger-egyenlőtlenség miatt

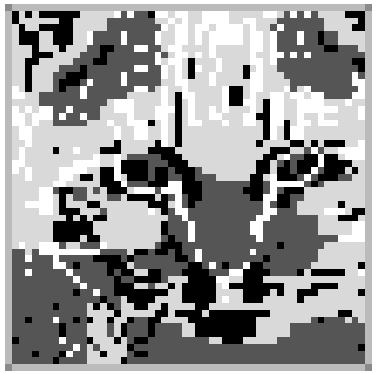
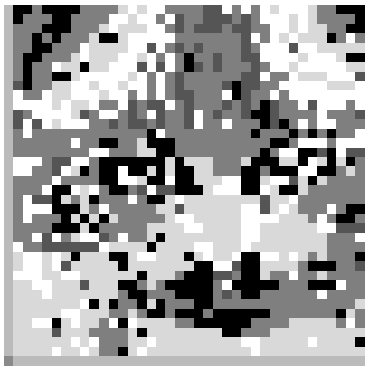
$$\frac{1-r_1}{2} \leq \min_{\substack{B\subset\mathbb{R}\,\text{Borel-h.}\\ \psi,\psi' \text{ i.d.}\\ \P_{\mathbb{D}}(\psi\in B)\leq 1/2}} \mathbb{P}_{\mathbb{W}}(\psi' \in \overline{B} | \psi \in B) \leq \sqrt{1-r_1^2}, \quad \text{ha} \quad r_1 > 0.$$

(Az $r_1 > 0$ feltétel ekvivalens a $\lambda_1' < 1$ feltétellel.)

Reprodukáló magú Hilbert-terek

Eredeti kép és a pixelek 3, 4, 5 színnel (klaszterrel)



$(48 \times 48$ pixel)

M_D strukturális sajátértékei:

$0.137259, \quad 0.014255, \quad 0.000925,$

$-0.0006707, \quad -0.0006706, \ldots$

Gauss-mag

image segmentation

Expander mixing lemma és megfordításai

$k = 1$ esetben „expander mixing lemma” irreguláris G gráfra:

Tétel (Chung-Graham, RSA (2008))

$$\mathtt{md}_1(G) \leq \|\mathbf{M}_D\| = |\mu_1| \iff \|K_G\|_{\boxtimes}\ (jumble\ norm) \leq \|K_G\|$$

$K_G : L_2([0,1]^2 \to \mathbb{R}$ magfv, mely a W_G grafonhoz tartozik (a csúcsok az általános fokokkal vannak súlyozva).
$K_{ij} = \frac{w_{ij}}{d_i d_j} - 1$
K_G s.értékei $=\mathbf{M}_D$ s.értékei