Spektrális klaszterezés (gráfok és kontingenciatáblák)

Bolla Marianna
BME Matematika Intézet
Sztochasztika Tanszék
(TÁMOP-4.2.2.C-11/1/KONV-2012-0001)
marib@math.bme.hu

Budapest, 2014. március 20.

machine learning. The fundamental object we are interested in is a vertex function or 0-chain, thus the linear operator we are looking for should operate on 0-chains. Notice, however, that a p^{th} order Laplacian only considers p-chains, and the structure of the Laplacian depends on the incidence relations between $p-1$, p and $p+1$ simplices. To operate on vertex functions, one needs a vertex Laplacian, which unfortunately only considers the incidence of 0-chains with 1-chains. Thus the vertex Laplacian for a k-uniform hypergraph will not consider any hyperedges, rendering it useless for the purposes of studying vertex functions. Indeed the Laplacian on a 3-uniform graph operates on 2-chains, functions defined on all pairs of vertices (Chung, 1993).

5. Hypergraph Learning Algorithms

A number of existing methods for learning from a hypergraph representation of data first construct a graph representation using the structure of the initial hypergraph. Then, they project the data onto the eigenvectors of the combinatorial or normalized graph Laplacian. Other methods define a hypergraph "Laplacian" using analogies from the graph Laplacian. These methods show that the eigenvectors of their Laplacians are useful for learning, and that there is a relationship between their hypergraph Laplacians and the structure of the hypergraph. In this section, we review these methods. In the next section, we compare these methods analytically.

5.1. Clique Expansion

The clique expansion algorithm constructs a graph $G^x(V, E^x \subseteq V^2)$ from the original hypergraph $G(V, E)$ by replacing each hyperedge $e = (u_1, ..., u_{\delta(e)}) \in E$ with an edge for each pair of vertices in the hyperedge (Zien et al., 1999): $E^x = \{(u, v) : u, v \in e, e \in E\}$.

Note that the vertices in hyperedge e form a clique in the graph G^x. The edge weight $w^x(u, v)$ minimizes the difference between the weight of the graph edge and the weight of each hyperedge e that contains both u and v:

criterion is simply

$$w^x(u, v) = \mu \sum_{e \in E: u, v \in e} w(e) = \mu \sum_e h(u, e)h(v, e)w(e). \tag{12}$$

Here μ is a fixed scalar. The combinatorial or normalized Laplacian of the constructed graph G^x is then used to partition the vertices.

5.2. Star Expansion

The star expansion algorithm constructs a graph $G^*(V^*, E^*)$ from hypergraph $G(V, E)$ by introducing a new vertex for every hyperedge $e \in E$, thus $V^* = V \cup E$ (Zien et al., 1999). It connects the new graph vertex e to each vertex in the hyperedge to it, i.e. $E^* = \{(u, e) : u \in e, e \in E\}$.

Note that each hyperedge in E corresponds to a star in the graph G^* and that G^* is a bi-partite graph. Star expansion assigns the scaled hyperedge weight to each corresponding graph edge:

$$w^*(u, e) = w(e)/\delta(e) \tag{13}$$

The combinatorial or normalized Laplacian of the constructed graph G^x is then used to partition the vertices.

5.3. Bolla's Laplacian

Bolla (Bolla, 1993) defines a Laplacian for an unweighted hypergraph in terms of the diagonal vertex degree matrix D_v, the diagonal edge degree matrix D_e, and the incidence matrix H, defined in Section 2.

$$L^o := D_v - HD_e^{-1}H^\top. \tag{14}$$

The eigenvectors of Bolla's Laplacian L^o define the "best" Euclidean embedding of the hypergraph. Here, the cost for embedding $\phi : V \to \mathbb{R}^k$ of the hypergraph is the total squared distance between pairs of embedded vertices in the same hyperedge

$$\sum_{u,v \in V} \sum_{e \in E: u, v \in e} \|\phi(u) - \phi(v)\|^2 \tag{15}$$

Bolla shows a relationship between the spectral prop-

1. Tézis

Általánosan, él- és csúcs-súlyozott gráfokra definiáltam a Laplace-mátrix normált változatait, melyek természetes módon adódtak a csúcsok optimális reprezentációját definiáló kvadratikus célfüggvény minimalizálásakor, különböző mellékfeltételekkel.

A reprezentációs technika alapját képező lineáris algebrai apparátus egyben lehetőséget nyújtott arra, hogy ezen Laplace-mátrixok spektrumával alsó becsléseket adjunk élsúlyozott gráfok minimális többszempontú vágásaira.

B és Tusnády G: *Spectra and optimal partitions of weighted graphs*, Discrete Math. 128 (1994), 1-20.

B: *Spectral Clustering and Biclustering*, Wiley (2013).

Jelölések

$G = (V, \mathbf{W})$ élsúlyozott gráf, $|V| = n$,

$\mathbf{W}$: $n \times n$-es mátrix (csúcsok páronkénti hasonlóságai)
$w_{ij} = w_{ji} \geq 0$ $(i \neq j)$ és $w_{ii} = 0$ (i=1,. . . ,n).

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

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

(Általában $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$ feltehető .)

Reprezentáció

Keresendők a csúcsok $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^k$ reprezentánsai, melyekkel

$$Q_k = \sum_{i<j} w_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2 \to min.$$

$$\sum_{i=1}^{n} \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k$$

mellett.

$$\mathbf{X}_{n\times k} := (\mathbf{x}_1, \ldots, \mathbf{x}_k) = (\mathbf{r}_1, \ldots, \mathbf{r}_n)^T$$

szubortogonális: $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$. Ezzel

$$Q_k = \mathrm{tr}[\mathbf{X}^T(\mathbf{D} - \mathbf{W})\mathbf{X}] = \mathrm{tr}[\mathbf{X}^T\mathbf{L}\mathbf{X}],$$

ahol **L**: Laplace-mátrix.
$\mathbf{L} \geq 0$ ($Q_1 \geq 0$), és a 0 sajátérték multiplicitása $= G$ összefüggő komponensei száma.
A továbbiakban feltesszük, hogy G összefüggő ($\mathbf{W}$ irreducibilis).

Reprezentációs tétel élsúlyozott gráfokra

Tétel

Legyen $G = (V, \mathbf{W})$ összefüggő élsúlyozott gráf. A fenti jelölésekkel legyenek $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1}$ az $\mathbf{L}$ mátrix sajátértékei az $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ ortonormált sajátvektorokkal, és legyen a $k < n$ pozitív egész olyan, hogy $\lambda_{k-1} < \lambda_k$. Akkor a Q_k célfüggvény minimuma a $\sum_{i=1}^{n} \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k$ kényszerfeltétel mellett

$$\sum_{i=0}^{k-1} \lambda_i = \sum_{i=1}^{k-1} \lambda_i,$$

és a minimum azokkal az ún. optimális $\mathbf{r}_1^*, \ldots, \mathbf{r}_n^*$ reprezentánsokkal éretik el, melyek az $\mathbf{X}^* = (\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1})$ mátrix sorvektorai.

A reprezentációból az $\mathbf{u}_o = \frac{1}{\sqrt{n}}\mathbf{1}$ azonos koordinátákat tartalmazó vektor elhagyható, és a reprezentánsok tetszőlegesen elforgathatók.

Ha a csúcsok is súlyozva vannak

Q_k minimumát

$$\sum_{i=1}^{n} s_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k$$

mellett keressük, ahol $s_1, \ldots, s_n > 0$ a csúcsok súlyai, $\mathbf{S} := \mathrm{diag}(s_1, \ldots, s_n)$. A megoldást az $\mathbf{L}_S = \mathbf{S}^{-1/2}\mathbf{L}\mathbf{S}^{-1/2}$ súlyozott Laplace-matrix spektrálfelbontása adja. Ha $\mathbf{S} = \mathbf{D}$:

Definíció

Az

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

mátrixot a $G = (V, \mathbf{W})$ élsúlyozott gráf normált Laplace-mátrixának nevezzük.

$\mathbf{L}_D$ sajátértékei a $[0, 2]$ intervallumban vannak (érzéketlenek az élsúlyok skálázására), a 0 sajátérték multiplicitása $=$ G összefüggő komponenseiszámával, és 2 pontosan akkor sajátérték, ha G páros.

Reprezentációs tétel él- és speciális csúcs-súlyozott gráfokra

Tétel

Legyen $G = (V, \mathbf{W})$ *összefüggő élsúlyozott gráf.* $\mathbf{L}_D$ *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 a* Q_{k-1} *célfüggvény minimuma*

$$\sum_{i=1}^{n} d_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_{k-1} \quad \textit{és} \quad \sum_{i=1}^{n} d_i \mathbf{r}_i = \mathbf{0}$$

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.*

Mivel $\mathbf{u}_0 = \sqrt{\mathbf{d}}$, a $\mathbf{D}^{-1/2}\mathbf{u}_0 = \mathbf{1}$ vektor automatikusan kimarad a reprezentációból a bővített kényszerfeltételek miatt.

Többszempontú arányos vágások

Definíció

A $G = (V, \mathbf{W})$ élsúlyozott gráf $U, T \subset V$ csúcshalmazai közti súlyozott vágás $w(U, T) = \sum_{i \in U} \sum_{j \in T} w_{ij}$.

Definíció

Legyen $G = (V, \mathbf{W})$ élsúlyozott gráf és $P_k = (V_1, \ldots, V_k)$ a csúcsok k-partíciója. G arányos k-vágása a P_k partíció tekintetében

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

minimális arányos k-vágása pedig $g_k(G) = \min_{P_k \in \mathcal{P}_k} g(P_k, G)$, ahol $\mathcal{P}_k$ az összes k-partíció halmaza.

Többszempontú normált vágások

Definíció

Az $U \subset V$ csúcshalmaz térfogata $\mathrm{Vol}(U) = \sum_{i \in U} d_i$.

Definíció

Legyen $G = (V, \mathbf{W})$ élsúlyozott gráf a $d_1, \ldots, d_n$ általánosított fokszámokkal. Az általánosság megszorítása nélkül feltesszük, hogy $\sum_{i=1}^{n} d_i = 1$. A G élsúlyozott gráf normált k-vágása a $P_k = (V_1, \ldots, V_k)$ partíció tekintetében

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

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

Megkeresésük NP-nehéz.

Sajátértékek kapcsolat a többszempontú vágásokkal

Tétel

Legyenek $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{n-1}$ *a* $G = (V, \mathbf{W})$ *élsúlyozott gráf Laplace-mátrixának sajátértékei. Akkor*

$$g_k(G) \geq \sum_{i=1}^{k-1} \lambda_i.$$

Tétel

Legyenek $0 = \lambda'_0 < \lambda'_1 \leq \cdots \leq \lambda'_{n-1} \leq 2$ *a* $G = (V, \mathbf{W})$ *összefüggő élsúlyozott gráf normált Laplace-mátrixának sajátértékei. Akkor*

$$f_k(G) \geq \sum_{i=1}^{k-1} \lambda'_i.$$

k-variancia

Biz: partícióvektorokkal.
Ezek közelségét a sajátvektorokhoz a k-variancia fejezi ki, amit a reprezentánsokra alkalmazott k-közép algoritmus minimalizál: spektrális relaxáció.

Definíció

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

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

ahol $P_k = (V_1, \ldots, V_k)$ *az* $\{1, \ldots, n\}$ *indexhalmaz* k*-partíciója és* $\mathbf{c}_a = \frac{1}{|V_a|} \sum_{j \in V_a} \mathbf{r}_j$ *az* a*-adik pontklaszter súlypontja* $(a = 1, \ldots, k)$.

Súlyozott k-variancia és rés a spektrumban

Definíció

Legyenek az $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^{k-1}$ *pontok a* $d_1, \ldots, d_n$ *pozitív súlyokkal ellátva, ahol* $\sum_{i=1}^{n} d_i = 1$. *A pontrendszer súlyozott* k*-varianciája*

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

ahol $\mathbf{c}_a = \frac{1}{\sum_{j \in V_a} d_j} \sum_{j \in V_a} d_j \mathbf{r}_j$ *az a-adik pontklaszter súlypontja.*

Tétel

$G = (V, \mathbf{W})$ *élsúlyozott gráf,* $\sum_{i=1}^{n} d_i = 1$. *Legyenek* $0 = \lambda'_0 < \lambda'_1 \leq \cdots \leq \lambda'_{n-1}$ $\mathbf{L}_D$ *sajátértékei az* $\mathbf{u}'_0, \mathbf{u}'_1, \ldots, \mathbf{u}'_{n-1}$ *ortonormált sajátvektorokkal. Akkor az optimális* $r_1^*, \ldots, r_n^* \in \mathbb{R}$ *csúcs-reprezentánsokkal, melyek a* $\mathbf{D}^{-1/2}\mathbf{u}'_1$ *vektor koordinátái:* $\tilde{S}^2(r_1^*, \ldots, r_n^*) = \lambda'_1 / \lambda'_0$

2. Tézis

Felső becslést adtam egy élsúlyozott gráf minimális normált k-vágására $\mathbf{L}_D$ k legkisebb sajátértéke segítségével a $(k-1)$-dimenziós reprezentánsok 'jó' klaszteresedése esetén. A $k = 2$ esetben az előbbi Tétel mutatta, hogy ehhez elégséges a λ_1' és λ_2' közti rés. A $k > 2$ bonyolultabb esetről a 8. Tézisben szólok. Az izoperimetrikus számra javított felső becslést adtunk λ_1' segítségével élsúlyozott gráfokra. Ennek következményeivel az ún. szimmetrikus maximálkorrelációra vonatkozóan az 5. Tézisben foglalkozom.

B és Molnár-Sáska G: *Isoperimetric properties of weighted graphs related to the Laplacian spectrum and canonical correlations*, Studia Sci. Math. Hung. 39 (2002), 425–441.

Felső becslés f_k-ra

Tétel

Legyenek $0 = \lambda'_0 < \lambda'_1 \leq \cdots \leq \lambda'_{n-1} \leq 2$ *a* $G = (V, \mathbf{W})$ *összefüggő élsúlyozott gráf normált Laplace-mátrixának sajátértékei az* $\mathbf{u}'_0, \mathbf{u}'_1, \ldots, \mathbf{u}'_{n-1}$ *ortonormált sajátvektorokkal. Tegyük fel, hogy a csúcsok optimális* $(k-1)$*-dimenziós reprezentánsai, melyek az* $\mathbf{X}^* = \mathbf{D}^{-1/2}(\mathbf{u}_1, \ldots, \mathbf{u}_{k-1})$ *mátrix sorai, a súlyozott k-közép algoritmussal 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* $c' = 1/\min_i \sqrt{\mathrm{Vol}(V_i)}$.

Az izoperimetrikus szám

Definíció

A $G = (V, \mathbf{W})$ élsúlyozott gráf ($\sum_{i=1}^{n} d_i = 1$) izoperimetrikus száma (Cheeger-konstans)

$$h(G) = \min_{\substack{U \subset V \\ \mathrm{Vol}(U) \leq \frac{1}{2}}} \frac{w(U, \overline{U})}{\mathrm{Vol}(U)}. \qquad (f_2(G) \leq 2h(G))$$

Tétel

Ha $\mathbf{L}_D$ legkisebb pozitív sajátértéke $\lambda_1' \leq 1$, akkor

$$\frac{\lambda_1'}{2} \leq h(G) \leq \sqrt{\lambda_1'(2 - \lambda_1')}.$$

B, Bullins B, Chaturapruek S, Chen S, Friedl K: *When the largest eigenvalue of the modularity and the normalized modularity matrix is zero* (2013), ArXiv:1305.2147.

3. Tézis

Még általánosabban, bevezettem kontingenciatáblák (nemnegatív elemű téglalapmátrixok, mint pl. microarray-k) sorainak és oszlopainak optimális alacsony-dimenziós reprezentációját és vizsgáltam annak kapcsolatát a korrespondencia-mátrix szinguláris felbontásával. Ezután kontingenciatáblák normált kétszempontú vágásait becsültem e mátrix szinguláris értékeivel.

A kiterjesztás alapján követetéseket vontam le a szimmetrikus esetre: amennyiben valamely k-ra egy élsúlyozott gráf normált modularitás-mátrixának $k-1$ nagy abszolút értékű sajátértéke mind pozitív ill. negatív, akkor k olyan csúcsklasztert találhatunk, melyeknél a klaszterek közti, ill. a klasztereken belüli élsűrűség kicsi (a fizikusok szóhasználatával élve, 'community', ill. 'anti-community structure'). Ezek a struktúrák speciális esetei a 8. Tézisben vizsgált kis diszkrepanciájú, ún. reguláris vágásoknak.

B: *Spectral Clustering and Biclustering*, Wiley (2013).

Normált kontingenciatábla, korrespondanciaanalízis

$\mathbf{C}_{m\times n}$: kontingenciatábla $(c_{ij} \geq 0)$
$d_{row,i} = \sum_{j=1}^{n} c_{ij}$, $\mathbf{D}_{row} = \mathrm{diag}(d_{row,1}, \ldots, d_{row,m})$
$d_{col,j} = \sum_{i=1}^{m} c_{ij}$, $\mathbf{D}_{col} = \mathrm{diag}(d_{col,1}, \ldots, d_{col,n})$

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

Adott $1 \leq k \leq \mathrm{rang}(\mathbf{C})$ egészhez keresendők a sorok $\mathbf{r}_1, \ldots, \mathbf{r}_m \in \mathbb{R}^k$ és az oszlopok $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^k$ reprezentánsai, melyekre

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

$$\sum_{i=1}^{m} d_{row,i}\mathbf{r}_i\mathbf{r}_i^T = \mathbf{I}_k \quad \text{és} \quad \sum_{j=1}^{n} d_{col,j}\mathbf{q}_j\mathbf{q}_j^T = \mathbf{I}_k$$

mellett.

Reprezentációs tétel kontingenciatáblákra

Tétel

Tfh. $\mathbf{C}\mathbf{C}^T$ *irreducibilis.* Q_k *minimuma a kényszerfeltételek mellett* $2k - \sum_{i=0}^{k-1} s_i$, *ahol* $1 = s_0 > s_1 \geq \cdots \geq s_{r-1} > 0$ $\mathbf{C}_{corr}$ *szinguláris értékei és a minimumot az első* k *korrespondancia-vektorpár adja.*

'Biclusterek':

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

$$\nu_k(\mathbf{C}) = \min_{P_{row}, P_{col}, \sigma} \nu_k(P_{row}, P_{col}, \sigma) \geq 2k - \sum_{i=0}^{k-1} s_i$$

B: *SVD, discrepancy, and regular structure of contingency tables*, Discrete Applied Math. (megjelenés előtt).

4. Tézis

Definiáltam a fizikusok által az ezredforduló után bevezetett Newman–Girvan modularitás normált változatait, és vizsgáltam azok kapcsolatát a minimális többszempontú vágásokkal és a modularitás-mátrix sajátértékeivel.

Bevezettem az ún. normált modularitás-mátrixot, mely a normált Laplace-mátrix alkalmas transzformációjával kapható, és sajátértékei az 5. Tézisben bevezetendő feltételes várható érték képzés operátora sajátértékeinek feleltethetők meg. Ennek a mátrixnak a spektruma $[-1, 1)$-beli és nagy abszolút értékű (ún. strukturális) sajátértékei a hozzájuk tartozó sajátvektorokkal együtt fontos szerepet kapnak a 8. Tézisbeli diszkrepanciák becslésében.

B: *Penalized versions of the Newman–Girvan modularity and their relation to normalized cuts and k-means clustering*, Physical Review E 84, 016108 (2011).

Newman–Girvan modularitás

Legyen $G = (V, \mathbf{W})$ élsúlyozott gráf, $\sum_{i=1}^{n} d_i = 1$.

$$M_k(G) = \max_{P_k \in \mathcal{P}_k} \sum_{a=1}^{k} \sum_{i,j \in V_a} (w_{ij} - d_i d_j)$$

$\mathbf{M} = \mathbf{W} - \mathbf{d}\mathbf{d}^T$: modularitás-mátrix

max–min kapcsolata a pozitív–negatív sajátértékekkel

0: vízválasztó

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

Ha $0 = \lambda'_0 < \lambda'_1 \leq \cdots \leq \lambda'_{n-1} \leq 2$ az $\mathbf{L}_D$ mátrix sajátértékei 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 tehát $[-1, 1)$-beli; 1 nem lehet sajátérték, amennyiben G összefüggő; $\mathbf{M}, \mathbf{M}_D \leq 0$ karakterizálása.

5. Tézis

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 korrespondanciaanalízis 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).

B: *Spectral Clustering and Biclustering*, Wiley (2013).

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:

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

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

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

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

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

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

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

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

(ψ_i és ψ_i' azonos eloszlásúak $\mathbb{W}$ együttes eloszlással).

Maximálkorrelácó (Gebelein és Rényi)

Keresendő 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. 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* $(\phi_1, \ldots, \phi_k)$ *optimális reprezentánsokkal érhető el.*

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 $(\psi_1, \ldots, \psi_k)$ 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):

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

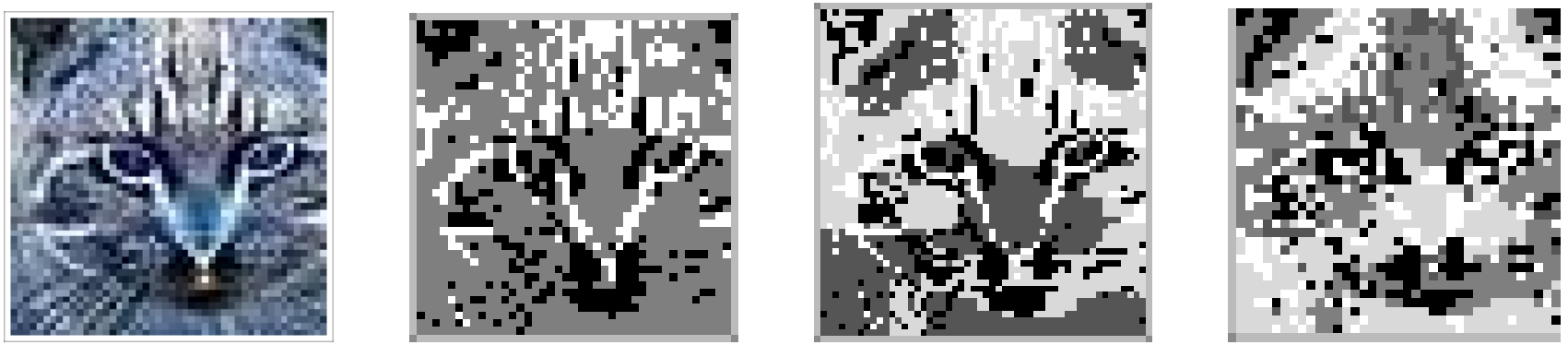
A Cheeger-egyenlőtlenség miatt

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

(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 × 48 pixel)

M_D strukturális sajátértékei:

0.137259, 0.014255, 0.000925,

−0.0006707, −0.0006706, . . .

Gauss-mag

image segmentation

6. Tézis

Nagyméretű véletlen hálózatok sajátértékeinek és spektrális klasztereinek aszimptotikus viselkedését vizsgáltam (a csúcsok száma növekszik, miközben a köztük levő kapcsolatok is perturbálódnak). Rögzített klaszterszám (k) esetén általános Wigner-zajjal terhelt blokkmátrixok sajátértékeinek és sajátaltereinek aszimptotikus tulajdonságait karakterizáltam, miközben a csúcsok száma (n) és ezzel együtt a klaszterméretek is tartottak a végtelenbe. Beláttam, hogy a zajos mátrixnak majdnem biztosan van k strukturális ($\Theta(n)$) sajátértéke, és a hozzájuk tartozó sajátvektorokkal reprezentálva, a reprezentánsok k-varianciája majdnem biztosan $\mathcal{O}(\frac{1}{n})$. Az általánosított véletlen gráfok szomszédsági mátrixa egy speciális zajos mátrixnak felel meg, így az ilyen gráfok spektrális karakterizációját is megadtam.

B: *Recognizing linear structure in noisy matrices*, Lin. Alg. Appl. 402 (2005), 228-244.

B: *Noisy random graphs and their Laplacians*, Discrete Math. 308 (2008), 4221-4230.

Szimmetrikus Wigner-zaj és felfújt mátrix

Definíció

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

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

Definíció

Az $n \times n$*-es* $\mathbf{B}$ *mátrix szimmetrikus felfújt mátrix, ha van olyan* $k < n$ *pozitív egész és* $\mathbf{P}$ $k \times k$*-as, szimmetrikus ún. mintázat-mátrix* $0 < p_{ij} < 1$ *elemekkel, továbá* $n_1, \ldots, n_k$ *pozitív egészek* $(\sum_{i=1}^{k} n_i = n)$*, hogy a* $\mathbf{B}$ *mátrix sorait és oszlopait ugyanúgy permutálva,* $\mathbf{B}$ *blokkmátrix* p_{ij} *elemekkel* $(n_i \times n_j$*-n).*

Szimmetrikus zajos mátrix

k rögzített, $\mathbf{P}$-t egyre nagyobb $n \times n$-es $\mathbf{B}_n$ blokkmátrixszá fújjuk fel, és vizsgáljuk az $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$ zajos mátrixsorozatot, amint $n_1, \ldots, n_k \to \infty$ úgy, hogy

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

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

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

akkor $\mathbf{A}_n$ elemei [0,1]-beliek, és $G_n = (V, \mathbf{A}_n)$ növekvő véletlen gráfsorozat.
Alkalmas Wigner-zajjal el tudom érni, hogy G_n általánosított véletlen gráf: csúcsainak létezik olyan $(V_1, \ldots, V_k)$ partíciója, hogy $\mathbb{P}(i \sim j) = p_{ab}$ egymástól függetlenül, ha $i \in V_a$, $j \in V_b$ $(a, b = 1, \ldots, k)$.

Sajátértékek és sajátalterek perturbációja

$\mathbf{B}_n$ rangja legfeljebb k; tfh. $\operatorname{rang}(\mathbf{B}) = \operatorname{rang}(\mathbf{P}) = k$.
$\mathbf{B}_n$ nullától különböző sajátértékei: $|\beta_1|, \ldots, |\beta_k| = \Theta(n)$.

Tétel

Az $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$ zajos mátrixnak van $\lambda_1, \ldots, \lambda_k$ sajátértéke, melyekre

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

a maradék $n - k$ sajátértékre pedig

$$|\lambda_j| \leq 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3} \log n) \quad j = k+1, \ldots, n$$

teljesül majdnem biztosan, ha $n \to \infty$ a blokkméretek végtelenbe tartására tett feltétel mellett.

(Alon–Krivelevich–Vu tétel + Borel–Cantelli lemma: majdnem biztos eredmények.)

Következmények

- $\Delta - 2\varepsilon$ nagyságrendű spektrális rés $\mathbf{A}_n$ strukturális $(\lambda_1, \ldots, \lambda_k)$ és többi sajátértéke közt, ahol

$$\varepsilon = \|\mathbf{W}_n\| = 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n) \quad \text{és} \quad \Delta = \min_{1 \le i \le k} |\beta_i| = \Theta(n).$$

- $$S_k^2(\mathbf{r}_1^*, \ldots, \mathbf{r}_n^*) \le k\frac{\varepsilon^2}{(\Delta - \varepsilon)^2} = \mathcal{O}(\frac{1}{n})$$

  majdnem biztosan, ha $n \to \infty$ a fenti feltételek mellett, ahol $\mathbf{r}_1^*, \ldots, \mathbf{r}_n^* \in \mathbb{R}^k$ az $\mathbf{A}_n$ mátrix strukturális sajátértékeihez tartozó sajátvektorokkal gyártott reprezentánsok.
- Gyenge szálak (Granovetter), randomizált SVD.

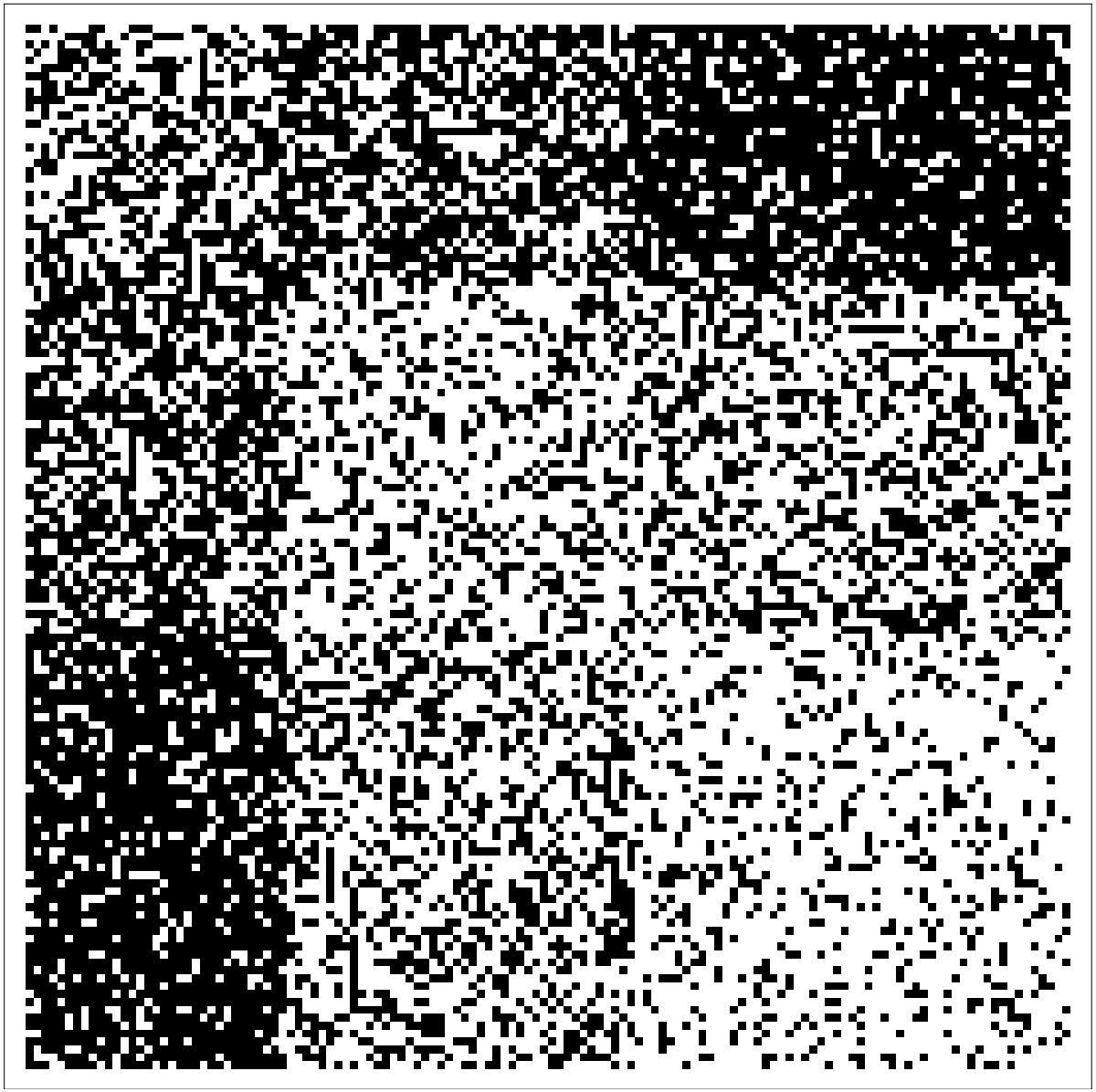
Perturbácós tétel a normált Laplace-mátrixra

Tétel

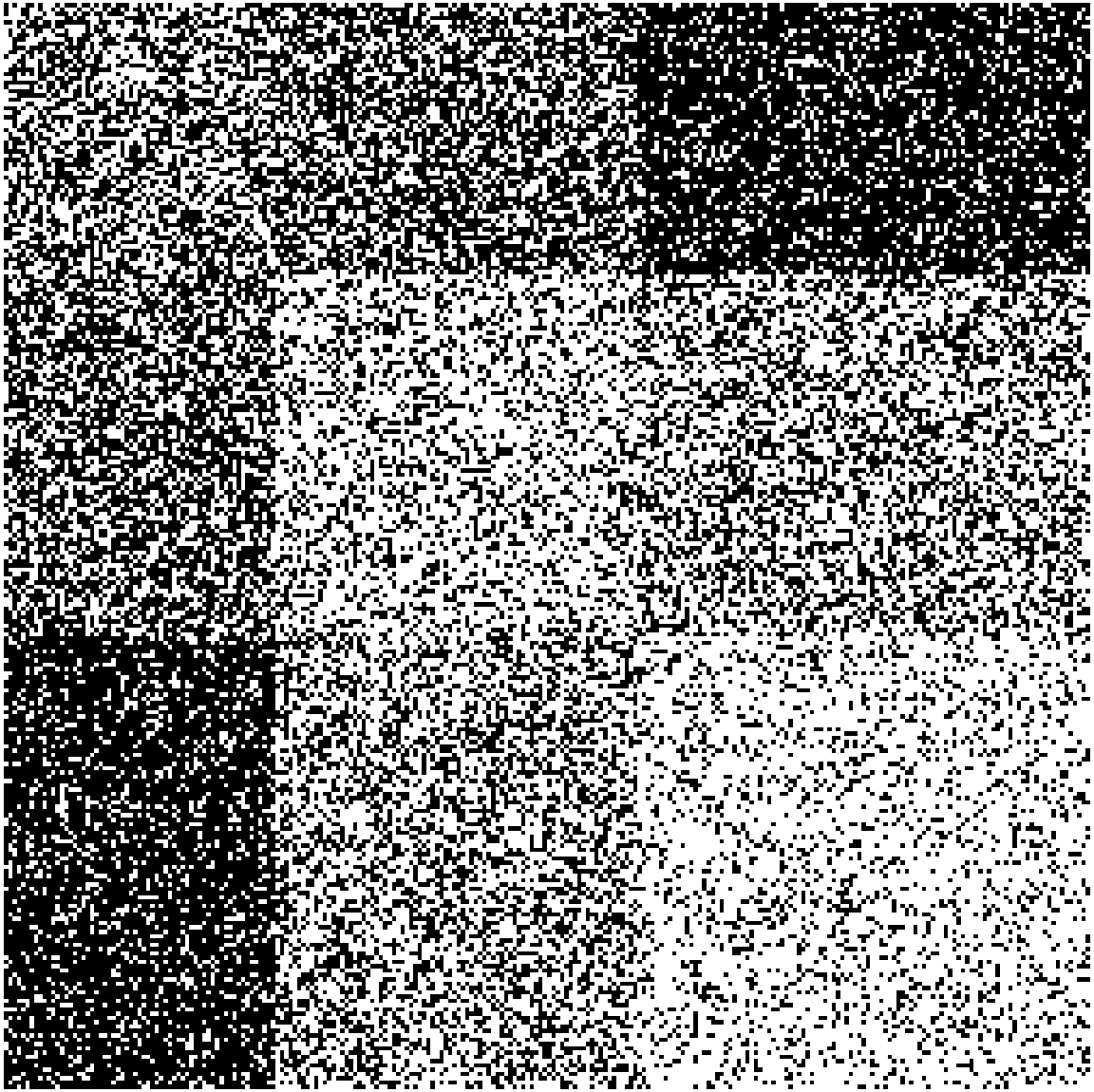
Legyen $G_n = (V, \mathbf{A}_n)$ *véletlen él-súlyozott gráf,* $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$, *ahol a* $\mathbf{B}_n$ *mátrix a* k*-rangú* $\mathbf{P}$ *mátrix felfújtja,* $\mathbf{W}_n$ *pedig szimmetrikus Wigner-zaj (az elemek* K *egyenletes korlátjára tett feltételek mellett) . Akkor (*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 blokméretek végtelenbe tartására tett feltételek mellett.*

Ha $\epsilon = n^{-\tau}$, akkor a normált Laplace-mátrix 1-től elszeparált sajátértékeihez tartozó sajátvektorokkal reprezentálva a csúcsokat, azok súlyozott k-varianciája legfeljebb $\frac{k}{(\frac{\delta}{\epsilon} - 1)^2}$ majdnem biztosan, a szokásos feltételek mellett.

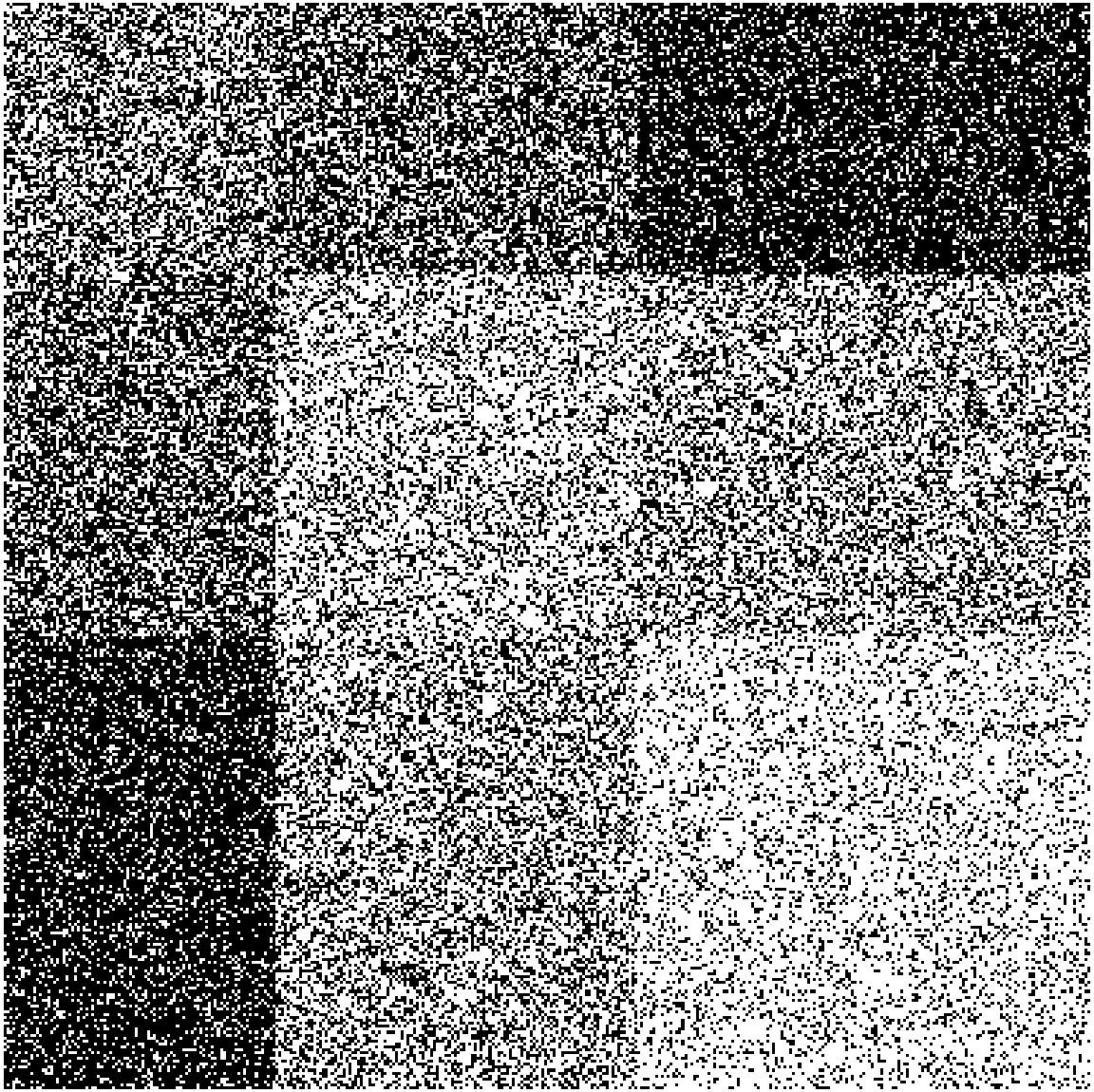
10-szeres felfújás + zaj



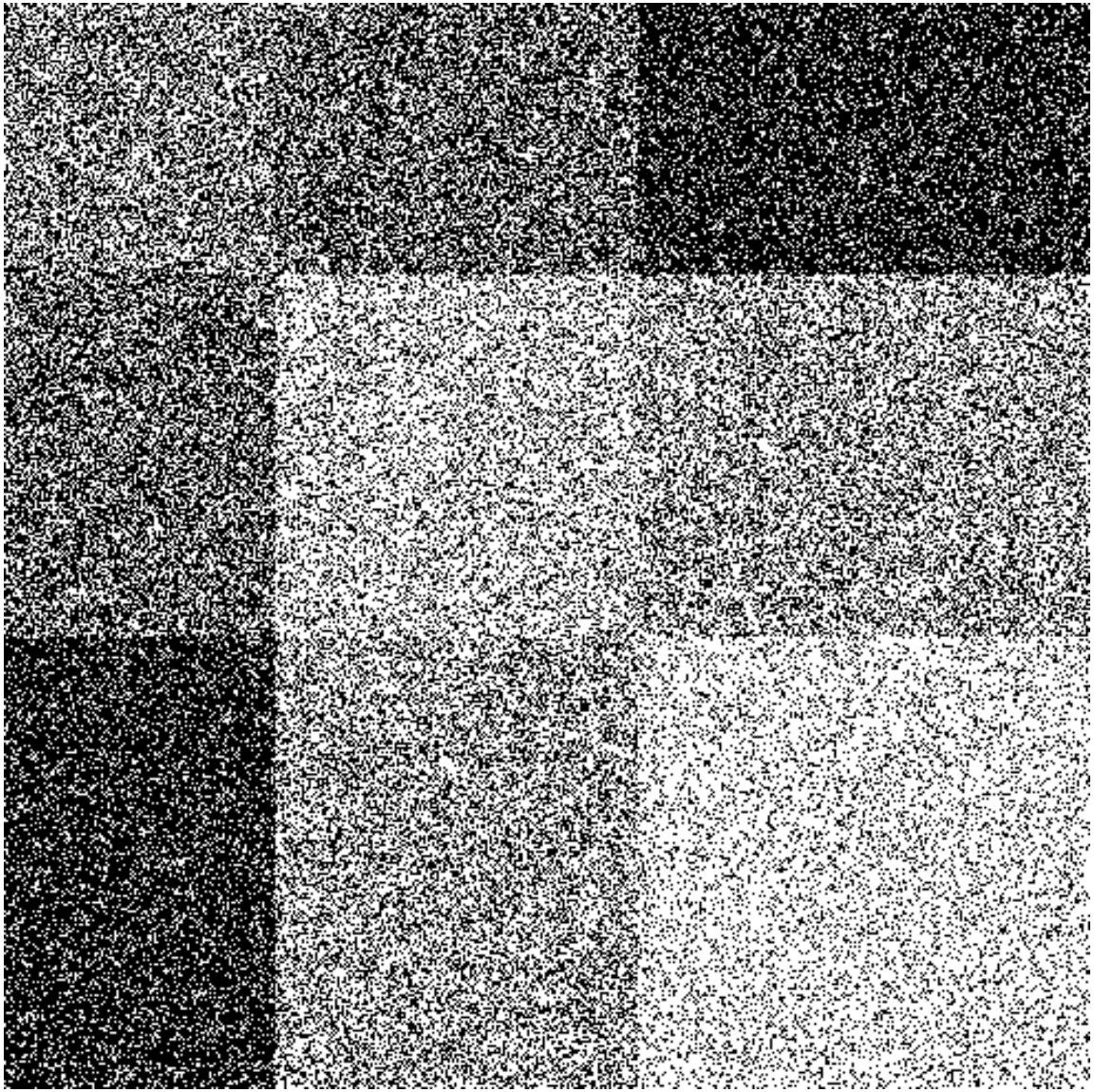
20-szoros felfújás + zaj



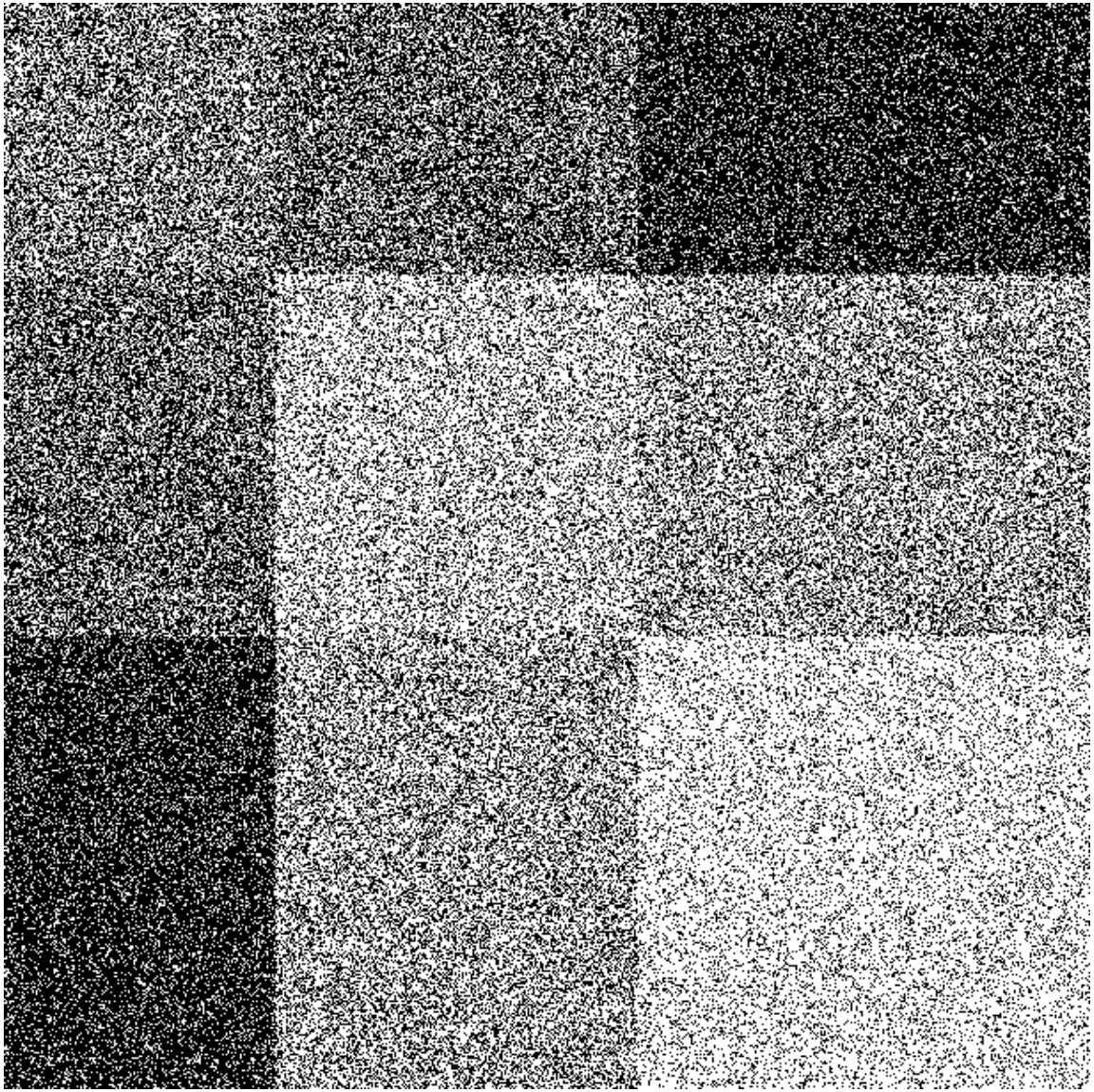
30-szoros felfújás + zaj



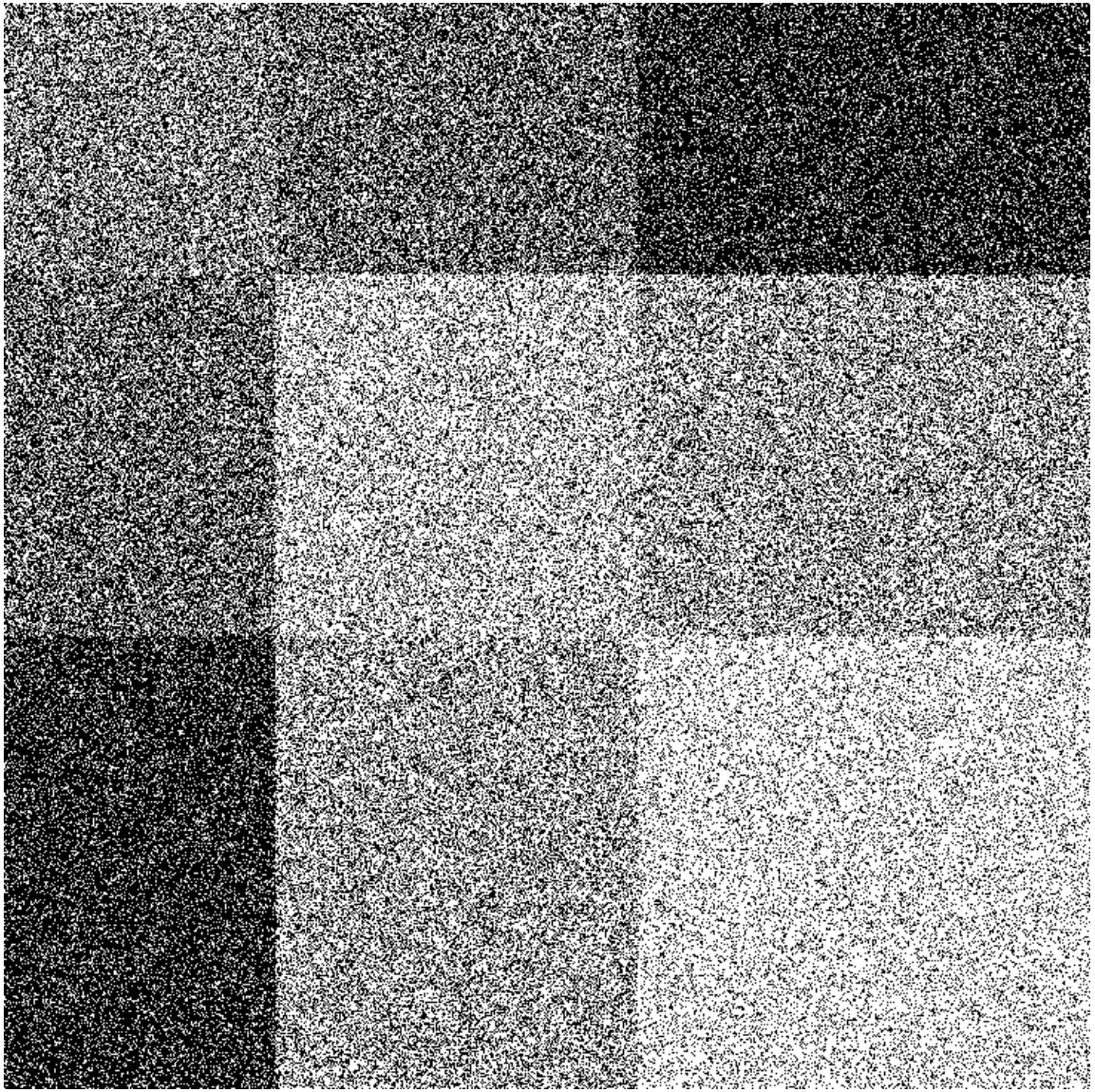
40-szeres felfújás + zaj



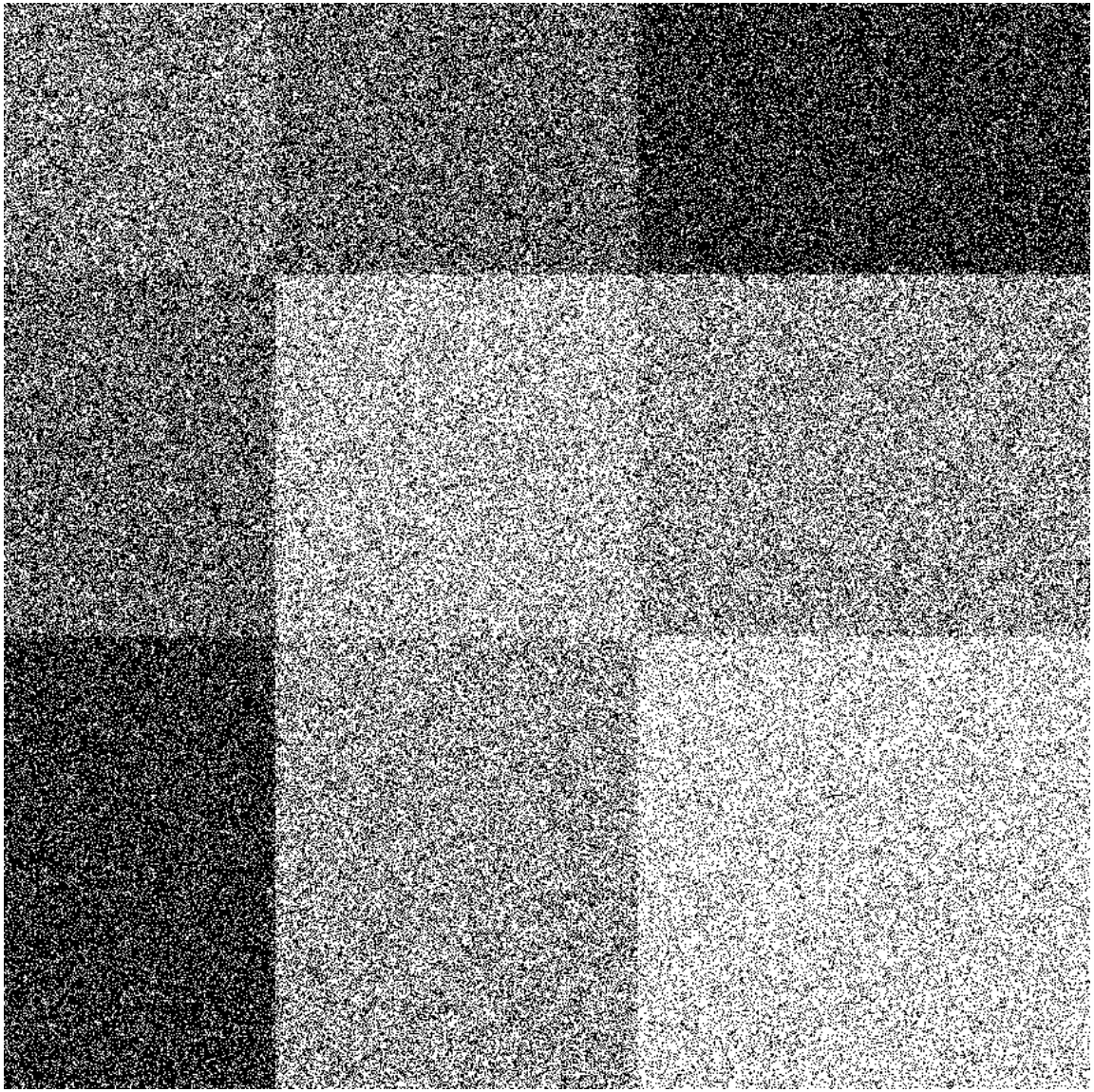
50-szeres felfújás + zaj



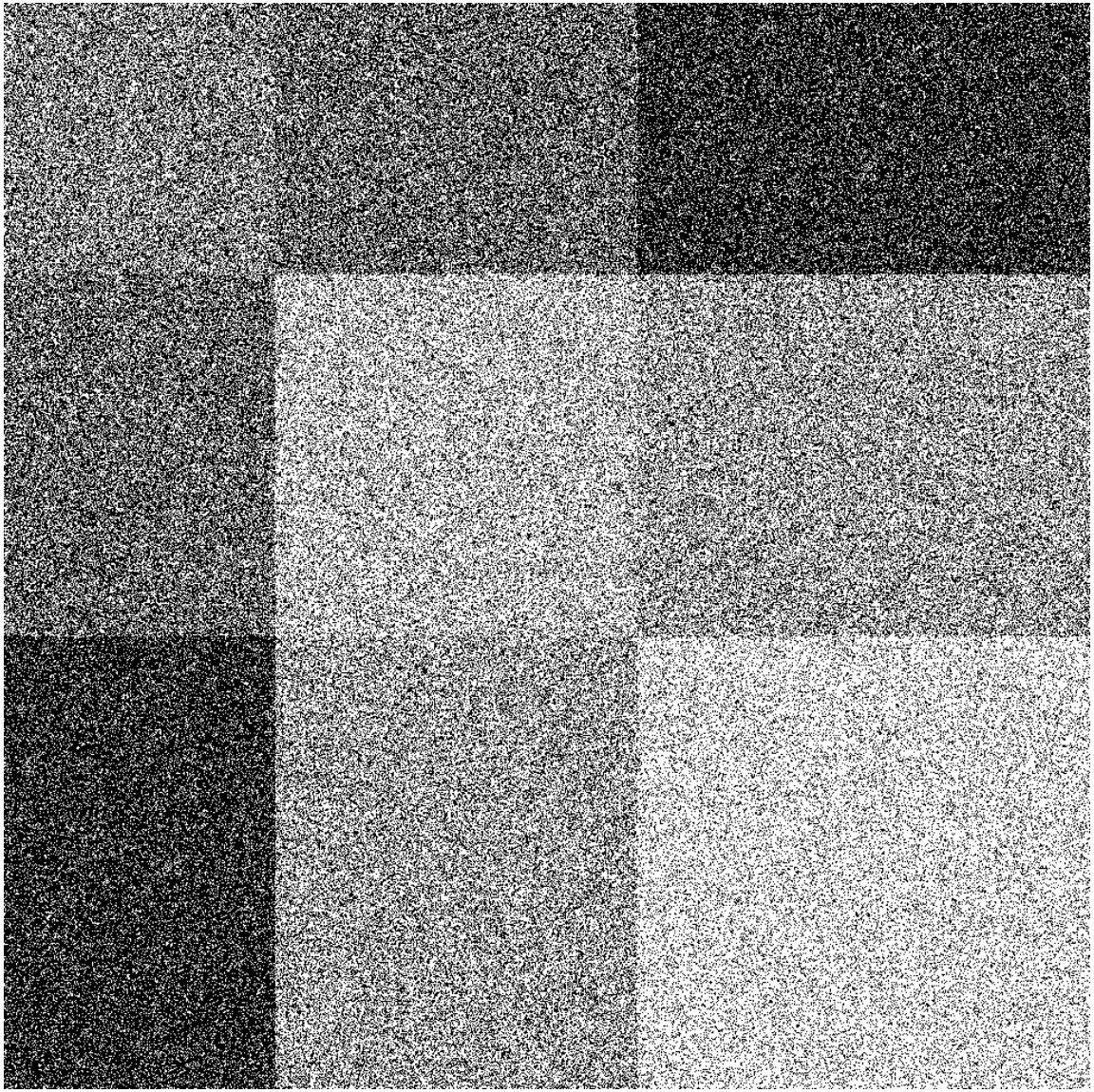
60-szoros felfújás + zaj



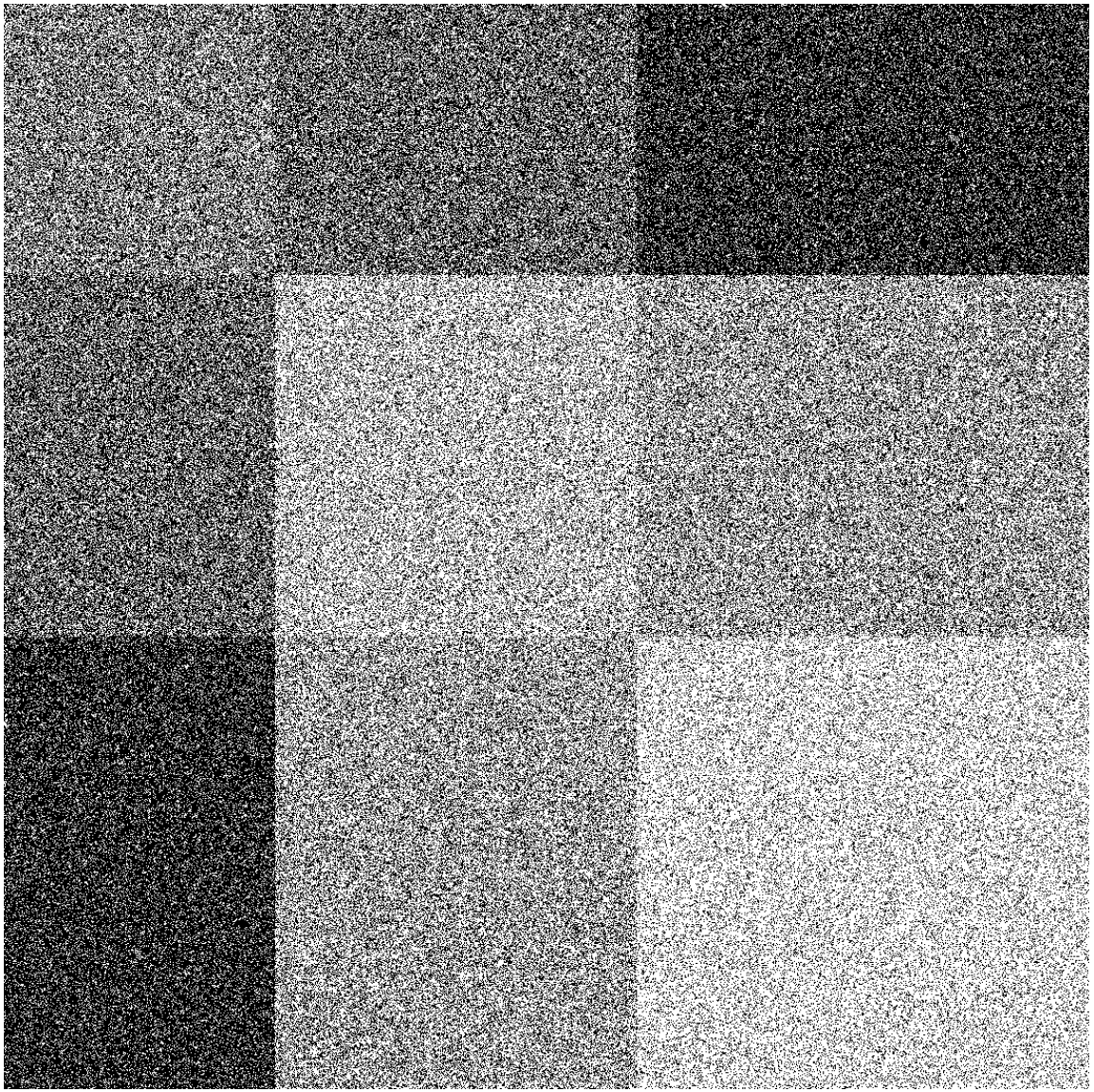
70-szeres felfújás + zaj



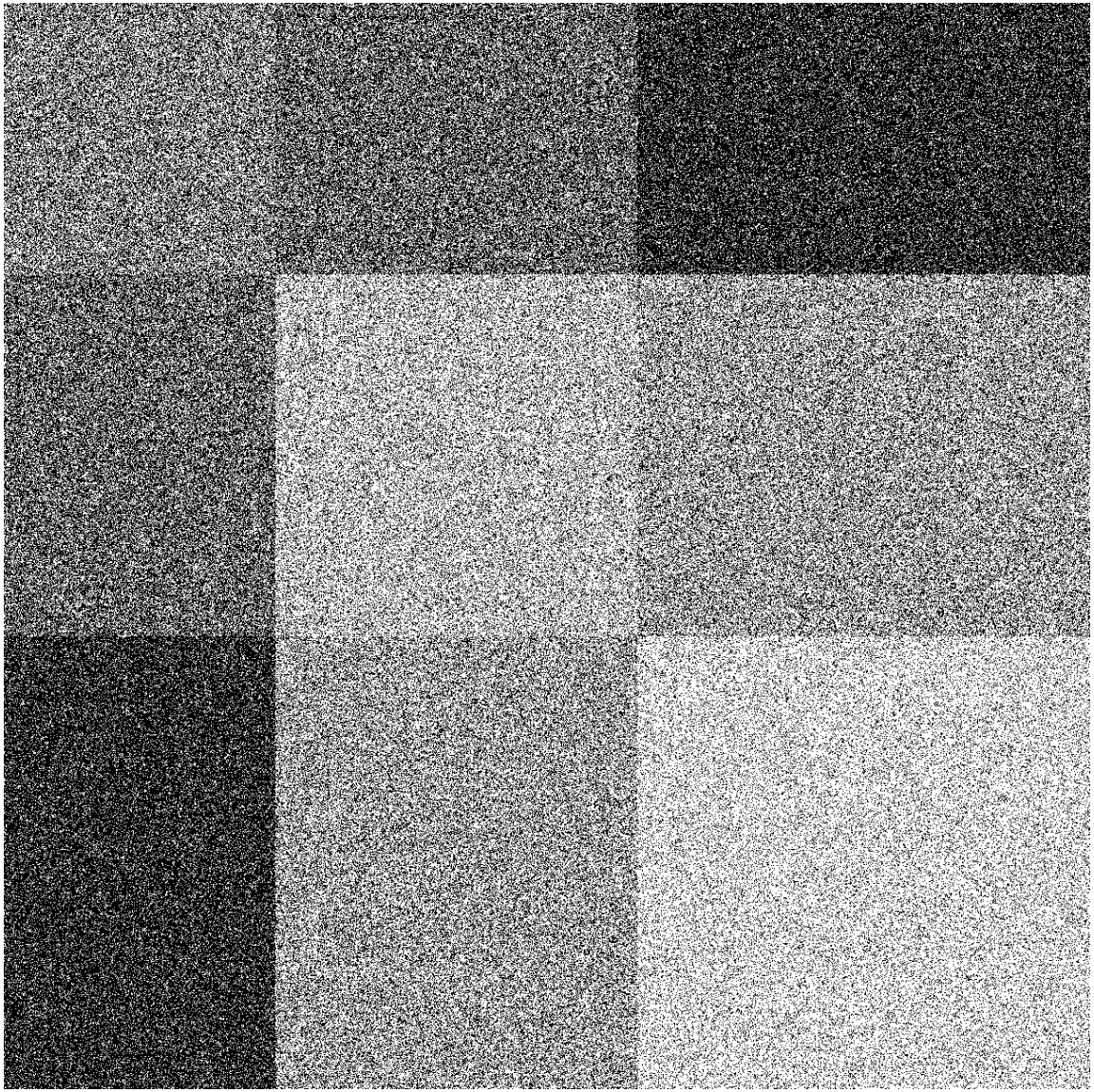
80-szoros felfújás + zaj



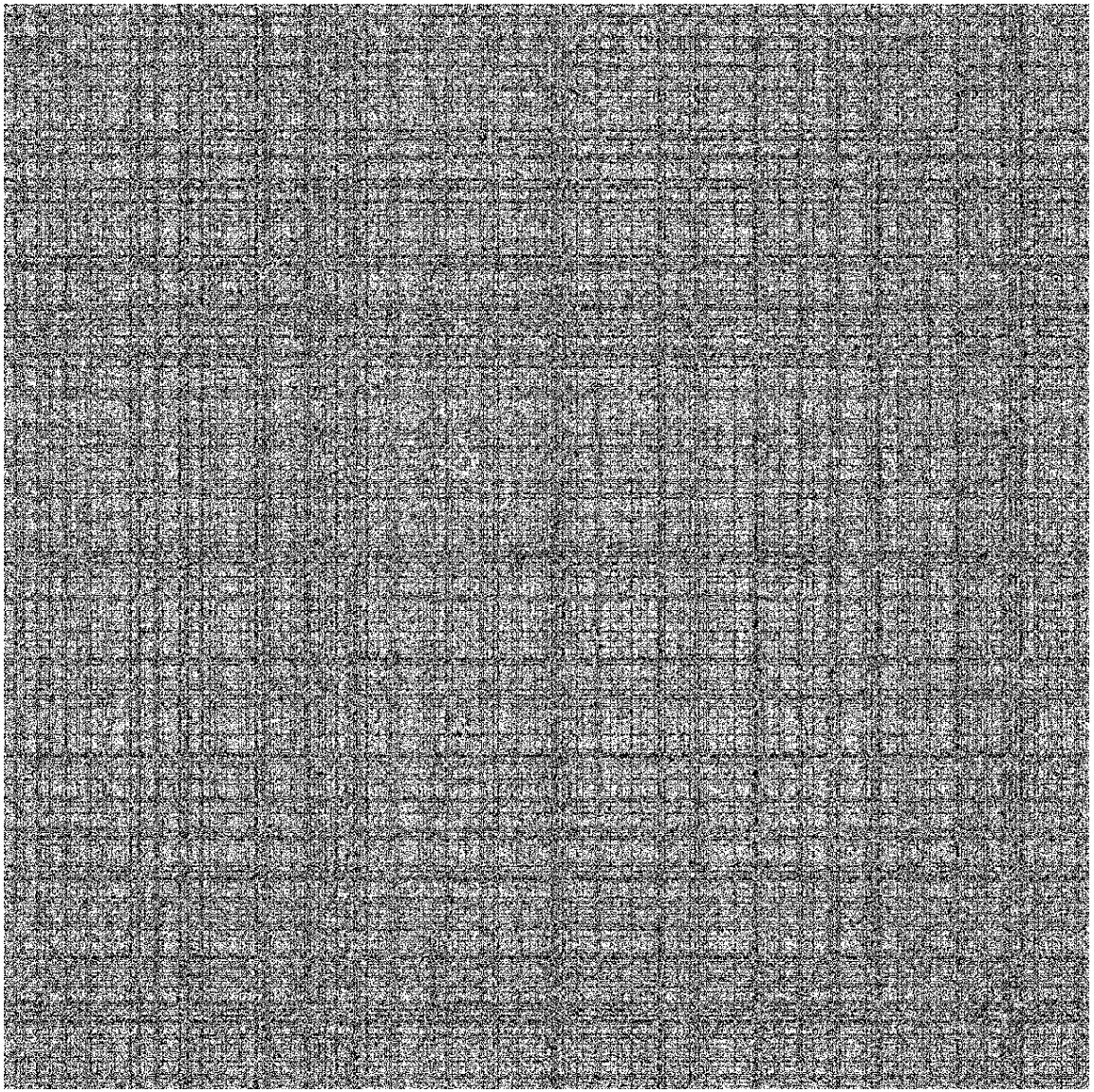
90-szeres felfújás + zaj



100-szoros felfújás + zaj



A csúcsok klaszterezése és permutálása előtti állapot



7. Tézis

A 6. Tézis eredményeit kiterjesztettük kontingenciatáblák perturbációira. Rögzített a, b pozitív egészek esetén egy $m \times n$-esre felfújt, $a \times b$ blokkból álló, téglalap alakú Wigner-zajjal terhelt mátrixnak majdnem biztosan van $k = \min\{a, b\}$ strukturális ($\Theta(\sqrt{mn})$) szinguláris értéke, míg a többi $\mathcal{O}(\sqrt{m+n})$ rendű. A strukturális szinguláris értékekhez tarozó szinguláris vektorpárok alapján reprezentálva a sorokat és oszlopokat, a sor- és oszlop-reprezentánsok a- és b-varianciája majdnem biztosan 0-hoz tart, amennyiben a blokkméretek végtelenbe tartanak bizonyos arányossági feltételek mellett. Az eredményeket kiterjesztettük zajos kontingenciatáblákból számolt korrespondancia-mátrixokra is. Rögzített a, b esetén, amennyiben még azt is feltesszük, hogy m, n közel azonos nagyságrendben tart a végtelenbe, a zajos korrespondancia-mátrixnak van pontosan $k = \min\{a, b\}$ darab 0-tól δ-val elszeparált szinguláris értéke (δ nem függ m, n-től).

B, Friedl K, Krámli A: *Singular value decomposition of large random matrices*, J. Multivariate Anal. 101 (2010), 434-446.

Zajos téglalap-mátrix

a, b rögzített, a $\mathbf{P}_{a\times b}$ mintázat-mátrixot egyre nagyobb $\mathbf{B}_{m\times n}$ blokkmátrixszá fújjuk fel ($m_1, \ldots, m_a$ és $n_1, \ldots, n_b$ blokk-méretekkel).

$\mathbf{A}_{m\times n} = \mathbf{B}_{m\times n} + \mathbf{W}_{m\times n}$ zajos mátrixsorozat, ahol $\mathbf{W}_{m\times n}$ Wigner-zaj (elemei független, egyenletesen korlátos, 0 várható értékű val. változók).

$m = \sum_{i=1}^{a} m_i \to \infty$, $n = \sum_{i=1}^{b} n_i \to \infty$ úgy, hogy

F1 Van olyan $0 < c \leq \frac{1}{a}$ konstans, hogy $\frac{m_i}{m} \geq c$ $(i = 1, \ldots, a)$ és olyan $0 < d \leq \frac{1}{b}$ konstans, hogy $\frac{n_i}{n} \geq d$ $(i = 1, \ldots, b)$.

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

Szinguláris értékek és sajátalterek perturbációja

$\mathbf{B}_n$ rangja legfeljebb $k = \min\{a, b\}$; tfh. $\operatorname{rang}(\mathbf{B}) = \operatorname{rang}(\mathbf{P}) = k$.
$\mathbf{B}_n$ nullától különböző szinguláris értékei:
$s_1, \ldots, s_k = \Theta(\sqrt{mn})$.

Tétel

Az $\mathbf{A}_{m\times n} = \mathbf{B}_{m\times n} + \mathbf{W}_{m\times n}$ mátrixnak vannak $z_1, \ldots, z_k$ szinguláris értékei, melyekre

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

többi szinguláris értékére pedig

$$z_j = \mathcal{O}(\sqrt{m+n}), \quad j = k+1, \ldots, \min\{m, n\}$$

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

(Alon–Krivelevich–Vu tétel téglalapra+Borel–Cantelli lemma: m.b.)

Következmények

- $\Delta - 2\varepsilon$ nagyságrendű spektrális rés $\mathbf{A}_{m\times n}$ strukturális $(z_1, \ldots, z_k)$ és többi szinguláris értéke közt, ahol

$$\varepsilon := \|\mathbf{W}_{m\times n}\| = \mathcal{O}(\sqrt{m+n}) \quad \text{és} \quad \Delta := \min_{1\le i\le r} s_i = \Theta(\sqrt{mn}).$$

- $$S_a^2(\mathbf{r}_1^*, \ldots, \mathbf{r}_m^*) = \mathcal{O}\left(\frac{m+n}{mn}\right), \quad S_b^2(\mathbf{q}_1^*, \ldots, \mathbf{q}_n^*) = \mathcal{O}\left(\frac{m+n}{mn}\right)$$

  majdnem biztosan, ha $m, n \to \infty$ az **F1** feltétel mellett, ahol $\mathbf{r}_1^*, \ldots, \mathbf{r}_m^* \in \mathbb{R}^k$ és $\mathbf{q}_1^*, \ldots, \mathbf{q}_n^* \in \mathbb{R}^k$ az $\mathbf{A}_{m\times n}$ mátrix strukturális szinguláris értékeihez tartozó szinguláris vektorpárokkal gyártott sor- és oszlop-reprezentánsok.

Perturbácós tétel a normált kontingenciatáblára

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

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

akkor az $\mathbf{A}_{m\times n}$ mátrix elemei [0,1]-beliek. (Ez fontos lesz a 9. Tézisben, ahol majd valószínűségeknek tekintem őket.)

Tétel

A fenti jelölésekkel és a K-ra tett megállapodással, van olyan $0 < \delta < 1$ konstans (m-től és n-től függetlenül), hogy tetszőleges $0 < \tau < 1/2$ választással: az $\mathbf{A}_{m\times n}$ mátrixból nyert korrespondancia-mátrix k legnagyobb szinguláris értéke a $[\delta - \max\{n^{-\tau}, m^{-\tau}\}, 1]$ intervallumba esik, míg a többi 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.*

8. Tézis

Általánosabb típusú klaszterezéseket, ún. térfogat-reguláris klasztereket kerestem úgy, hogy a klaszterpárok közti és a klasztereken belüli diszkrepanciát próbáltam minimalizálni. Élsúlyozott gráfoknál beláttam, hogy amenyiben nincsenek domináns csúcsok, a k-klaszteres esetben a klaszterpárok közti maximális diszkrepancia felülről becsülhető a normált modularitás-mátrix $k-1$ legnagyobb abszolút értékű sajátértéke és a többi közötti réssel, valamint e $k-1$ strukturális sajátértékhez tarozó sajátvektorok által definiált reprezentánsok súlyozott k-varianciájával. A tételt általánosítottam kontingenciatáblákra és ún. térfogat-reguláris sor–oszlop klaszterpárokra, és alkalmaztam irányított gráfok kimeneti és bemeneti klasztereinek keresésére is, melyek közti információáramlás a lehető leghomogénebb.

B: *Modularity spectra, eigen-subspaces and structure of weighted graphs*, European J. Comb. 35 (2014), 105-116.

Expander mixing lemma élsúlyozott gráfokra

Tétel

Legyen $G = (V, W)$ élsúlyozott gráf és tegyük fel, hogy $\mathrm{Vol}(V) = 1$. Akkor tetszőleges $X, Y \subset V$ esetén

$$|w(X, Y) - \mathrm{Vol}(X)\mathrm{Vol}(Y)| \leq \|\mathbf{M}_D\| \cdot \sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}$$

teljesül, ahol $\|\mathbf{M}_D\|$ a G gráf normált modularitás-mátrixának spektrálnormája.

Mivel a G-beli ún. spektrális rés $1 - \|\mathbf{M}_D\|$, a fenti tétel értelmében, a 'nagy' spektrális rés azt jelzi, hogy bármely két csúcshalmaz közti súlyozott vágás közel van ahhoz, mint amit a csúcsok független, általános fokszámaikkal arányos valószínűséggel való kapcsolódása esetén remélnénk, azaz 'kicsi' a diszkrepanciájuk (kvázirandom tulajdonság).
Még általánosabban előfordul, hogy a rés nem a spektrum végein valósul meg.

Térfogat-reguláris klaszterpárok

Definíció

Legyen $G = (V, \mathbf{W})$ *élsúlyozott gráf,* $\mathrm{Vol}(V) = 1$*. A diszjunkt* $A, B \subseteq V$ *klaszterpár* α*-térfogat reguláris, ha tetszőleges* $X \subset A$*,* $Y \subset B$ *csúcshalmazra*

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

ahol $\rho(A, B) = \frac{w(A,B)}{\mathrm{Vol}(A)\mathrm{Vol}(B)}$ *az* A, B *klaszterpár relatív sűrűsége.*

Diszkrepanca és moularitás-spektrum

Tétel

Tegyük fel, hogy G összefüggő, $\mathrm{Vol}(V) = 1$, és nincsenek domináns csúcsok. Legyenek G normált modularitás-mátrixának sajátértékei csökkenő abszolút értékek szerint

$$1 \geq |\mu_1| \geq \cdots \geq |\mu_{k-1}| > \varepsilon \geq |\mu_k| \geq \cdots \geq |\mu_n| = 0.$$

Definiáljuk a csúcsok $(V_1, \ldots, V_k)$ partícióját úgy, hogy az minimalizálja a $\mu_1, \ldots, \mu_{k-1}$ sajátértékekhez tartozó sajátvektorok által legyártott $(k-1)$-dimenziós reprezentánsok súlyozott k-varianciáját, és jelölje a minimumot s^2. Tegyük fel, hogy valamely $0 < K \leq \frac{1}{k}$ konstanssal $|V_i| \geq Kn$, $i = 1, \ldots, k$. Akkor a (V_i, V_j) párok $\mathcal{O}(\sqrt{2ks^2} + \varepsilon)$-térfogat regulárisak $(i \neq j)$ és magukra a V_i klaszterekre: tetszőleges $X, Y \subset V_i$ esetén

$$|w(X, Y) - \rho(V_i)\mathrm{Vol}(X)\mathrm{Vol}(Y)| = \mathcal{O}(\sqrt{2ks^2} + \varepsilon)\mathrm{Vol}(V_i).$$

Az Expander mixing lemma kiterjesztése kontingenciatáblákra és irányított gráfokra

$\mathbf{W}$ $n \times n$-es élsúly mátrix (zéró diagonálissal), ahol w_{ij} az $i \to j$ irányított él súlya $(i, j = 1, \ldots, n;\ i \neq j)$.
Ekkor az általánosított be- és kifokok:

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

$\mathbf{D}_{in} = \mathrm{diag}(d_{in,1}, \ldots, \mathbf{d}_{in,n})$, $\mathbf{D}_{out} = \mathrm{diag}(d_{out,1}, \ldots, \mathbf{d}_{out,n})$,
$\mathbf{W}_{corr} = \mathbf{D}_{out}^{-1/2}\mathbf{W}\mathbf{D}_{in}^{-1/2}$. A fenti irányított gráf csúcsainak V_{in}, V_{out} be- és kimeneti klaszterei α-térfogat regulárisak, ha tetszőleges $X \subset V_{out}$ és $Y \subset V_{in}$ választással

$$|w(X, Y) - \rho(V_{out}, V_{in})\mathrm{Vol}_{out}(X)\mathrm{Vol}_{in}(Y)| \leq \alpha\sqrt{\mathrm{Vol}_{out}(V_{out})\mathrm{Vol}_{in}(V_{in})},$$

ahol $w(X, Y)$ az $X \to Y$ irányított vágás,
$\rho(V_{out}, V_{in}) = \frac{w(V_{out}, V_{in})}{\mathrm{Vol}(V_{out})\mathrm{Vol}(V_{in})}$.
A minimális α becslése $\mathbf{W}_{corr}$ SVD-jével.

9. Tézis

A Lovász László és társszerzői által kifejlesztett gráfkonvergencia és gráfparaméter tesztelhetőségi elméletet alkalamaztuk csúcs- és élsúlyozott gráfokra, továbbá kiterjesztettem kontingenciatáblákra. A minimális vágások általában nem tesztelhetők, de beláttuk, hogy a klaszterméretekre tett különböző kiegyensúlyozottsági feltételek mellett bizonyos minimális többszempontú vágássűrűségek tesztelhetők. 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 a csúcsreprezentánsok k-varianciája tesztelhető. Megmutattuk továbbá, hogy a 6. és 7. Tézisekben vizsgált zajos gráf- és kontingenciatábla-sorozatok a homomorfizmussűrűségekkel definiált értelemben konvergálnak.

B, Kói T, Krámli A: *Testability of minimum balanced multiway cut densities* Discrete Appl. Math. 160 (2012), 1019-1027.

B: *Modularity spectra, eigen-subspaces and structure of weighted graphs*, European J. Comb. 35 (2014), 105-116.

Konvergens gráfsorozatok

$G = G_n$ súlyozott, irányítatlan gráf n csúccsal.
Az éleket és a csúcsokat is súlyozzuk.
Élsúlyok: $\beta_{ij} = \beta_{ji} \in [0, 1]$ (az i és j csúcsok közti hasonlóság, hurokél megengedett).
Csúcssúlyok: $\alpha_i > 0$ (a csúcsok relatív fontosságát fejezi ki).
$\mathcal{G}$: az összes ilyen gráf halmaza.
$G \in \mathcal{G}$ térfogata: $\alpha_G = \sum_{i=1}^{n} \alpha_i$.
Az $S \subset V$ csúcshalmaz térfogata: $\alpha_S = \sum_{i \in S} \alpha_i$.
Egy konvergens gráfsorozat növekvő csúcsszámú elemei apró részleteikben egyre hasonlóbakká válnak egymáshoz,
Cauchy-sorozatok a $\delta_\square$ ún. vágás-metrikában, és határértékük egy $[0, 1] \times [0, 1] \to [0, 1]$ kétváltozós, szimmetrikus függvény, ún. grafon, ami csak a $[0, 1]$-en mértéktartó transzformációk erejéig egyértelmű (összhangban azzal, hogy a csúcsok átcímkézése sem befolyásolja a konvergenciát).

Mintavételezés

'Nagy' n-re G_n-ből egy $k < n$ csúccsal rendelkező F egyszerű gráfot sorsolunk ki a következőképpen: k csúcsot visszatevéssel választunk G csúcsai közül, α_i/α_G valószínűségekkel; a kiválasztott csúcspárokat ezután olyan valószínűséggel kötjük össze, mint ami annak az élnek a súlya volt, ami G-ben összekötötte őket. Az így nyert véletlen egyszerű gráfot $\xi(k, G)$-vel jelöljük. (Megjegyezzük, hogy $k \ll n$ esetén akár visszatevés nélkül is sorsolhatnánk.)

Egy $f : \mathcal{G} \to \mathbb{R}$ függvényt gráfparaméternek nevezünk, ha izomorf gráfokra ugyanazt az értéket veszi fel (pl. a spektrum is ilyen).

Legyen $(G_n) \subset \mathcal{G}$ olyan gráfsorozat, hogy tagjaiban nincsenek domináns csúcsok, azaz $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0$, ha $n \to \infty$.
Ilenekre, az egyszerű gráfok analógiájára, kiterjeszthető a tesztelhetőség fogalma.

Tesztelhető súlyozott gráfparaméterek

Definíció

Az f súlyozott gráfparaméter tesztelhető, ha minden $\varepsilon > 0$ esetén van olyan k pozitív egész, hogy amennyiben $G \in \mathcal{G}$ teljesíti a

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \leq \frac{1}{k}$$

feltételt, akkor

$$\mathbb{P}(|f(G) - f(\xi(k, G))| > \varepsilon) \leq \varepsilon$$

is teljesül.

A fenti tesztelhetőséggel ekvivalens, hogy

- f folytonos a $\delta_\square$-metrikában,
- konvergens G_n súlyozott gráfsorozatokra $f(G_n)$ is konvergens.

Minimális kiegyensúlyozott vágás-sűrűségek

Jól ismert tény, hogy a maximális vágássűrűség tesztelhető. A minimális vágások esetében ez nincsen így, hiszen például ha egyetlen csúcsot egy kis súlyú él köt egy teljes gráf egyetlen csúcsához, akkor az ennek az élnek az elmozdításával kapott minimális vágássűrűség 'kicsi', míg egy kisebb részt randomizálva, annak 'nagy' a minimális vágássűrűsége. Azonban, ha több részre vágás esetén ún. kiegyensúlyozottsági feltételeket teszünk a csúcsklaszterek méretére vagy térfogatára, akkor az így kapott minimális kiegyensúlyozott vágás-sűrűségek tesztelhetők (használtuk a Lovász László és társszerzői által kidolgozott statisztikus fizikai apparátust).
A blokkméretekre tett kiegyensűlyozttsági feltételek mellett a 6. Tézisben vizsgált zajos $G_{\mathbf{A}_n}$ $(\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n)$ gráfsorozatok konvergálnak. (Az ilyen általánosított véletlen gráfsorozatok determinisztikus megfelelői az általánosított kvázirandom gráfsorozatok, melyeket Lovász László és T. Sós Vera vezettek be.)

Domináns csúcssúlyok hiányában a normált modularitás-spektrum tesztelhető

Tétel

Legyen $G_n = (V_n, \mathbf{W}_n)$ egy konvergens súlyozott gráfsorozat általános tagja [0,1]-beli élsúlyokkal, a csúcsok súlyai az általánosított fokszámok (nincsenek domináns csúcsok). Jelölje W a (G_n) sorozat limit-grafonját,
$1 \geq |\mu_{n,1}| \geq |\mu_{n,2}| \geq \cdots \geq |\mu_{n,n}| = 0$ *pedig G_n normált modularitás-mátrixának spektrumát. Jelölje $\mu_i(P_{\mathbb{W}})$ 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 (mint az 5. Tézisben). Akkor $\forall i \geq 1$:*

$$\mu_{n,i} \to \mu_i(P_{\mathbb{W}}), \quad ha \quad n \to \infty.$$

A spektrális klaszterek is tesztelhetők

Tétel

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 sajátértékekhez tartozó ortonormált sajátvektorokat. Akkor, domináns csúcsok hiányában, a transzformált* $\mathbf{D}_n^{-1/2}\mathbf{u}_{n,1}, \ldots, \mathbf{D}_n^{-1/2}\mathbf{u}_{n,k-1}$ *vektorok által kifeszített* $(k-1)$*-dimenziós altér konvergál a* $P_{\mathbb{W}}$ *operátor analóg alterérhez.*

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 $\|\mathbf{P}_{n,k-1} - \mathbf{P}_{k-1}\| \to 0$ $(n \to \infty)$. A k-variancia is tesztelhető, mivel ezen alterek folytonos függvénye $\Longrightarrow$ kisebb, randomizált rész is alkalmas klaszterezésre.

10. Tézis

Néhány algoritmus és alkalmazás, melyek az előző tézisekben kifejtett tételeken alapulnak egy kivételével, ahol az EM-algoritmust adaptáltam a sztochasztikus blokkmodell paramétereinek becslésére.

B: *Spectral Clustering and Biclustering*, Wiley (2013).

Spektrális klaszterezés a súlyozott k-közép algoritmussal

Az 1. és 8. Tézisek elmélete alapján a normált modularitás-mátrix spektrumának vizsgálata után teszek javaslatot a klaszterszám és a sajátvektorok választására az alábbiak szerint.

Legyenek

$$1 \geq |\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_n| = 0$$

az élsúlyozott gráfhoz tartozó $\mathbf{M}_D$ mátrix sajátértékei.

Ha n 'nagyon nagy', elég néhány vezető sajátértéket meghatározni (az erre a célra használt numerikus algoritmusok és randomizált eljárások egy része úgyis a nagy abszolút értékű sajátértékek meghatározására lett kifejlesztve).

Klaszterek típusai

Ezután válasszunk egy k egészet úgy, hogy $|\mu_{k-1}|$ és $|\mu_k|$ közt 'rés' van.

- Ha $\mu_1, \ldots, \mu_{k-1}$ mind pozitív, akkor a hozzájuk tartozó transzformált sajátvektorokkal reprezentálva és klaszteresítve a a gráf minimális normált k-vágására 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'.
- Ha $\mu_1, \ldots, \mu_{k-1}$ közt vannak pozitív és negatív előjelűek is, akkor a hozzájuk tartozó transzformált sajátvektorokkal reprezentálva és klaszteresítve 'kis' diszkrepanciájú klaszterpárokat kapunk: 'regular structure'.

Fél-paraméteres módszerek

Legyen a statisztikai minta egy n csúcson értelmezett egyszerű gráf $n \times n$-es, szimmetrikus szomszédsági mátrixa.
A következő, sztochasztikus blokk-modell paramétereit becsüljük:

- Adott k egészre $(1 < k < n)$ a csúcsok függetlenül tartoznak a V_a klaszterekbe π_a valószínűséggel, $a = 1, \ldots, k$; $\sum_{a=1}^{k} \pi_a = 1$.
- V_a és V_b csúcsai egymástól függetlenül,

$$\mathbb{P}(i \sim j | i \in V_a, j \in V_b) = p_{ab}, \quad 1 \leq a, b \leq k$$

valószínűséggel vannak összekötve.

A Dempster, Laird és Rubin által 1977-ben leírt EM (Expectation–Maximization) algoritmust alkalmaztam erre az esetre.