Spectral Clustering and Biclustering

Marianna Bolla
Institute of Mathematics
Budapest University of Technology and Economics
marib@math.bme.hu

Barcelona, July 17, 2012

Motivation

- To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.
- To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.
- To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.

Motivation

- To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.
- To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.
- To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.

Motivation

- To recover the structure of large edge-weighted graphs, for example: biological, social, economic, or communication networks.
- To find a clustering (partition) of the vertices such that the induced subgraphs on them and the bipartite subgraphs between any pair of them exhibit regular behavior of information flow within or between the vertex subsets.
- To find biclustering of a contingency table (e.g., microarray) such that clusters of equally functioning genes equally influence conditions of the same cluster.

Spectral clustering of edge-wighted graphs

$G = (V, \mathbf{W})$ edge-weighted graph, $|V| = n$, $\mathbf{W}$: weight matrix of edges

$w_{ij} = w_{ji} \geq 0$ $(i \neq j)$ and $w_{ii} = 0$ (i=1,. . . ,n).

$d_i := \sum_{j=1}^{n} w_{ij}$ $(i = 1, \dots, n)$ generalized degrees

$\mathbf{d} := (d_1, \dots, d_n)^T$: degree vector, $\sqrt{\mathbf{d}} := (\sqrt{d_1}, \dots, \sqrt{d_n})^T$

$\mathbf{D} := \texttt{diag}\,(d_1, \dots, d_n)$: degree matrix

w.l.g. $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$ will be supposed

Laplacian and modularity matrices

$\mathbf{L} = \mathbf{D} - \mathbf{W}$: Laplacian
$\mathbf{L}_D = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$: normalized Laplacian
$\mathrm{Spec}\,(\mathbf{L}_D) \in [0, 2]$
If G is connected ($\mathbf{W}$ is irreducible), then 0 is a single eigenvalue with corresponding unit-norm eigenvector $\sqrt{\mathbf{d}}$.

$\mathbf{M}_D = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$: normalized modularity matrix
B, Phys. Rev. E (2011) $\mathrm{Spec}\,(\mathbf{M}_D) \in [-1, 1]$
1 cannot be an eigenvalue if G is connected, and 0 is always an eigenvalue with eigenvector $\sqrt{\mathbf{d}}$.

The spectral gap of G: $1 - \|\mathbf{M}_D\|$ (spectral norm)

Quadratic placement problems

Fact: the spectral decomposition of either $\mathbf{L}_D$ or $\mathbf{M}_D$ solves the following quadratic placement problem.
for a given positive integer k $(1 < k < n)$, minimize

$$Q_k(\mathbf{X}) = \sum_{i<j} w_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2$$

on the conditions

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

where the vectors $\mathbf{r}_1, \dots, \mathbf{r}_n$ are $(k-1)$-dimensional representatives of the vertices, which form the row vectors of the $n \times (k-1)$ matrix $\mathbf{X}$.

Normalized Laplacian eigenvalues

G is connected, $0 = \lambda_0 < \lambda_1 \le \cdots \le \lambda_{n-1} \le 2$ eigenvalues of $\mathbf{L}_D$ with corresponding unit-norm, pairwise orthogonal eigenvectors $\mathbf{u}_0 = \sqrt{\mathbf{d}}, \mathbf{u}_1, \dots, \mathbf{u}_{n-1}$.

In B, Tusnády, Discrete Math. (1994): the minimum of $Q_k(\mathbf{X})$ under the constraints for the representatives is

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

and is attained by the following representation:
$\mathbf{r}_1^*, \dots, \mathbf{r}_n^*$ are row vectors of the matrix
$\mathbf{X}^* = (\mathbf{D}^{-1/2}\mathbf{u}_1, \dots, \mathbf{D}^{-1/2}\mathbf{u}_{k-1})$.

Explanation

Instead of $\mathbf{X}$ the augmented $n \times k$ matrix $\tilde{\mathbf{X}}$ can as well be used, which is obtained from $\mathbf{X}$ by inserting the column $\mathbf{x}_0 = \mathbf{1}$ of all 1's. In fact, $\mathbf{x}_0 = \mathbf{D}^{-1/2}\mathbf{u}_0 = 1$, where $\mathbf{u}_0 = \sqrt{\mathbf{d}}$ is the eigenvector belonging to the eigenvalue 0 of $\mathbf{L}_D$. Then

$$Q_k(\tilde{\mathbf{X}}) = Q_k(\mathbf{X}) = \mathtt{tr}\,(\mathbf{D}^{1/2}\tilde{\mathbf{X}})^T(\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})(\mathbf{D}^{1/2}\tilde{\mathbf{X}}),$$

and $Q_k(\mathbf{X})$ is minimized on the constraint $\tilde{\mathbf{X}}^T\mathbf{D}\tilde{\mathbf{X}} = \mathbf{I}_k$,
or equivalently,
$\mathbf{D}^{1/2}\tilde{\mathbf{X}}$ is suborthogonal.

Continuous relaxation of a discrete minimization

This problem is the continuous relaxation of minimizing

$$Q_k(\tilde{\mathbf{X}}(P_k)) = \mathtt{tr}\,(\mathbf{D}^{1/2}\tilde{\mathbf{X}}(P_k))^T(\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})(\mathbf{D}^{1/2}\tilde{\mathbf{X}}(P_k))$$

over the set of k-partitions $P_k = (V_1, \dots, V_k)$ of the vertices such that P_k is planted into $\tilde{\mathbf{X}}$ in the way that the columns of $\tilde{\mathbf{X}}(P_k)$ are so-called partition-vectors belonging to P_k:
the coordinates of the ith column are zeros, except those indexing vertices of V_i which are equal to
$\frac{1}{\sqrt{\mathtt{Vol}\,(V_i)}}, \quad i = 1, \dots, k.$

Representation, spectral relaxation

$Q_k(\tilde{\mathbf{X}}(P_k))$ is the normalized cut of $P_k = (V_1, \dots, V_k)$:

$$Q_k(\tilde{\mathbf{X}}(P_k)) = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \left(\frac{1}{\mathtt{Vol}\,(V_a)} + \frac{1}{\mathtt{Vol}\,(V_b)} \right) w(V_a, V_b)$$

$$= \sum_{a=1}^{k} \frac{w(V_a, \bar{V}_a)}{\mathtt{Vol}\,(V_a)} = k - \sum_{a=1}^{k} \frac{w(V_a, V_a)}{\mathtt{Vol}\,(V_a)}$$

Minimum k-way normalized cut of $G = (V, \mathbf{W})$:

$$f_k(G) = \min_{P_k \in \mathcal{P}_k} Q_k(\tilde{\mathbf{X}}(P_k)),$$

where $\mathtt{Vol}\,(U) = \sum_{i \in U} d_i$: volume of $U \subset V$
$w(X, Y) = \sum_{i \in X} \sum_{j \in Y} w_{ij}$: weighted cut between $X, Y \subset V$

Estimation, references

Because of the spectral relaxation:

$$f_k(G) \geq \sum_{i=0}^{k-1} \lambda_i = \sum_{i=1}^{k-1} \lambda_i$$

B, Tusnády, Discrete Math. (1994) general k, called weighted cut
Azran, Ghahramani, Siam J. Comput (2000) general k
Meila and Shi, NIPS (2001): $k = 2$
B, M-Sáska, Studia Sci. Math. Hun. (2002) general k
Upper estimate: depends on the corresponding eigenvectors.
Point of spectral clustering: optimizing over $\mathcal{P}_k$ is NP-hard.

Isoperimetric number

Definition

The Cheeger constant of the weighted graph $G = (V, W)$ is

$$h(G) = \min_{\substack{U \subset V \\ \mathtt{Vol}\,(U) \le 1/2}} \frac{w(U, \bar{U})}{\mathtt{Vol}\,(U)}$$

Theorem

(B, M-Sáska, Discrete Math. (2004)). Let λ_1 be the smallest positive eigenvalue of $\mathbf{L}_D$. Then

$$\frac{\lambda_1}{2} \le h(G) \le \min\{1, \sqrt{2\lambda_1}\}.$$

If $\lambda_1 \le 1$ (G is not the complete graph), then

$$h(G) \le \sqrt{\lambda_1(2 - \lambda_1)}.$$

Normalized Newman–Girvan modularity

Newman–Girvan, Physical Review E (2004) N-G mod.
B, Physical Review E (2011) normalized N-G mod.:

$$M_k(\mathbf{W}, P_k) = \sum_{a=1}^{k} \frac{1}{\mathtt{Vol}\,(V_a)} \sum_{i,j \in V_a} (w_{ij} - d_i d_j) = \sum_{a=1}^{k} \frac{w(V_a, V_a)}{\mathtt{Vol}\,(V_a)} - 1$$

Since

$$M_k(\mathbf{W}, P_k) = k - 1 - Q_k(\tilde{\mathbf{X}}(P_k)),$$

maximizing the k-way normalized Newman-Girvan modularity is equivalent to the normalized cut problem and it can be solved by the same spectral relaxation.

Spectral gap and variance

Weighted k-variance of the vertex representatives:

$$S_k^2(\mathbf{X}) = \min_{P_k=(V_1,\dots,V_k)} \sum_{a=1}^{k} \sum_{j\in V_a} d_j \|\mathbf{r}_j - \mathbf{c}_a\|^2$$

where $\mathbf{c}_a = \frac{1}{\mathtt{Vol}\,(V_a)} \sum_{j\in V_a} d_j \mathbf{r}_j$.

In B, Tusnády, Discrete Math. (1994)

Theorem

In the representation $\mathbf{X}^* = (\mathbf{D}^{-1/2}\mathbf{u}_0, \mathbf{D}^{-1/2}\mathbf{u}_1) = (\mathbf{1}, \mathbf{D}^{-1/2}\mathbf{u}_1)$*:*

$$S_2^2(\mathbf{X}^*) \le \frac{\lambda_1}{\lambda_2}$$

$f_2(G)$ is the symmetric version of $h(G)$: $f_2(G) \le 2h(G) \implies$

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

Theorem

Suppose that $G=(V,\mathbf{W})$ is connected, and λ_i's are the eigenvalues of $\mathbf{L}_D$. Then $\sum_{i=1}^{k-1} \lambda_i \le f_k(G)$ and in the case when the optimal k-dimensional representatives can be classified into k well-separated clusters in such a way that the maximum cluster diameter ε satisfies the relation $\varepsilon \le \min\{1/\sqrt{2k}, \sqrt{2}\min_i \sqrt{\mathtt{Vol}\,(V_i)}\}$ with k-partition $(V_1,\dots,V_k)$ induced by the clusters above, then

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

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

Normalized modularity eigenvalues

$\mathbf{M}_D = \mathbf{I} - \mathbf{L}_D - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$ eigenvalues:
$1 - \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq -1$ with the same eigenvectors and 0 with eigenvector $\sqrt{\mathbf{d}}$.
1 cannot be an eigenvalue if G is connected / W is irreducible

- Large absolute value positive eigenvalues of $\mathbf{M}_D$ are responsible for clusters with high intra- and low inter-cluster densities.
- If we minimize $M_k(\mathbf{W}, P_k)$ instead of maximizing over $\mathcal{P}_k$: small negative eigenvalues of $\mathbf{M}_D$ are responsible for clusters with low intra- and high inter-cluster densities.
- If we take into account eigenvalues from both ends of the normalized modularity spectrum, we can recover so-called regular cluster pairs.

Volume regularity

Lemma

Expander Mixing Lemma for weighted graphs: Supposing $\texttt{Vol}\,(V) = 1$, *for all* $X, Y \subset V$,

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

For simple graphs: Alon, Combinatorica (1986)
Hoory, Linial, Widgerson, Bulletin of AMS (2006)
For edge-weighted graphs: Chung, Graham, Random structures and algorithms (2008), in context of quasi-random properties.

What if the gap is not at the ends of the spectrum?

We want to partition the vertices into clusters so that a relation formulated in the Lemma (1-cluster case) between the edge-densities and volumes of the cluster pairs would hold.
We will use a slightly modified version of the volume regularity's notion introduced by Alon, Coja-Oghlan, Han, Kang, Rödl, and Schacht, Siam J. Comput. (2010):

Definition

Let $G = (V, \mathbf{W})$ be a weighted graph with $\texttt{Vol}\,(V) = 1$. The disjoint pair (A, B) is α-volume regular if for all $X \subset A$, $Y \subset B$ we have

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

where $\rho(A, B) = \frac{w(A,B)}{\texttt{Vol}\,(A)\texttt{Vol}\,(B)}$ is the relative inter-cluster density of (A, B).

Outline

For general deterministic edge-weighted graphs we'll prove that the existence of $k-1$ eigenvalues of $\mathbf{M}_D$ separated from 0 by ε, is indication of a k-cluster structure, while the eigenvalues accumulating around 0 are responsible for the pairwise regularities.

The clusters themselves can be recovered by applying the k-means algorithm for the vertex representatives obtained by the eigenvectors corresponding to the structural eigenvalues.

Our theorem bounds the volume regularity's constants of the different cluster pairs by means of ε and the k-variance of the vertex representatives (based on the structural eigenvectors). Estimates for the intra-cluster densities are also given.

Result

Theorem

$G = (V, \mathbf{W})$ is edge-weighted graph on n vertices, $\mathtt{Vol}\,(V) = 1$ and there are no dominant vertices: $d_i = \Theta(1/n)$, $i = 1, \dots, n$ as $n \to \infty$. The eigenvalues of $\mathbf{M}_D$ in decreasing absolute values are:

$$1 > |\mu_1| \geq \cdots \geq |\mu_{k-1}| > \varepsilon \geq |\mu_i|, \quad i \geq k.$$

The partition $(V_1, \dots, V_k)$ of V is defined so that it minimizes the weighted k-variance $s^2 = S_k^2(\mathbf{X}^*)$ of the vertex representatives. Suppose that there is a constant $0 < c \leq \frac{1}{k}$ such that $|V_i| \geq cn$, $i = 1, \dots, k$. Then the (V_i, V_j) pairs are $\mathcal{O}(\sqrt{2k}s + \varepsilon)$-volume regular $(i \neq j)$ and for the clusters V_i $(i = 1, \dots, k)$ the following holds: for all $X, Y \subset V_i$,
$|w(X, Y) - \rho(V_i)\mathtt{Vol}\,(X)\mathtt{Vol}\,(Y)| = \mathcal{O}(\sqrt{2k}s + \varepsilon)\mathtt{Vol}\,(V_i)$, where $\rho(V_i) = \frac{w(V_i, V_i)}{\mathtt{Vol}^2(V_i)}$ is the relative intra-cluster density of V_i.

Remark

The case $k = 2$ was treated separately in
B, International Journal of Combinatorics, 2011:

Under the same conditions and with notations $|\mu_1| = \theta$, $|\mu_2| = \varepsilon$, the (V_1, V_2) pair is $\mathcal{O}\left(\sqrt{\frac{1-\theta}{1-\varepsilon}}\right)$-volume regular.

Random graphs, Wigner-noise

Definition

The $n \times n$ symmetric real matrix $\mathbf{W}$ is a Wigner-noise if its entries w_{ij}, $1 \le i \le j \le n$, are independent random variables, $\mathbb{E} w_{ij} = 0$, $\mathtt{Var}\, w_{ij} \le \sigma^2$ with some $0 < \sigma < \infty$ and the w_{ij}'s are uniformly bounded (there is a constant $K > 0$ such that $|w_{ij}| \le K$).

Füredi, Komlós, Combinatorica (1981):

$$\max_{1 \le i \le n} |\lambda_i(\mathbf{W})| \le 2\sigma\sqrt{n} + O(n^{1/3} \log n)$$

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

Sharp concentration theorem

Theorem

$\mathbf{W}$ 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 $\mathbf{W}$.
For any $t > 0$:

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

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)

Previous results imply:

Lemma

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

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

Borel–Cantelli Lemma $\Longrightarrow$
The spectral norm of $\mathbf{W}$ is $\mathcal{O}(\sqrt{n})$ almost surely.

Perturbation results for weighted graphs

$\mathbf{A} = \mathbf{B} + \mathbf{W}$, where

$\mathbf{W}$: $n \times n$ Wigner-noise

$\mathbf{B}$: $n \times n$ blown-up matrix of $\mathbf{P}$ with blow-up sizes $n_1, \dots, n_k$, $\sum_{i=1}^{k} n_i = n$.

$\mathbf{P}$: $k \times k$ pattern matrix

k is kept fixed as $n_1, \dots, n_k \to \infty$ "at the same rate": there is a constant c such that

$\frac{n_i}{n} \geq c, \quad i = 1, \dots k.$

growth rate condition: g.r.c.

Adjacency spectrum of a noisy graph

$G_n = (V, \mathbf{A})$, $\mathbf{A} = \mathbf{B} + \mathbf{W}$ is $n \times n$, $n \to \infty$
$\mathbf{B}$ induces a planted partition $P_k = (V_1, \dots, V_k)$ of V.
Weyl's perturbation theorem $\Longrightarrow$
Adjacency spectrum of G_n: under g.r.c. there are k structural eigenvalues of order n (in absolute value) and the others are $\mathcal{O}(\sqrt{n})$, almost surely.
The eigenvectors $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ corresponding to the structural eigenvalues are "not far" from the subspace of stepwise constant vectors on P_k $\Longrightarrow$

$$S_k^2(\mathbf{X}) \le S_k^2(P_k, \mathbf{X}) = \mathcal{O}(\frac{1}{n}), \quad \text{almost surely as} \quad n \to \infty.$$

This extends over the normalized Laplacian and modularity spectra.

Noisy graph is simple with appropriate noise

The uniform bound K on the entries of $\mathbf{W}$ is such that $\mathbf{A} = \mathbf{B} + \mathbf{W}$ has entries in [0,1].
With an appropriate Wigner-noise the noisy matrix $\mathbf{A}$ is a generalized random graph: edges between V_a and V_b exist with probability $0 < p_{ab} < 1$.
For $1 \le a \le b \le k$ and $i \in V_a$, $j \in V_b$:

$$w_{ij} := \begin{cases} 1 - p_{ab}, & \text{with probability} \quad p_{ab} \\ -p_{ab} & \text{with probability} \quad 1 - p_{ab} \end{cases}$$

be independent random variables, otherwise $\mathbf{W}$ is symmetric. The entries have zero expectation and bounded variance:

$$\sigma^2 = \max_{1 \le a \le b \le k} p_{ab}(1 - p_{ab}) \le \frac{1}{4}.$$

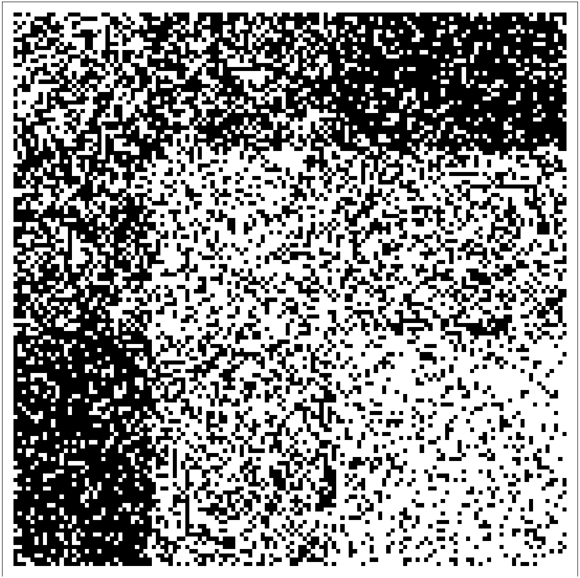
Generalized random graphs

Ideal k-cluster case: given the partition $(V_1, \dots, V_k)$ of V, vertices $i \in V_a$ and $j \in V_b$ are connected with probability p_{ab}, independently of each other, $1 \le a, b \le k$.
Generalized random graph: random simple graph= edge-weighted graph with a special block-structure+random noise $\Longrightarrow$ Spectral characterization in B, Discrete Math. (2008):
If k is fixed and $n \to \infty$ such that $\frac{|V_a|}{n} \ge c$ $(a = 1, \dots, k)$ with some $0 < c \le \frac{1}{k}$, then there exists a positive number $0 < \theta \le 1$, independent of n, such that for every $0 < \tau < 1/2$

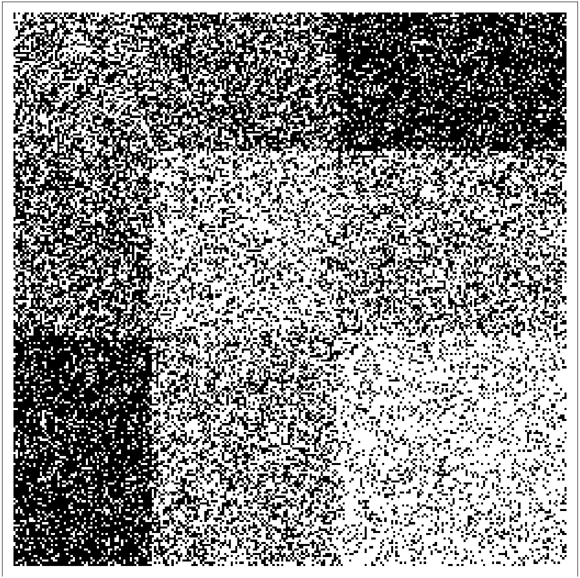
- there are exactly $k-1$ eigenvalues of $\mathbf{M}_D$ greater than $\theta - n^{-\tau}$, while all the others are at most $n^{-\tau}$ in absolute value,
- the k-variance of the vertex representatives constructed by the $k-1$ transformed structural eigenvectors is $\mathcal{O}(n^{-2\tau})$,
- with any "small" $\alpha > 0$, the V_a, V_b pairs are α-volume regular,

almost surely.

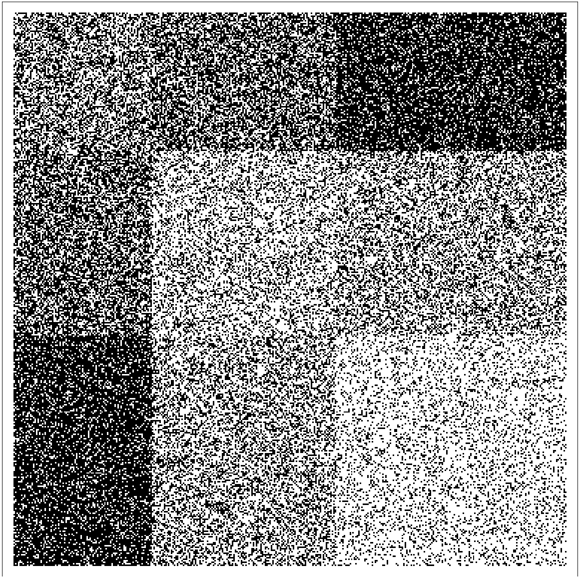
10-fold blow up



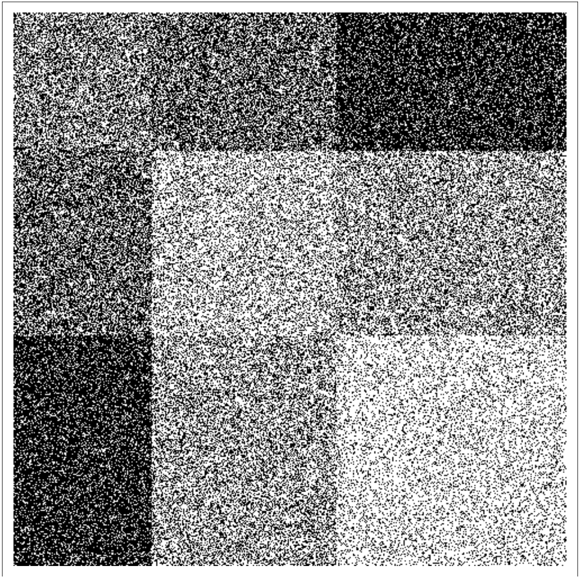
20-fold blow up



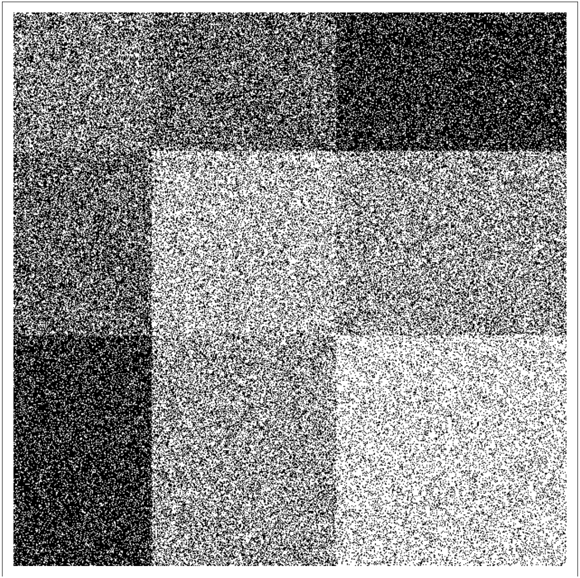
30-fold blow up



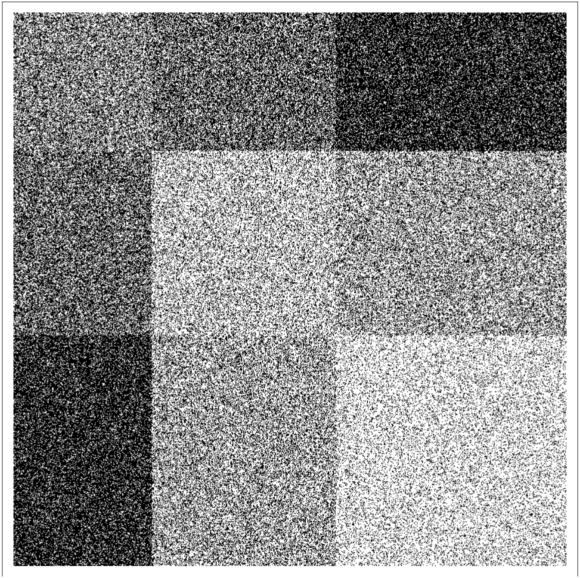
40-fold blow up



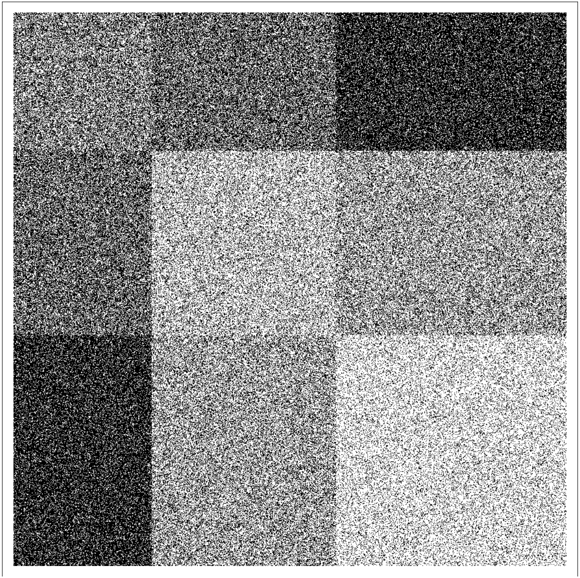
50-fold blow up



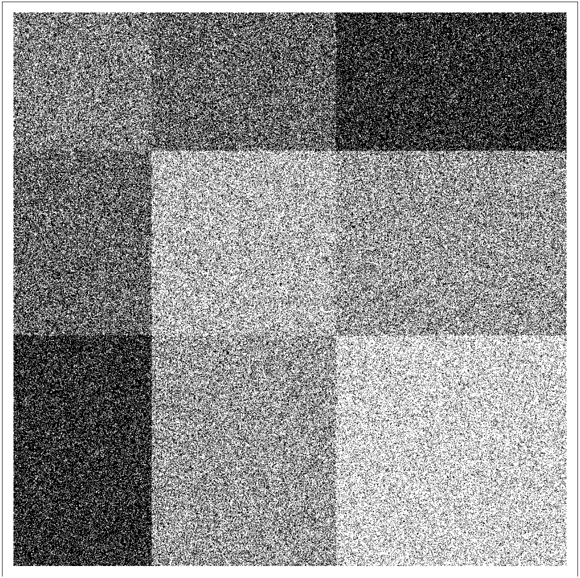
60-fold blow up



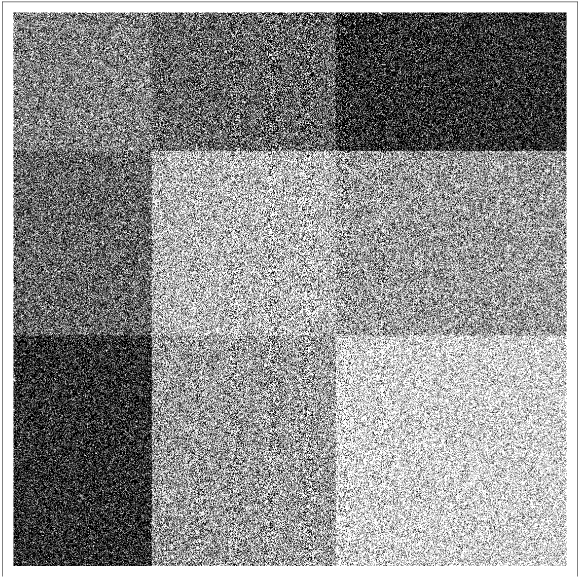
70-fold blow up



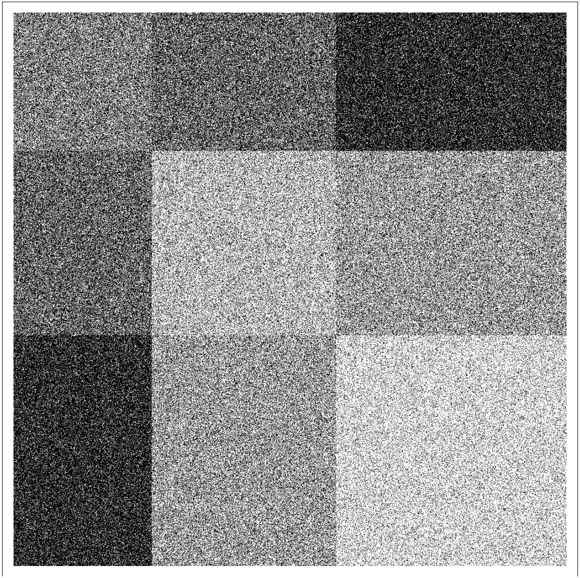
80-fold blow up



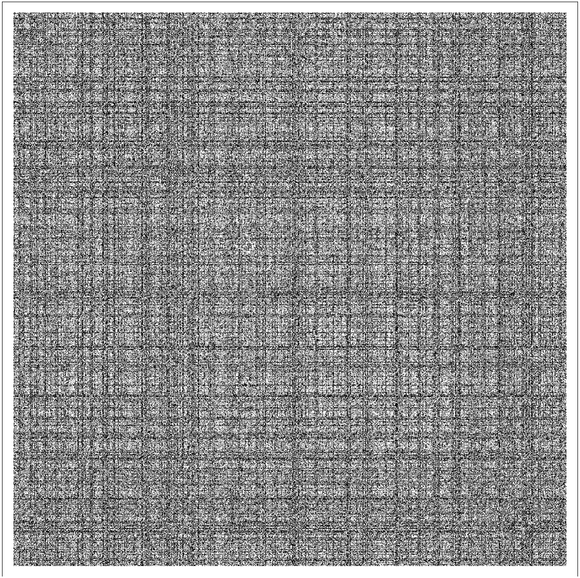
90-fold blow up



100-fold blow up



Before sorting and clustering the vertices



Biclustering of contingency tables

$(Row, Col, \mathbf{C})$ contingency table
Row set: $Row = \{1, \dots, n\}$
Column set: $Col = \{1, \dots, m\}$
$\mathbf{C}$: $n \times m$ matrix of entries $c_{ij} \geq 0$.
c_{ij}: some kind of interaction between the objects representing row i and column j, where 0 means no interaction at all.

$$d_{row,i} = \sum_{j=1}^{m} c_{ij}, \quad i = 1, \dots, n$$

$$d_{col,j} = \sum_{i=1}^{n} c_{ij}, \quad j = 1, \dots, m$$

$$\mathbf{D}_{row} = \mathtt{diag}\,(d_{row,1}, \dots, d_{row,n}), \quad \mathbf{D}_{col} = \mathtt{diag}\,(d_{col,1}, \dots, d_{col,m})$$

Quadratic placement problem

Given the integer $1 \leq k \leq \min\{n, m\}$: find k-dimensional representatives $\mathbf{r}_1, \dots, \mathbf{r}_n \in \mathbb{R}^k$ of the rows and $\mathbf{c}_1, \dots, \mathbf{c}_m \in \mathbb{R}^k$ of the columns such that they minimize

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

under the conditions

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

Equivalence to the correspondence analysis

$\mathbf{X} := (\mathbf{r}_1^T, \dots, \mathbf{r}_n^T)^T = (\mathbf{x}_1, \dots, \mathbf{x}_k) \quad n \times k$
$\mathbf{Y} := (\mathbf{c}_1^T, \dots, \mathbf{c}_m^T)^T = (\mathbf{y}_1, \dots, \mathbf{y}_k) \quad m \times k$
Constraints:

$$\mathbf{X}^T \mathbf{D}_{row} \mathbf{X} = \mathbf{I}_k, \quad \mathbf{Y}^T \mathbf{D}_{col} \mathbf{Y} = \mathbf{I}_k.$$

$$Q_k = \sum_{i=1}^{n} d_{row,i} \|\mathbf{r}_i\|^2 + \sum_{j=1}^{m} d_{col,j} \|\mathbf{c}_j\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \mathbf{r}_i^T \mathbf{c}_j$$

$$= 2k - \mathrm{tr}\, \mathbf{X}^T \mathbf{C} \mathbf{Y} = 2k - \mathrm{tr}\, (\mathbf{D}_{row}^{1/2} \mathbf{X})^T (\mathbf{D}_{row}^{-1/2} \mathbf{C} \mathbf{D}_{col}^{-1/2})(\mathbf{D}_{col}^{1/2} \mathbf{Y}),$$

where $\mathbf{D}_{row}^{1/2} \mathbf{X}$ and $\mathbf{D}_{col}^{1/2} \mathbf{Y}$ are suborthogonal matrices.

Correspondence matrix / normalized contingency table

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

$$\mathbf{C}_{corr} = \sum_{l=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^T ,$$

where $r \le \min\{n, m\}$ is the rank of $\mathbf{C}_{corr}$, or equivalently (as there are not identically zero rows or columns), that is the rank of $\mathbf{C}$.
$1 = s_1 \ge s_2 \ge \cdots \ge s_r > 0$: non-zero singular values of $\mathbf{C}_{corr}$ with singular vector pairs $\mathbf{v}_i, \mathbf{u}_i$ $(i = 1, \ldots, r)$.
1 is a single singular value if $\mathbf{C}_{corr}$ (or equivalently, $\mathbf{C}$) is irreducible. In this case
$\mathbf{v}_1 = (\sqrt{d_{row,1}}, \ldots, \sqrt{d_{row,n}})^T$ and $\mathbf{u}_1 = (\sqrt{d_{col,1}}, \ldots, \sqrt{d_{col,m}})^T$.

Representation theorem for contingency tables

Theorem

Let $(Row, Col, \mathbf{C})$ be an irreducible contingency table with the above SVD of its correspondence matrix $\mathbf{C}_{corr}$. Let $k \le r$ be a positive integer such that $s_k > s_{k+1}$. Then the minimum of Q_k under the given constraints is $2k - \sum_{i=1}^{k} s_i$ and it is attained with the optimum row representatives $\mathbf{r}_1^*, \dots, \mathbf{r}_n^*$ and column representatives $\mathbf{c}_1^*, \dots, \mathbf{c}_m^*$, the transposes of which are row vectors of $\mathbf{X}^* = \mathbf{D}_{row}^{-1/2}(\mathbf{v}_1, \dots, \mathbf{v}_k)$ and $\mathbf{Y}^* = \mathbf{D}_{col}^{-1/2}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, respectively.

Remark: if 1 is a single singular value, the first columns of $\mathbf{X}^*$ and $\mathbf{Y}^*$: $\mathbf{D}_{row}^{-1/2}\mathbf{v}_1$ and $\mathbf{D}_{col}^{-1/2}\mathbf{u}_1$ are the constantly $\mathbf{1}$ vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.

Normalized two-way cuts of a contingency table

$(Row, Col, \mathbf{C})$: $n \times m$ contingency table
k $(0 < k \leq r)$: fixed integer
Partition simultaneously the rows and columns into disjoint, nonempty subsets

$$Row = R_1 \cup \cdots \cup R_k, \quad Col = C_1 \cup \cdots \cup C_k$$

such that the cuts

$$c(R_a, C_b) = \sum_{i \in R_a} \sum_{j \in C_b} c_{ij}, \quad a, b = 1, \dots, k$$

between the row-column cluster pairs be as homogeneous as possible.

The normalized two-way cut of the contingency table with respect to the above k-partitions $P_{row} = (R_1, \dots, R_k)$ and $P_{col} = (C_1, \dots, C_k)$ of its rows and columns and to the collection of signs σ:

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

where

$$\mathtt{Vol}\,(R_a) = \sum_{i \in R_a} \sum_{j=1}^{m} c_{ij}, \quad \mathtt{Vol}\,(C_b) = \sum_{j \in C_b} \sum_{i=1}^{n} c_{ij}$$

are volumes of the clusters, and $\sigma = (\sigma_{11}, \dots, \sigma_{kk})$ with $\sigma_{aa} = \pm 1$ $(a = 1, \dots, k)$, whereas σ_{ab} has no relevance if $a \neq b$.
The objective function also penalizes clusters of extremely different volumes.

Theorem

The normalized two-way cut of the contingency table **C***:*

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

Proof: $\nu_k(P_{row}, P_{col}, \sigma)$ is Q_k in the special representation, where the column vectors of $\mathbf{X}$ and $\mathbf{Y}$ are partition vectors belonging to P_{row} and P_{col}:

$$x_{ia} := \frac{1}{\sqrt{\mathtt{Vol}\,(R_a)}} \text{ if } i \in R_a \text{ and } 0 \,\mathtt{otherwise}\, (a = 1, \dots k)$$

$$y_{jb} := \frac{\sigma_{bb}}{\sqrt{\mathtt{Vol}\,(C_b)}} \text{ if } j \in C_b \text{ and } 0 \,\mathtt{otherwise}\, (b = 1, \dots, k)$$

$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$ satisfy the conditions imposed on the representatives and

$$\|\mathbf{r}_i - \mathbf{c}_j\|^2 = \frac{1}{\mathtt{Vol}\,(R_a)} + \frac{1}{\mathtt{Vol}\,(C_b)} + \frac{2\delta_{ab}\sigma_{bb}}{\sqrt{\mathtt{Vol}\,(R_a)\mathtt{Vol}\,(C_b)}} \text{ if } i \in R_a,\ j \in C_b.$$

Symmetric contingency table = edge-weighted graph

- If the $k-1$ largest absolute value eigenvalues of the normalized modularity matrix are all positive: the $k-1$ largest singular values (apart of the 1) of $\mathbf{C}_{corr}$ are identical to the $k-1$ largest eigenvalues of $\mathbf{M}_D$, and the left and right singular vectors are identical to the corresponding eigenvector with the same orientation $\Longrightarrow$ $\mathbf{r}_i = \mathbf{c}_i$ for all $(k-1)$-dimensional row and column representatives; $\nu_k(\mathbf{C}) = 2f_k(G) \Longrightarrow$ the normalized two-way cut favors k-partitions with low inter-cluster edge-densities.
- If all the $k-1$ largest absolute value eigenvalues of the normalized modularity matrix are negative: $\mathbf{r}_i = -\mathbf{c}_i$, and any (but only one) of them can be the corresponding vertex representative; $\nu_k(\mathbf{C})$ differs from $f_k(G)$ in that it also counts the edge-weights within the clusters. Here, minimizing $\nu_k(\mathbf{C})$, rather a so-called anti-community structure is detected.

Regular row-column cluster pairs

In the generic case, for given k, if the clusters are formed via applying the k-means algorithm for the row- and column representatives, respectively, then the so obtained row–column cluster pairs are homogeneous in the sense, that they form equally dense parts of the contingency table.

Definition

The row–column cluster pair $R \subset Row$, $C \subset Col$ of the contingency table $(Row, Col, \mathbf{C})$ (where the sum of the entries is 1) is γ-volume regular, if for all $X \subset R$ and $Y \subset C$ the relation

$$|c(X,Y) - \rho(R,C)\mathtt{Vol}\,(X)\mathtt{Vol}\,(Y)| \le \gamma\sqrt{\mathtt{Vol}\,(R)\mathtt{Vol}\,(C)}$$

holds, where $\rho(R,C) = \frac{c(R,C)}{\mathtt{Vol}\,(R)\mathtt{Vol}\,(C)}$ is the relative inter-cluster density of the row–column pair R, C.

Weighted k-variances

The weighted k-variance of the k-dimensional row representatives:

$$S_k^2(\mathbf{X}) = \min_{P_{row,k} \in \mathcal{P}_{row,k}} S_k^2(P_k, \mathbf{X}) = \min_{(R_1,\dots,R_k)} \sum_{a=1}^{k} \sum_{j \in R_a} d_{row,j} \|\mathbf{r}_j - \mathbf{b}_a\|^2,$$

where $\mathbf{b}_a = \frac{1}{\mathtt{Vol}\,(R_a)} \sum_{j \in R_a} d_{row,j} \mathbf{r}_j \quad (a = 1, \dots, k)$.
The weighted k-variance of the k-dimensional column representatives:

$$S_k^2(\mathbf{Y}) = \min_{Q_{col,k} \in \mathcal{P}_{col,k}} S_k^2(Q_k, \mathbf{Y}) = \min_{(C_1,\dots,C_k)} \sum_{b=1}^{k} \sum_{j \in C_b} d_{col,j} \|\mathbf{c}_j - \mathbf{b}_b\|^2,$$

where $\mathbf{b}_b = \frac{1}{\mathtt{Vol}\,(C_b)} \sum_{j \in C_b} d_{col,j} \mathbf{c}_j \quad (b = 1, \dots, k)$.
Observe, that the trivial vector components can be omitted, and the k-variance of the so obtained $(k-1)$-dimensional representatives will be the same.

Volume regularity versus spectral properties

Theorem

Let $(Row, Col, \mathbf{C})$ be a contingency table of n rows and m columns, with row- and column sums $d_{row,1}, \dots, d_{row,n}$ and $d_{col,1}, \dots, d_{col,m}$, respectively. Suppose that $\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} = 1$ and there are no dominant rows and columns: $d_{row,i} = \Theta(1/n)$, $(i = 1, \dots, n)$ and $d_{col,j} = \Theta(1/m)$, $(j = 1, \dots, m)$ as $n, m \to \infty$. Let the singular values of $\mathbf{C}_{corr}$ be

$$1 = s_1 > s_2 \geq \cdots \geq s_k > \varepsilon \geq s_i, \quad i \geq k + 1.$$

The partition $(R_1, \dots, R_k)$ of Row and $(C_1, \dots, C_k)$ of Col are defined so that they minimize the weighted k-variances $S_k^2(\mathbf{X}^*)$ and $S_k^2(\mathbf{Y}^*)$ of the row and column representatives. Suppose that there are constants $0 < K_1, K_2 \leq \frac{1}{k}$ such that $|R_i| \geq K_1 n$ and $|C_i| \geq K_2 m$ $(i = 1, \dots, k)$, respectively. Then the R_i, C_j pairs are $\mathcal{O}(\sqrt{2k}(S_k(\mathbf{X}^*) + S_k(\mathbf{Y}^*)) + \varepsilon)$-volume regular $(i, j = 1, \dots, k)$.

Noisy contingency table sequences

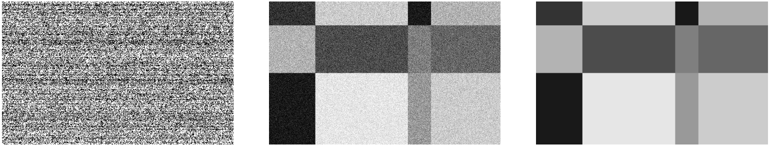


Figure: noisy table; table close to the limit; approximation by SVD

THE END