Spectra and Structure of Weighted Graphs

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

EUROCOMB'2011
Budapest, August 29, 2011

Motivation

- To recover the structure of large edge-weighted graphs, for example: metabolic, 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 estimate the constants bounding the volume regularity of the cluster pairs by means of spectral gaps and classification properties of eigenvectors.

Motivation

- To recover the structure of large edge-weighted graphs, for example: metabolic, 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 estimate the constants bounding the volume regularity of the cluster pairs by means of spectral gaps and classification properties of eigenvectors.

Motivation

- To recover the structure of large edge-weighted graphs, for example: metabolic, 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 estimate the constants bounding the volume regularity of the cluster pairs by means of spectral gaps and classification properties of eigenvectors.

Notation

$G = (V, \mathbf{W})$ edge-weighted graph, $|V| = n$, $\mathbf{W}$: weight matrix
$w_{ij} = w_{ji} \geq 0$ $(i \neq j)$ and $w_{ii} = 0$ (i=1,. . . ,n).

$d_i := \sum_{j=1}^{n} w_{ij}$ $(i = 1, \ldots, n)$ generalized degrees
$\mathbf{d} := (d_1, \ldots, d_n)^T$: degree vector, $\sqrt{\mathbf{d}} := (\sqrt{d_1}, \ldots, \sqrt{d_n})^T$

$\mathbf{D} := \mathtt{diag}\,(d_1, \ldots, d_n)$: 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 $\sqrt{\mathbf{d}}$.

$\mathbf{B}_D = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$: normalized modularity matrix
$\mathrm{Spec}\,(\mathbf{B}_D) \in [-1, 1]$
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: $1 - \|\mathbf{B}_D\|$ (spectral norm)

Volumes, weighted cuts, and mixing

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

Lemma

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

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

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 $\mathtt{Vol}(V) = 1$. The disjoint pair (A, B) is α-volume regular if for all $X \subset A$, $Y \subset B$ we have

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

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

Euclidean representation

Vertex representatives $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^{k-1}$:
row vectors of the $n \times (k-1)$ matrix
$\mathbf{X} = (\mathbf{D}^{-1/2}\mathbf{u}_1, \ldots, \mathbf{D}^{-1/2}\mathbf{u}_{k-1})$,
where $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ are unit-norm eigenvectors belonging to the *structural* (well separated from 0) *eigenvalues* of $\mathbf{B}_D$

Weighted k-variance of the $(k-1)$-dimensional representatives:

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

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

Generalized random graphs

Ideal k-cluster case: given the partition $(V_1, \ldots, V_k)$ of V, vertices $i \in V_a$ and $j \in V_b$ are connected with probability p_{ab}, independently of each other, $1 \leq a, b \leq k$.

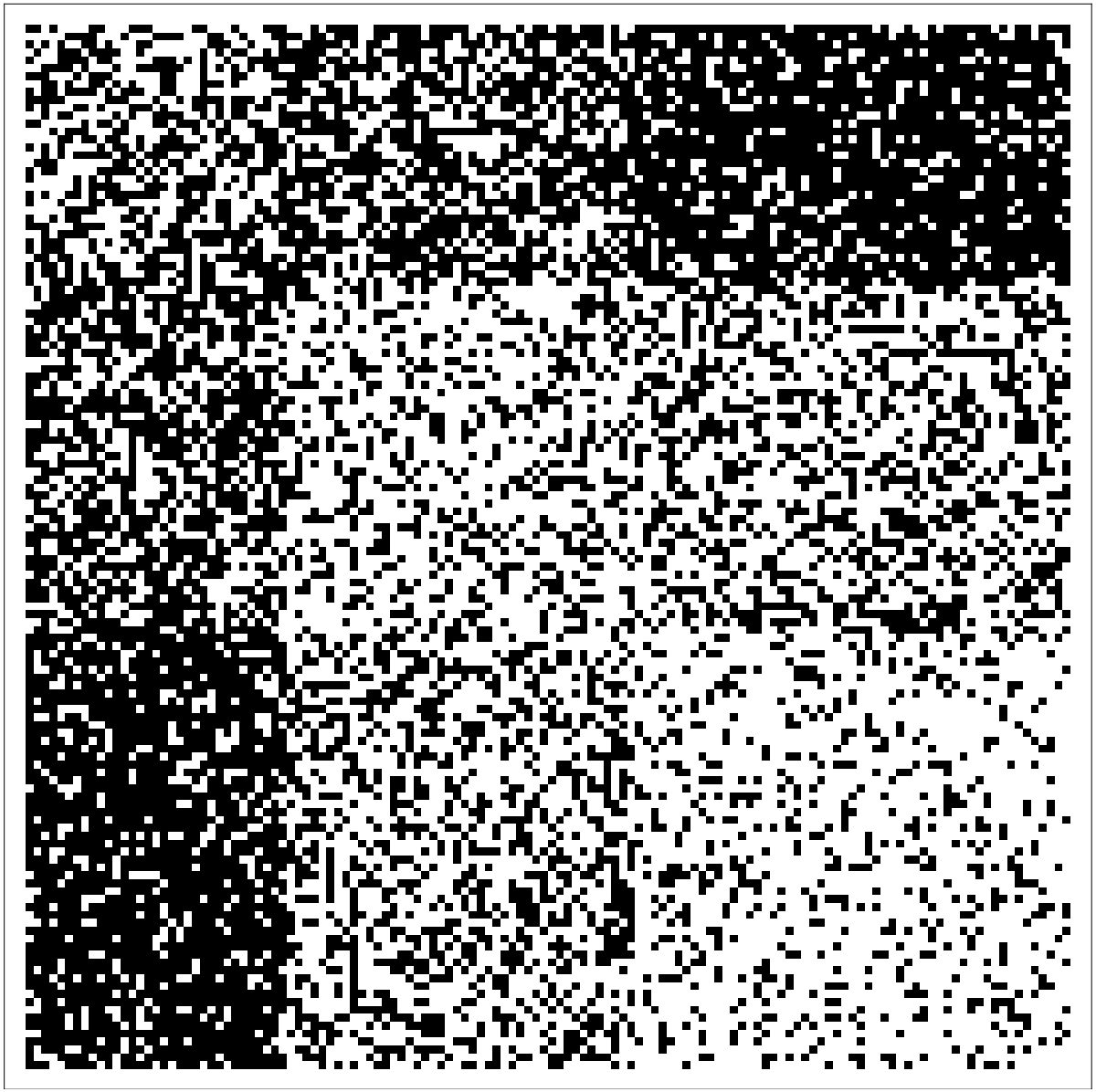
Generalized random graphs are edge-weighted graphs with a special block-structure burdened with random noise $\Longrightarrow$ Spectral characterization in B, Discrete Math. (2008):

If k is fixed and $n \to \infty$ such that $\frac{|V_i|}{n} \geq c$ $(i = 1, \ldots, k)$ with some $0 < c \leq \frac{1}{k}$, then there exists a positive number $0 < \theta \leq 1$, independent of n, such that for every $0 < \tau < 1/2$

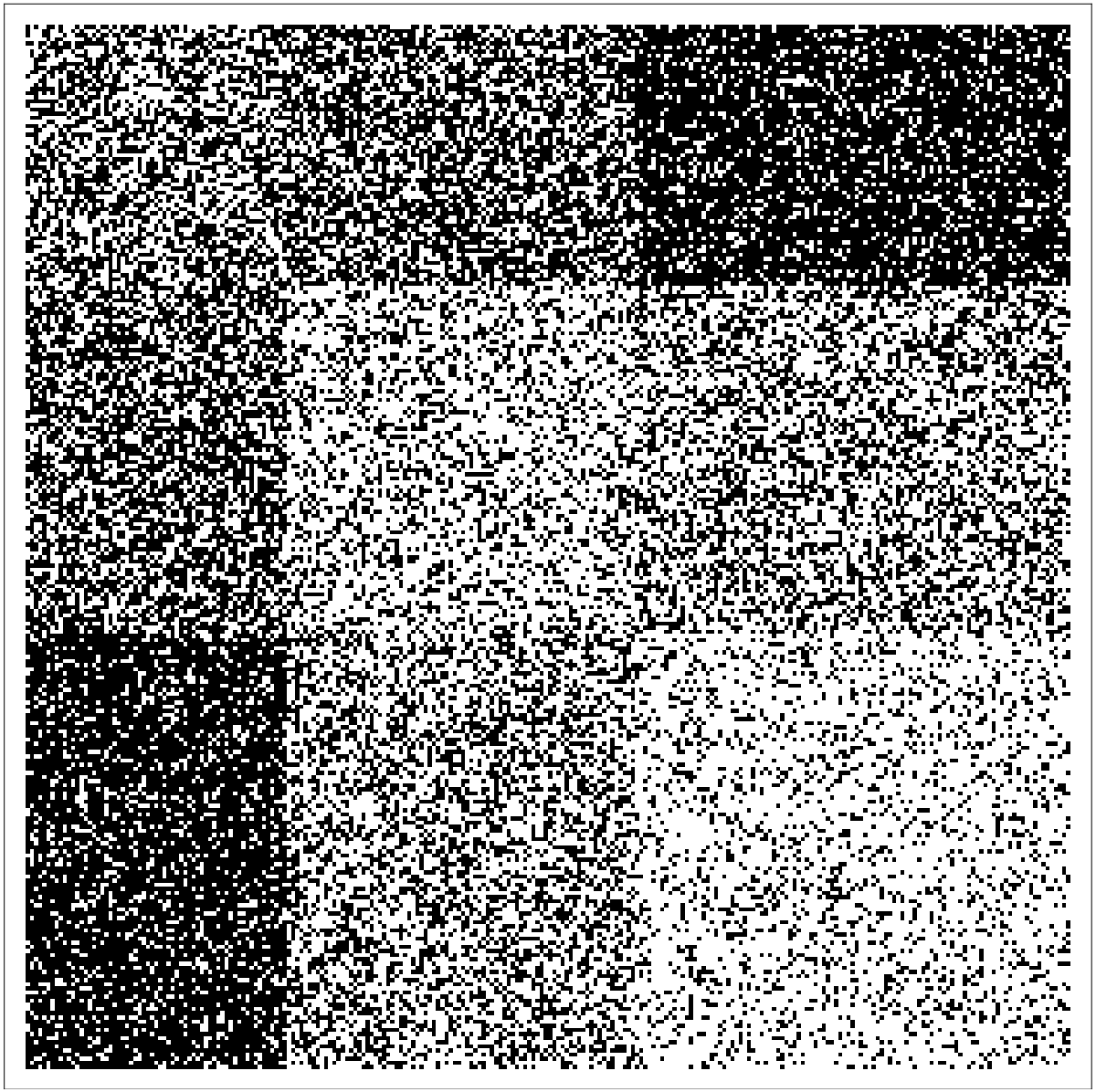
- there are exactly $k-1$ eigenvalues of $\mathbf{B}_D$ greater than $\theta - n^{-\tau}$, while all the others are at most $n^{-\tau}$ in absolute value,
- the k-variance of the vertex representatives constructed by the $k-1$ transformed structural eigenvectors is $\mathcal{O}(n^{-2\tau})$,
- with any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

almost surely.

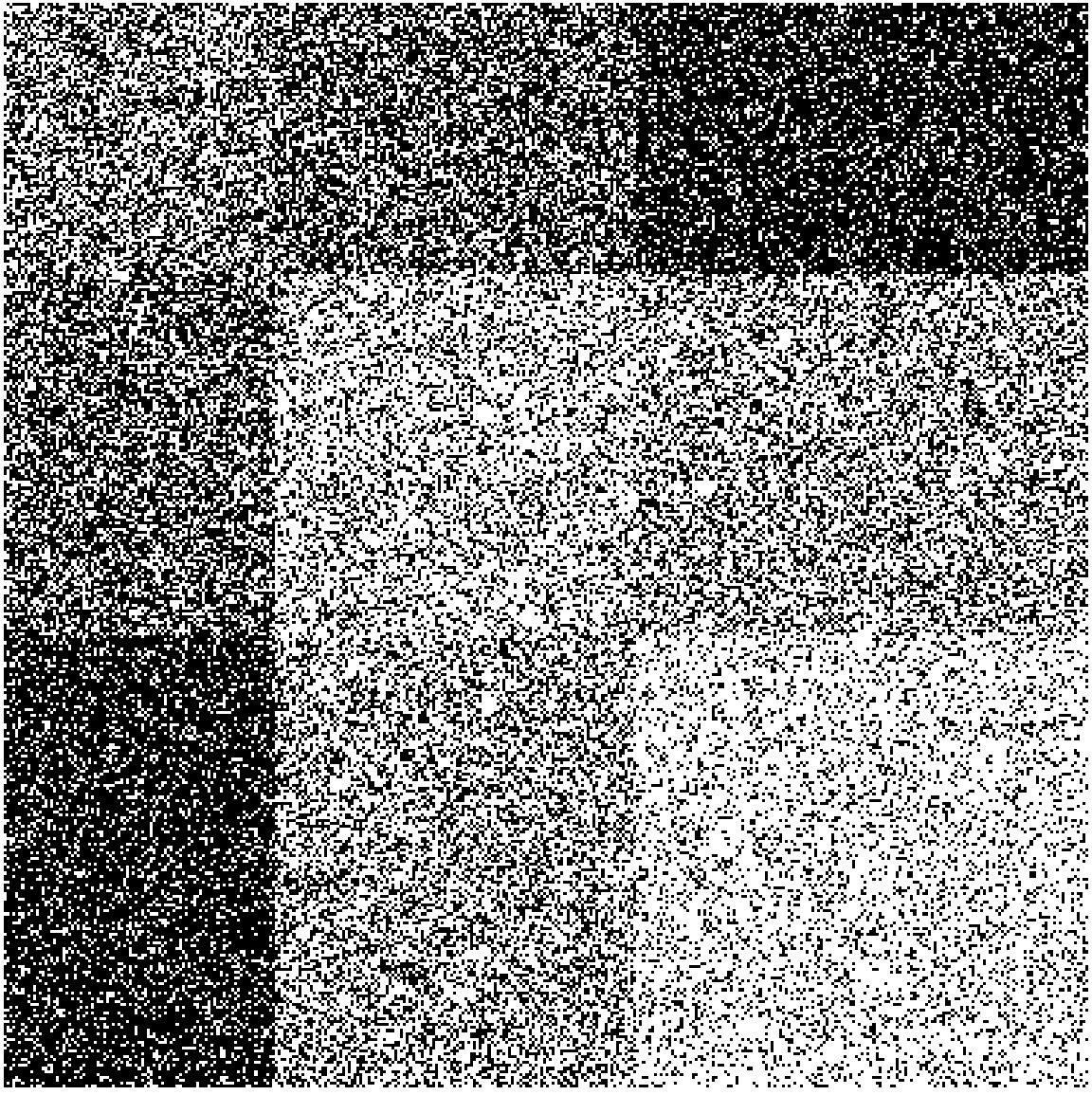
10-fold blow up



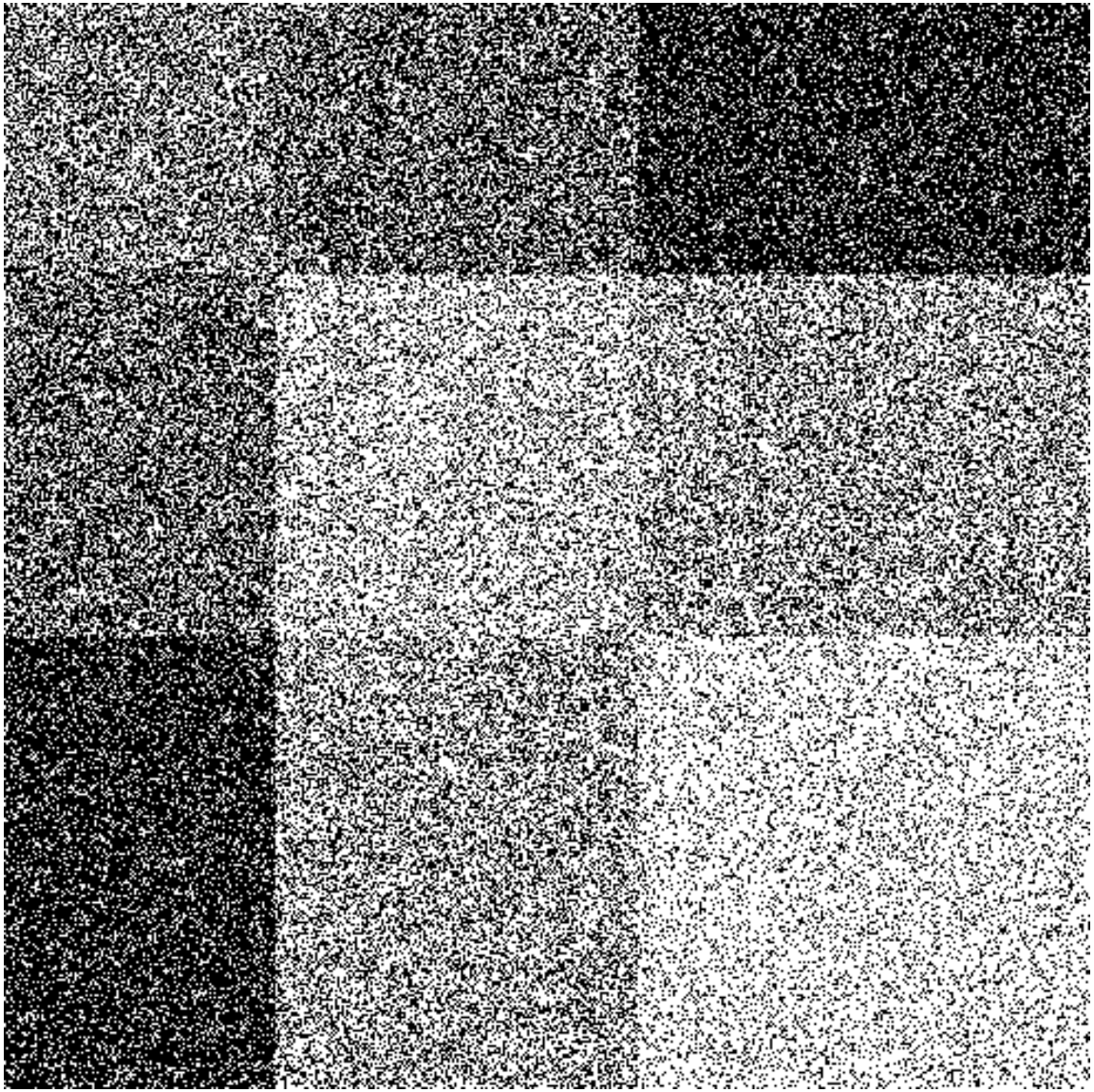
20-fold blow up



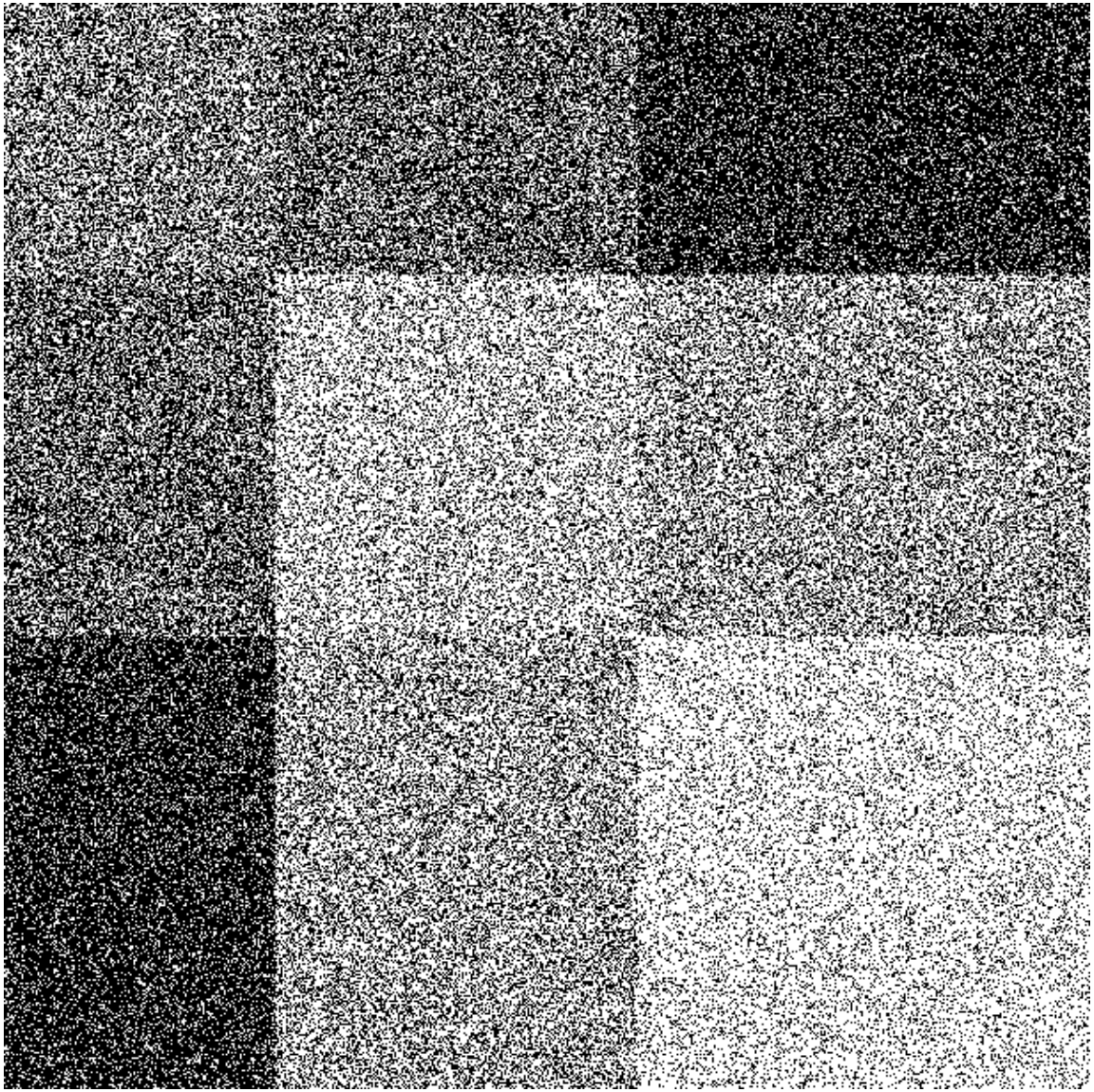
30-fold blow up



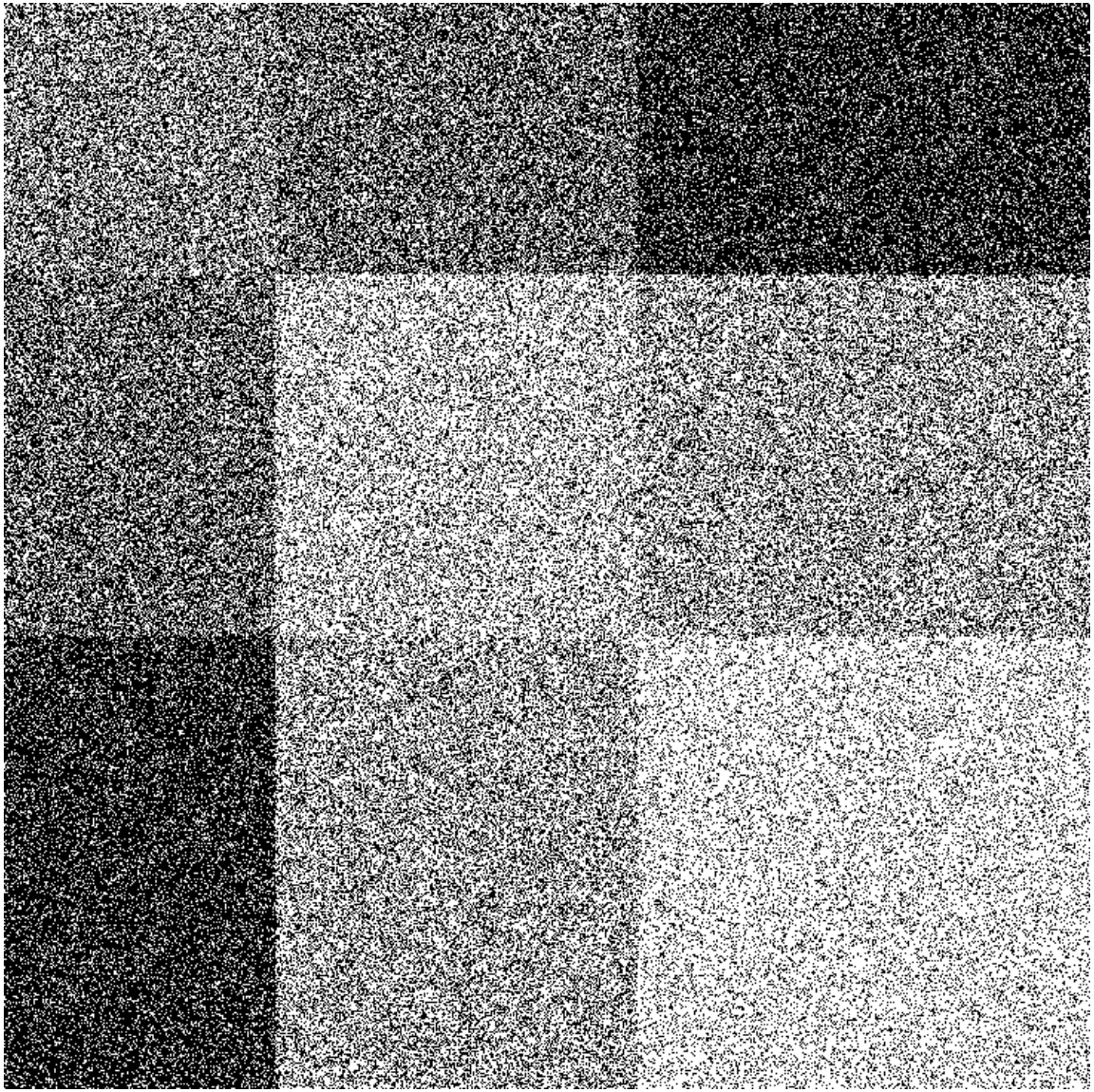
40-fold blow up



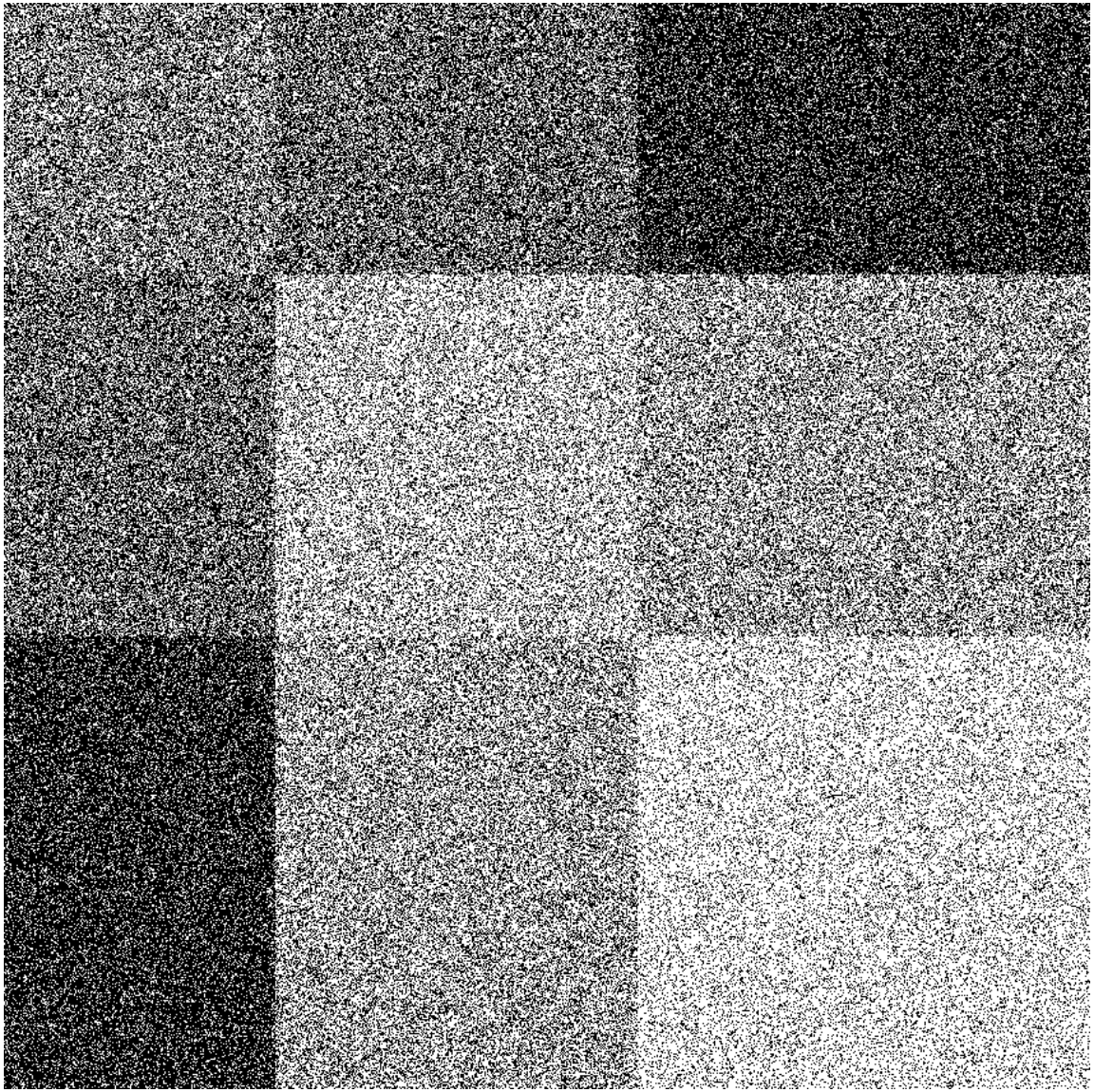
50-fold blow up



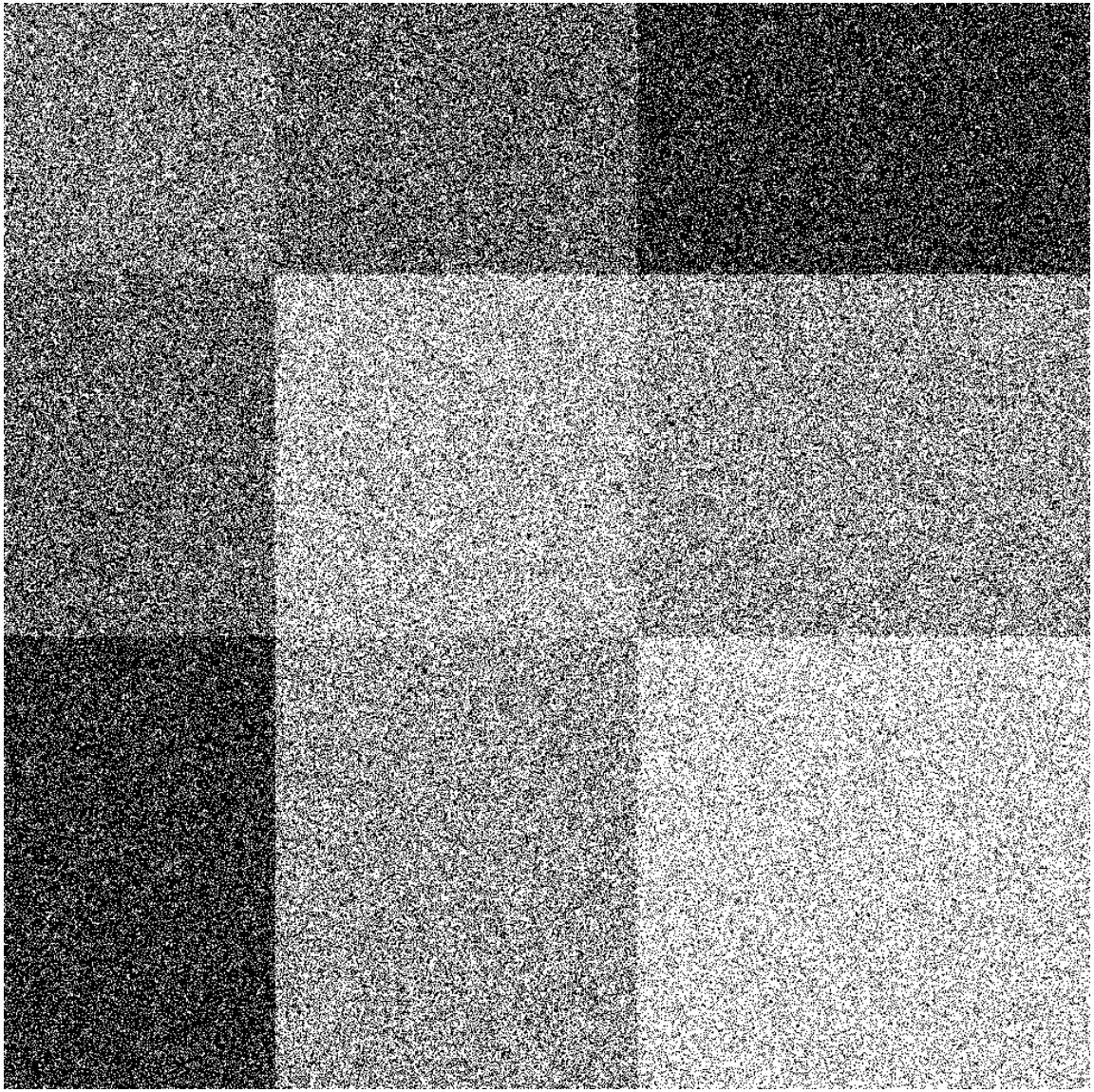
60-fold blow up



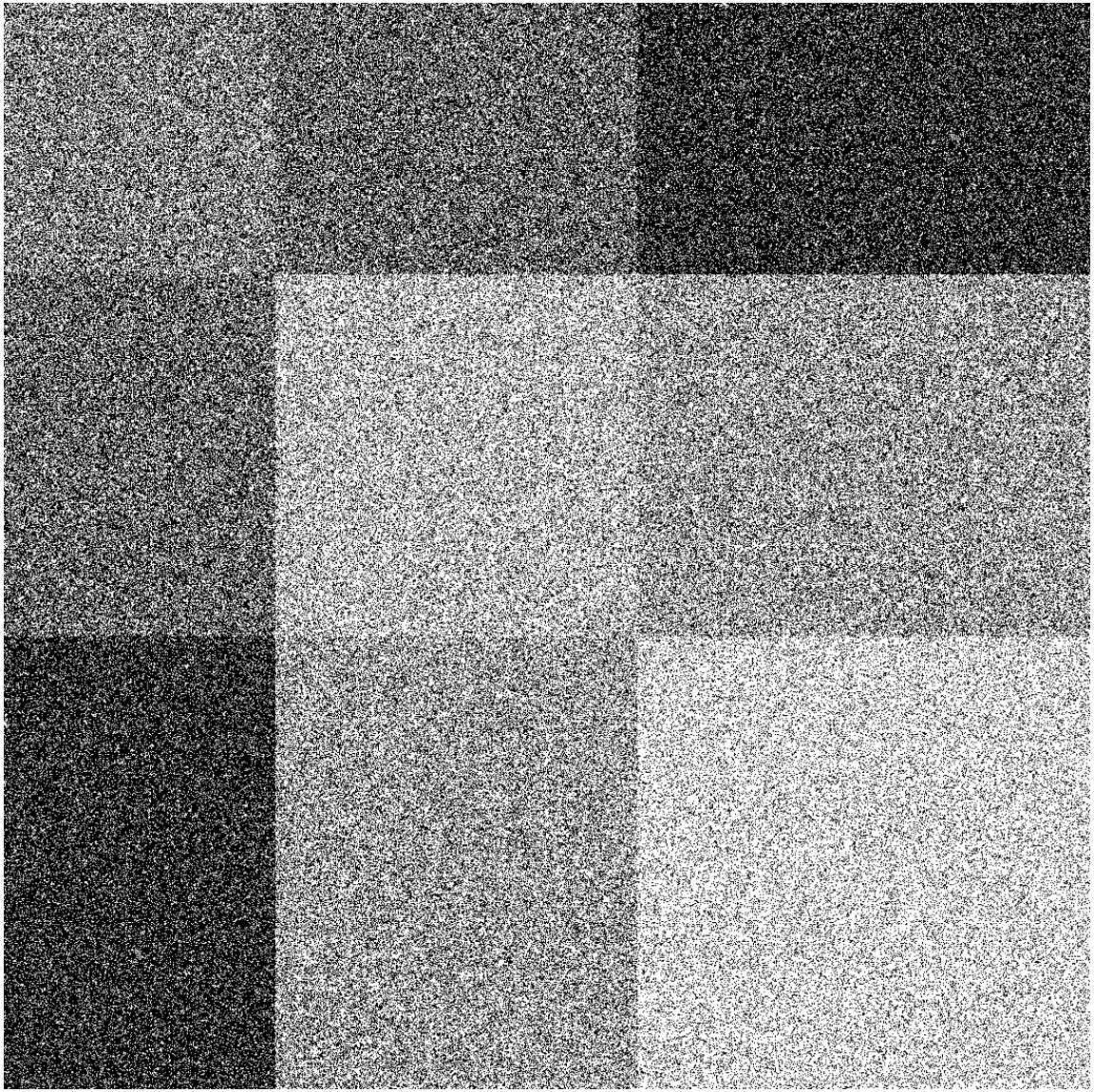
70-fold blow up



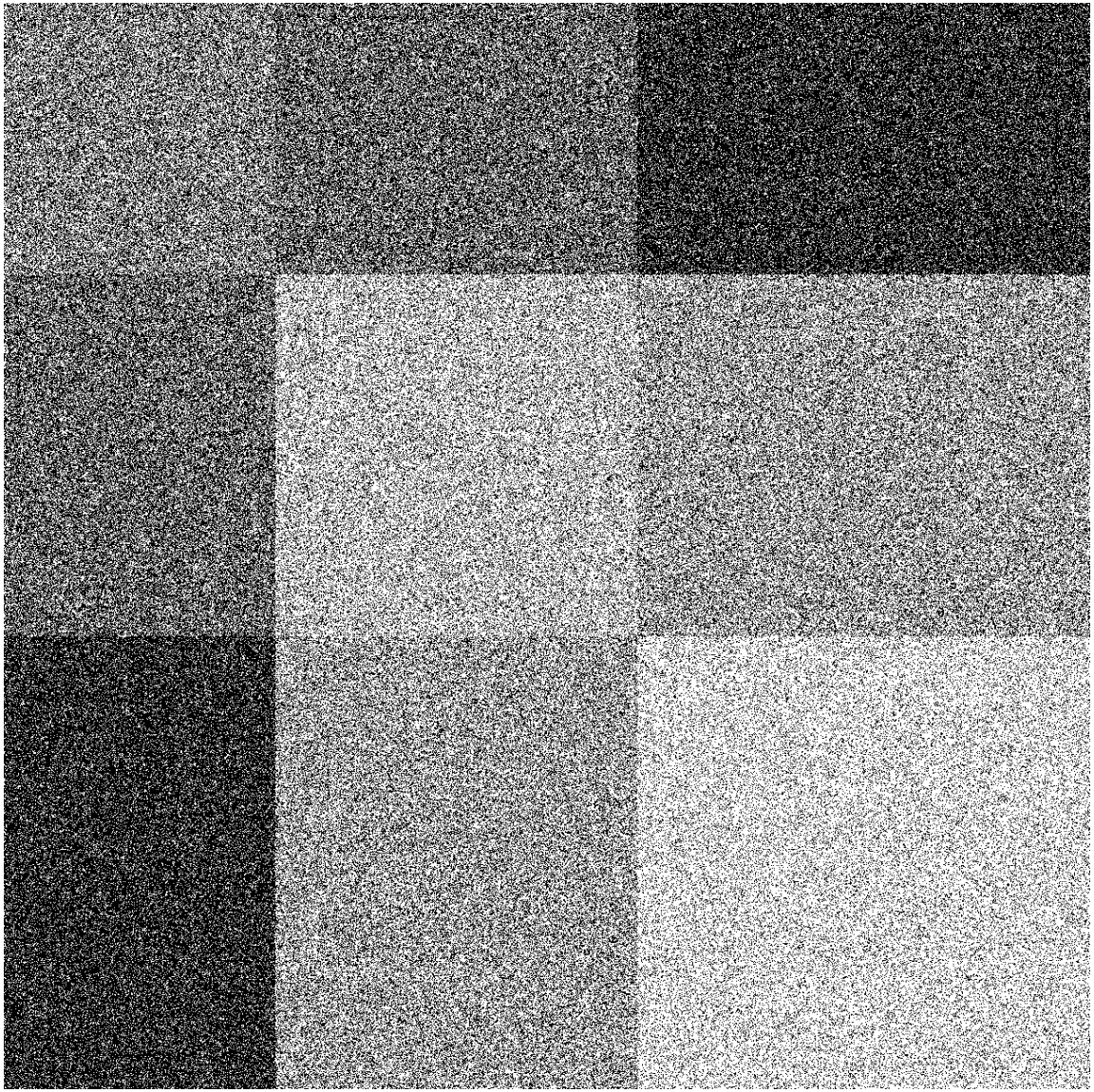
80-fold blow up



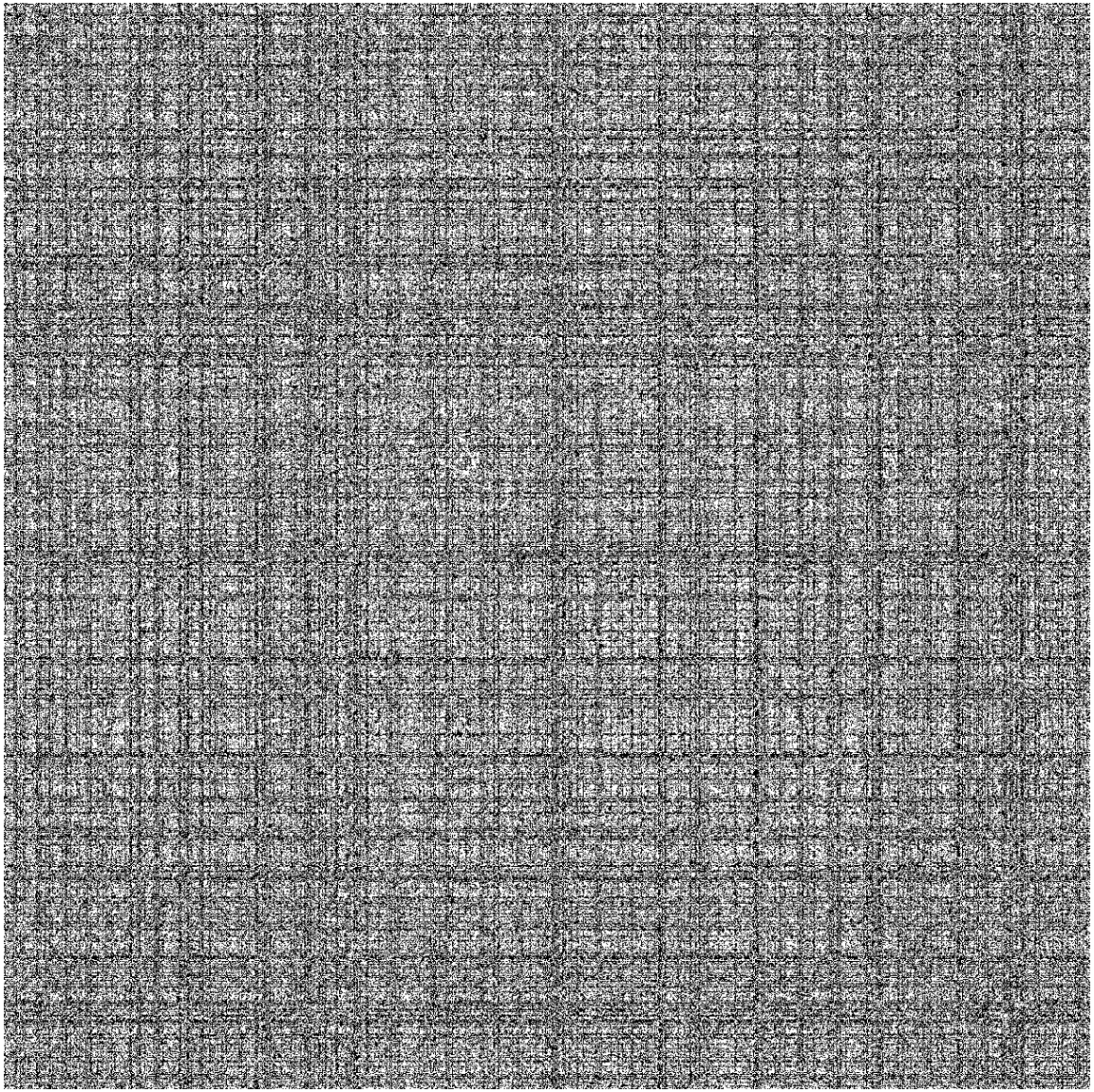
90-fold blow up



100-fold blow up



Before sorting and clustering the vertices



Generalized quasirandom graphs

Generalized quasi-random graph sequences are introduced in Lovász and T. Sós, J. Comb. Theory (2008):
given a model graph graph H on k vertices (both vertices and edges have weights), (G_n) is H-quasirandom if $G_n \longrightarrow W_H$ as $n \longrightarrow \infty$ (left-convergence).
Left-convergence also implies convergence of the spectra $\Longrightarrow$ generalized quasirandom graphs have the same spectral properties as generalized random graphs.

The spectrum itself does not carry enough information for the cluster structure, but together with some classification properties of the eigenvectors it does.

Purpose

For general deterministic edge-weighted graphs we'll prove that the existence of $k-1$ eigenvalues of $\mathbf{B}_D$ 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: $d_i = \Theta(1/n)$, $i = 1, \ldots, n$ as $n \to \infty$. The eigenvalues of $\mathbf{B}_D$ in decreasing absolute values are:

$$(1) > |\rho_2| \geq \cdots \geq |\rho_k| > \varepsilon \geq |\rho_i|, \quad i \geq k+1.$$

The partition $(V_1, \ldots, V_k)$ 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 $|V_i| \geq cn$, $i = 1, \ldots, k$. Then the (V_i, V_j) pairs are $\mathcal{O}(\sqrt{2ks} + \varepsilon)$-volume regular $(i \neq j)$ and for the clusters V_i $(i = 1, \ldots, k)$ the following holds: for all $X, Y \subset V_i$,
$|w(X, Y) - \rho(V_i)\mathtt{Vol}(X)\mathtt{Vol}(Y)| = \mathcal{O}(\sqrt{2ks} + \varepsilon)\mathtt{Vol}(V_i)$,
where $\rho(V_i) = \frac{w(V_i, V_i)}{\mathtt{Vol}^2(V_i)}$ is the relative intra-cluster density of V_i.

Proof

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 $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \sum_{i=1}^{n} \rho_i \mathbf{u}_i \mathbf{u}_i^T$ ($\rho_1 = 1$, $\mathbf{u}_1 = \sqrt{\mathbf{d}}$)
by the rank k matrix $\sum_{i=1}^{k} \rho_i \mathbf{v}_i \mathbf{v}_i^T$ with the following accuracy (in spectral norm):

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

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 $(V_1, \ldots, V_k)$, $i = 1, \ldots, k$

$\Longrightarrow \sum_{i=1}^{k} \rho_i \mathbf{y}_i \mathbf{y}_i^T$ 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 $\mathbf{E}$:

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

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

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 $\mathbf{E}_{ab}$, is estimated as follows:

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

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 $X \subset V_a$, $Y \subset V_b$:

$$
\begin{aligned}
&|w(X,Y) - \rho(V_a, V_b)\mathtt{Vol}(X)\mathtt{Vol}(Y)| = \\
&\left| \sum_{i \in X} \sum_{j \in Y} (d_i d_j \widetilde{w}_{ab} + \eta_{ij}^{ab}) - \frac{\mathtt{Vol}(X)\mathtt{Vol}(Y)}{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)} \sum_{i \in V_a} \sum_{j \in V_b} (d_i d_j \widetilde{w}_{ab} + \eta_{ij}^{ab}) \right| \\
&= \left| \sum_{i \in X} \sum_{j \in Y} \eta_{ij}^{ab} - \frac{\mathtt{Vol}(X)\mathtt{Vol}(Y)}{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)} \sum_{i \in V_a} \sum_{j \in V_b} \eta_{ij}^{ab} \right| \\
&\leq 2C(\sqrt{2k}s + \varepsilon)\sqrt{\mathtt{Vol}(V_a)\mathtt{Vol}(V_b)}
\end{aligned}
$$

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

Remark

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

Under the same conditions and with notations $|\rho_2| = \theta$, $|\rho_3| = \varepsilon$, the (V_1, V_2) pair is $\mathcal{O}\left(\sqrt{\frac{1-\theta}{1-\varepsilon}}\right)$-volume regular.

This also follows from the $k \geq 2$ case, as in B, Tusnády, Discrete Math (1994) we proved that

$$S_2^2(\mathbf{D}^{-1/2}\mathbf{u}_2) = \mathcal{O}\left(\frac{1-\theta}{1-\varepsilon}\right).$$