SPECTRAL CLUSTERING, Lessons 4-5.

Large networks, perturbation of block structures, discrepancy

Marianna Bolla, DSc. Prof. BME Math. Inst.

October 16, 2020

We apply the results of the previous lessons for the spectral clustering of large networks. Networks are modeled either by edge-weighted graphs or contingency tables, and usually subject to random errors due to their evolving and flexible nature. Asymptotic properties of SD and SVD of the involved matrices are discussed when not only the number of the graph's vertices or that of the rows and columns of the contingency table tends to infinity, but the cluster sizes also grow proportionally with them. Mostly, perturbation results for the SD and SVD of blown-up matrices burdened with a Wigner-type error matrix are investigated. Conversely, given a weight-matrix or rectangular array of non-negative entries, we are looking for the underlying block-structure. We will show that under very general circumstances, clusters of a large graph's vertices and simultaneously those of the rows and columns of a large contingency table can be identified with high probability. In this framework, so-called volume regular cluster pairs are also considered with homogeneous information flow within the clusters and between the pairs of them.

1 Symmetric block structures burdened with random noise

First the notion of a very general kind of random noise will be introduced.

Definition 1. Let w_{ij} $(1 \le i \le j \le n)$ be independent, real-valued random variables defined on the same probability space, $w_{ji} = w_{ij}$, $\mathbb{E}(w_{ij}) = 0$ $(\forall i, j)$, and the w_{ij}'s are uniformly bounded, i.e. there is a constant $K > 0$ – that does not depend of n – such that $|w_{ij}| \le K$, $\forall i, j$. Then the $n \times n$ real symmetric random matrix $\boldsymbol{W}_n = (w_{ij})_{1 \le i,j \le n}$ is called symmetric Wigner-noise.

This random matrix is the generalization of that introduced by E. Wigner when formulated his famous semicircle law. Note that the condition of uniform boundedness of the entries could be relaxed: the entries may have Gaussian distribution or sub-Gaussian moments. However, most of the subsequent results and the frequently used Füredi–Komlós theorem rely on Definition 1. This theorem implies that for the spectral norm of an $n \times n$ symmetric Wigner noise

$$\|\boldsymbol{W}_n\| = \max_{1 \le i \le n} |\lambda_i(\boldsymbol{W}_n)| \le 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n) \tag{1}$$

holds with probability tending to 1 as $n \to \infty$, where σ^2 is the uniform bound for the variances of the w_{ij}'s.

Definition 2. The $n \times n$ matrix $\boldsymbol{B}$ is a symmetric blown-up matrix if there is a positive integer $k < n$, a $k \times k$ symmetric pattern matrix $\boldsymbol{P}$ with entries p_{ij} $(0 < p_{ij} < 1)$, and there are positive integers $n_1, \dots, n_k$, $\sum_{i=1}^k n_i = n$ such that – after rearranging its rows and columns in the same way – the matrix $\boldsymbol{B}$ can be divided into k^2 blocks, where the block (i,j) is an $n_i \times n_j$ matrix with entries all equal to p_{ij} $(1 \le i,j \le k)$.

Let us fix $\boldsymbol{P}$, blow it up to an $n \times n$ matrix $\boldsymbol{B}_n$, and consider the noisy matrix $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{W}_n$ as $n_1, \dots, n_k \to \infty$, roughly speaking, at the same rate. For this purpose, an exact growth rate condition is formulated as follows.

Definition 3. Under Growth Rate Condition 1, briefly, **GC1**, the following is understood: $n = \sum_{i=1}^k n_i \to \infty$ in such a way that $\frac{n_i}{n} \ge c$ with some constant $0 < c \le \frac{1}{k}$.

While perturbing $\boldsymbol{B}_n$ by $\boldsymbol{W}_n$, assume that for the uniform bound of the entries of $\boldsymbol{W}_n$ the condition

$$K \le \min\{ \min_{i,j \in \{1,\dots,k\}} p_{ij} \,,\, 1 - \max_{i,j \in \{1,\dots,k\}} p_{ij} \} \tag{2}$$

is satisfied. In this way, the entries of $\boldsymbol{A}_n$ are in the [0,1] interval, and $\boldsymbol{A}_n$ defines an edge-weighted graph $G_n = (V, \boldsymbol{A}_n)$ on n vertices. We are interested in the spectral properties of the expanding random graph sequence G_n. In fact, G_n is the noisy version of the deterministic edge-weighted graph $(V, \boldsymbol{B}_n)$ the vertices of which are partitioned into clusters $V_1, \dots, V_k$ with $|V_a| = n_a$ $(a = 1, \dots k)$ such that vertices of any pair V_a, V_b are connected with an edge of the same weight p_{ab}. Hence, loops are also present.

With an appropriate Wigner-noise we can achieve that the noisy matrix $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{W}_n$ contains 1's in the (a,b)-th block with probability p_{ab}, and 0's otherwise. Indeed, for indices $1 \le a < b \le k$ and $i \in V_a$, $j \in V_b$ let

$$w_{ij} := \begin{cases} 1 - p_{ab} & \text{with probability} \quad p_{ab} \\ -p_{ab} & \text{with probability} \quad 1 - p_{ab} \end{cases} \tag{3}$$

be independent random variables; further, for $a = 1, \dots, k$ and $i,j \in V_a$ $(i \le j)$ let

$$w_{ij} := \begin{cases} 1 - p_{aa} & \text{with probability} \quad p_{aa} \\ -p_{aa} & \text{with probability} \quad 1 - p_{aa} \end{cases} \tag{4}$$

be also independent, otherwise $\boldsymbol{W}$ is symmetric. This $\boldsymbol{W}$ satisfies the conditions of Definition 1 with uniformly bounded entries of zero expectation and variance bounded by

$$\sigma^2 = \max_{1 \le i \le j \le k} p_{ij}(1 - p_{ij}) \le \frac{1}{4}.$$

So, the noisy weighted graph $G_n = (V, \boldsymbol{A}_n)$ becomes a *generalized random graph* on the *planted partition* $V_1, \dots, V_k$ of the vertices such that vertices of V_a and V_b are connected independently, with probability p_{ab}, $1 \le a \le b \le k$; see e.g. [48, 57, 60, 62]. This so-called *stochastic block-model* was first mentioned in [43], and discussed much later in [23] as a special case of an inhomogeneous random graph. [53] show that the best cluster sizes are not very large in a diverse set of real-life networks. This fact motivates the idea to increase the number of clusters with the number of vertices. In the stochastic block-model of [60, 26] this idea is also used. Note that this model is the generalization of the classical Erdős–Rényi random graph (the first random graph of the history, introduced in [35] and also discussed in [19]) which corresponds to the $k = 1$ case.

Definition 4. Let $\boldsymbol{A}_n$ be an $n \times n$ symmetric matrix and $\mathcal{P}_n$ a property which mostly depends on the SD of $\boldsymbol{A}_n$. Then $\boldsymbol{A}_n$ can have the property $\mathcal{P}_n$ in the two following senses. If $\boldsymbol{A}_n$ is the weight matrix of an edge-weighted graph G_n, then we equivalently say that G_n has the property $\mathcal{P}_n$, and denote this fact by $\boldsymbol{A}_n \in \mathcal{P}_n$ or $G_n \in \mathcal{P}_n$.

WP1 The property $\mathcal{P}_n$ holds for $\boldsymbol{A}_n$ *with probability tending to 1* if

$$\lim_{n\to\infty} \mathbb{P}\left(\boldsymbol{A}_n \in \mathcal{P}_n\right) = 1.$$

AS The property $\mathcal{P}_n$ holds for $\boldsymbol{A}_n$ *almost surely* if

$$\mathbb{P}\left(\exists\, n_0 \in \mathbb{N}\, \text{s.t. for}\, n \geq n_0 \;:\; \boldsymbol{A}_n \in \mathcal{P}_n\right) = 1.$$

Here we may assume **GC1** of Definition 3 for the growth of n together with the cluster sizes.

Note that **AS** always implies **WP1**. Conversely, if in addition to **WP1**

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\boldsymbol{A}_n \notin \mathcal{P}_n\right) < \infty$$

also holds, then, by the Borel–Cantelli lemma, $\boldsymbol{A}_n$ has $\mathcal{P}_n$ **AS**.

In view of sharp concentration of the eigenvalues, for the spectral norm, i.e. the largest absolute value eigenvalue, of an $n \times n$ Wigner-noise $\boldsymbol{W}_n$, the following relation holds with every positive real number t:

$$\mathbb{P}\left(|\|\boldsymbol{W}_n\| - \mathbb{E}(\|\boldsymbol{W}_n\|)| > t\right) \leq \exp\left(-\frac{(1-o(1))t^2}{32K^2}\right),$$

where K is the uniform bound for the entries of $\boldsymbol{W}_n$.

This inequality and the fact that, in view of (1), $\|\boldsymbol{W}_n\| = \mathcal{O}(\sqrt{n})$ together imply that $\mathbb{E}\|\boldsymbol{W}_n\| = \mathcal{O}(\sqrt{n})$. Therefore, there exist positive constants c_1 and c_2 such that they do not depend on n (they only depend on K) and

$$\mathbb{P}\left(\, \|\boldsymbol{W}\| > c_1\sqrt{n}\,\right) \leq e^{-c_2 n}. \tag{5}$$

As the right-hand side of (5) is the general term of a convergent series, by the Borel–Cantelli lemma it follows that the spectral norm of $\boldsymbol{W}_n$ is of order $\sqrt{n}$, **AS**. This observation will provide the base of the almost sure results of this chapter. Note that in combinatorics literature, sometimes **WP1** is called **AS**. However, from probabilistic point of view, **AS** is much stronger than **WP1**, and it makes a difference in practice too: an almost sure property means that no matter how $\boldsymbol{A}_n$ is selected, it must have property $\mathcal{P}_n$ if n is large enough.

2 General blown-up structures

The spectrum of a symmetric blown-up matrix is characterized as follows.

Proposition 1. *Under the growth rate condition* ***GC1****, all the non-zero eigenvalues of the* $n \times n$ *blown-up matrix* $\boldsymbol{B}_n$ *of the* $k \times k$ *symmetric pattern matrix* $\boldsymbol{P}$ *are of order* n *in absolute value.*

Proof. As there are at most k linearly independent rows of $\boldsymbol{B}_n$, $r = \text{rank}(\boldsymbol{B}_n) \le k$. Let $\beta_1, \dots, \beta_r > 0$ be the non-zero eigenvalues of $\boldsymbol{B}_n$ with corresponding orthonormal eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_r \in \mathbb{R}^n$. For notational convenience, we discard the subscripts: let $\beta \neq 0$ be an eigenvalue with corresponding eigenvector $\mathbf{u}$, $\|\mathbf{u}\| = 1$. It is easy to see that $\mathbf{u}$ has a piecewise constant structure: it has n_i coordinates equal to $u(i)$ $(i = 1, \dots, k)$, where $n_1, \dots, n_k$ are the blow-up sizes. Then, with these coordinates, the eigenvalue–eigenvector equation

$$\boldsymbol{B}\mathbf{u} = \beta\mathbf{u}$$

has the form

$$\sum_{j=1}^{k} n_j p_{ij} u(j) = \beta u(i), \quad i = 1, \dots, k. \tag{6}$$

With the notation

$$\tilde{\mathbf{u}} = (u(1), \dots, u(k))^T, \qquad \boldsymbol{N} = \text{diag}(n_1, \dots, n_k), \tag{7}$$

(6) can be rewritten in the form

$$\boldsymbol{P}\boldsymbol{N}\tilde{\mathbf{u}} = \beta\tilde{\mathbf{u}}. \tag{8}$$

Further, introducing the transformation

$$\mathbf{v} = \boldsymbol{N}^{1/2}\tilde{\mathbf{u}}, \tag{9}$$

Equation (8) is equivalent to

$$\boldsymbol{N}^{1/2}\boldsymbol{P}\boldsymbol{N}^{1/2}\mathbf{v} = \beta\mathbf{v}. \tag{10}$$

It is easy to see that the transformation (9) results in a unit-norm vector. Furthermore, applying the transformation (9) to the $\tilde{\mathbf{u}}_i$ vectors obtained from the $\mathbf{u}_i$ $(i = 1, \dots, r)$, the orthogonality is also preserved. Consequently, $\mathbf{v}_i = \boldsymbol{N}^{1/2}\tilde{\mathbf{u}}_i$ is an eigenvector corresponding to the eigenvalue β_i of the $k \times k$ matrix $\boldsymbol{N}^{1/2}\boldsymbol{P}\boldsymbol{N}^{1/2}$, $i = 1, \dots, r$. With the shrinking

$$\widetilde{\boldsymbol{N}} = \frac{1}{n}\boldsymbol{N}, \tag{11}$$

(10) is also equivalent to

$$\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{v} = \frac{\beta}{n}\mathbf{v},$$

that is the $k \times k$ matrix $\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}$ has nonzero eigenvalues $\frac{\beta_i}{n}$ with orthonormal eigenvectors $\mathbf{v}_i$ $(i = 1, \dots, r)$.

Now we want to establish relations between the eigenvalues of $\boldsymbol{P}$ and $\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}$. Since we are interested in the absolute values of the nonzero eigenvalues, we will use singular values (recall that the singular values of a symmetric matrix are the absolute values of its real eigenvalues). Also, we are interested only in the first r eigenvalues, where $r = \text{rank}(\boldsymbol{B}) = \text{rank}(\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2})$, therefore, it suffices to consider vectors $\mathbf{x}$, for which $\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x} \neq \mathbf{0}$ and apply the minimax principle. In view of this theorem, for $i \in \{1, \dots, r\}$ and an arbitrary i-dimensional subspace $H \subset \mathbb{R}^n$:

$$\min_{\mathbf{x}\in H} \frac{\|\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\mathbf{x}\|} = \min_{\mathbf{x}\in H} \frac{\|\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|} \cdot \frac{\|\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|} \cdot \frac{\|\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\mathbf{x}\|}$$

$$\geq s_k(\widetilde{\boldsymbol{N}}^{1/2}) \cdot \min_{\mathbf{x}\in H} \frac{\|\boldsymbol{N}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|} \cdot s_k(\widetilde{\boldsymbol{N}}^{1/2}) \geq c \cdot \min_{\mathbf{x}\in H} \frac{\|\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{N}}^{1/2}\mathbf{x}\|},$$

with the constant c of the growth rate condition **GC1** (see Definition 3). Now taking the maximum for all possible i-dimensional subspace H we obtain that $|\lambda_i(\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2})| \geq c|\lambda_i(\boldsymbol{P})| > 0$. On the other hand,

$$|\lambda_i(\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2})| \leq \|\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}\| \leq \|\widetilde{\boldsymbol{N}}^{1/2}\| \cdot \|\boldsymbol{P}\| \cdot \|\widetilde{\boldsymbol{N}}^{1/2}\| \leq \|\boldsymbol{P}\| \leq k.$$

These together imply that $\lambda_i(\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2})$ can be bounded from below and from above with a positive constant that does not depend on n and n_i's, it only depends on c and $\|\boldsymbol{P}\|$. Hence, because of $\lambda_i(\widetilde{\boldsymbol{N}}^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{N}}^{1/2}) = \frac{\beta_i}{n}$, we obtain that $\beta_1, \dots, \beta_r = \Theta(n)$. □

For simplicity, in the sequel, we will assume that $\text{rank}(\boldsymbol{P}) = k$, consequently, $\text{rank}(\boldsymbol{B}) = k$ too.

Theorem 1. *Let $\boldsymbol{B}_n$ be an $n \times n$ blown-up matrix of the $k \times k$ symmetric pattern matrix $\boldsymbol{P}$ with non-zero eigenvalues $\beta_1, \dots, \beta_k$, and $\boldsymbol{W}_n$ be an $n \times n$ Wigner-noise. Then there are k eigenvalues $\lambda_1, \dots, \lambda_k$ of the noisy random matrix $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{W}_n$ such that*

$$|\lambda_i - \beta_i| \leq 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n), \quad i = 1, \dots, k \tag{12}$$

and for the other $n - k$ eigenvalues

$$|\lambda_j| \leq 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n), \quad j = k+1, \dots, n \tag{13}$$

*holds **AS** as $n \to \infty$ under **GC1**.*

Proof. The statement immediately follows by applying the Weyl's perturbation theorem for the spectrum of the symmetric matrix $\boldsymbol{B}_n$ characterized in Proposition 1, where the spectral norm of the perturbation $\boldsymbol{W}_n$ is estimated by (1). This proves the order of eigenvalues **WP1**. In view of (5) and the Borel–Cantelli lemma, it implies that this is an **AS** property as well. □

Consequently, taking into consideration the order $\Theta(n)$ of the non-zero eigenvalues of $\boldsymbol{B}_n$, there is a spectral gap between the k largest absolute value and the other eigenvalues of $\boldsymbol{A}_n$, and this is of order $\Delta - 2\varepsilon$, where

$$\varepsilon = 2\sigma\sqrt{n} + \mathcal{O}(n^{1/3}\log n) \quad \text{and} \quad \Delta = \min_{1\leq i\leq k} |\beta_i|. \tag{14}$$

Then Theorem 1 guarantees the existence of k protruding, so-called *structural* eigenvalues of $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{W}_n$. With the help of this theorem we are also able to estimate the distances between the corresponding eigen-subspaces of the matrices $\boldsymbol{B}_n$ and $\boldsymbol{A}_n$.

Let us denote the unit-norm eigenvectors corresponding to the largest eigenvalues $\beta_1, \dots, \beta_k$ of $\boldsymbol{B}_n$ by $\mathbf{u}_1, \dots, \mathbf{u}_k$ and those corresponding to the largest eigenvalues $\lambda_1, \dots, \lambda_k$ of $\boldsymbol{A}_n$ by $\mathbf{x}_1, \dots, \mathbf{x}_k$. Let $F := \mathrm{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} \subset \mathbb{R}^n$ be the k-dimensional eigen-subspace, and let $\mathrm{dist}(\mathbf{x}, F)$ denote the Euclidean distance between the vector $\mathbf{x} \in \mathbb{R}^n$ and the subspace F.

Proposition 2. *With the above notation, the following estimate holds* $\boldsymbol{AS}$ *for the sum of the squared distances between* $\mathbf{x}_1, \dots, \mathbf{x}_k$ *and* F*:*

$$\sum_{i=1}^{k} \mathrm{dist}^2(\mathbf{x}_i, F) \le k \frac{\varepsilon^2}{(\Delta - \varepsilon)^2} = \mathcal{O}(\frac{1}{n}). \tag{15}$$

Proof. Let us choose one of the eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ of $\boldsymbol{A}_n$ and denote it simply by $\mathbf{x}$ with corresponding eigenvalue λ. To estimate the distance between $\mathbf{x}$ and F, we expand $\mathbf{x}$ in the basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ with coefficients $t_1, \dots, t_n \in \mathbb{R}$:

$$\mathbf{x} = \sum_{i=1}^{n} t_i \mathbf{u}_i.$$

The eigenvalues $\beta_1, \dots, \beta_k$ of the matrix $\boldsymbol{B}_n$ corresponding to $\mathbf{u}_1, \dots, \mathbf{u}_k$ are of order n (by Proposition 1), whereas the other eigenvalues are zeros.

Then, on the one hand

$$\boldsymbol{A}_n \mathbf{x} = (\boldsymbol{B}_n + \boldsymbol{W}_n)\mathbf{x} = \sum_{i=1}^{k} t_i \beta_i \mathbf{u}_i + \boldsymbol{W}_n \mathbf{x}, \tag{16}$$

and on the other hand

$$\boldsymbol{A}_n \mathbf{x} = \lambda \mathbf{x} = \sum_{i=1}^{n} t_i \lambda \mathbf{u}_i. \tag{17}$$

Equating the right-hand sides of (16) and (17) we get that

$$\sum_{i=1}^{k} t_i (\lambda - \beta_i) \mathbf{u}_i + \sum_{i=k+1}^{n} t_i \lambda \mathbf{u}_i = \boldsymbol{W}_n \mathbf{x}.$$

The Pythagorean theorem yields

$$\sum_{i=1}^{k} t_i^2 (\lambda - \beta_i)^2 + \sum_{i=k+1}^{n} t_i^2 \lambda^2 = \|\boldsymbol{W}_n \mathbf{x}\|^2 = \mathbf{x}^T \boldsymbol{W}_n^T \boldsymbol{W}_n \mathbf{x} \le \varepsilon^2, \tag{18}$$

since $\|\mathbf{x}\| = 1$ and the largest eigenvalue of $\boldsymbol{W}_n^T \boldsymbol{W}_n$ is ε^2.

The squared distance between $\mathbf{x}$ and F is $\mathrm{dist}^2(\mathbf{x}, F) = \sum_{i=k+1}^{n} t_i^2$. In view of $|\lambda| \ge \Delta - \varepsilon$,

$$\begin{aligned} (\Delta - \varepsilon)^2 \mathrm{dist}^2(\mathbf{x}, F) = (\Delta - \varepsilon)^2 \sum_{i=k+1}^{n} t_i^2 &\le \sum_{i=k+1}^{n} t_i^2 \lambda^2 \\ &\le \sum_{i=1}^{k} t_i^2 (\lambda - \beta_i)^2 + \sum_{i=k+1}^{n} t_i^2 \lambda^2 \le \varepsilon^2 \end{aligned}$$

where in the last inequality we used (18). From here,

$$\text{dist}^2(\mathbf{x}, F) \leq \frac{\varepsilon^2}{(\Delta - \varepsilon)^2} = \mathcal{O}(\frac{1}{n}) \tag{19}$$

where the order of the estimate follows from the order of ε and Δ of (14).

Applying (19) for the eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ of $\boldsymbol{A}_n$, and adding the k inequalities together, we obtain the same order of magnitude for the sum of the squared distances, which finishes the proof. □

Let $G_n = (V, \boldsymbol{A}_n)$ be the random edge-weighted graph on the n-element vertex set and edge-weights in $\boldsymbol{A}_n$. Denote by $V_1, \dots, V_k$ the partition of V with respect to the blow-up of $\boldsymbol{B}_n$ (it defines a clustering of the vertices). Proposition 2 implies the well-clustering property of the representatives of the vertices of G_n in the following representation. Let $\boldsymbol{X}$ be the $n \times k$ matrix containing the eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ in its columns. Let the k-dimensional representatives of the vertices be the row vectors of $\boldsymbol{X}$ and $S_k^2(P_k; \boldsymbol{X})$ denote the k-variance of these representatives in the clustering $P_k = (V_1, \dots, V_k)$.

Theorem 2. *For the k-variance of the above representation of the noisy weighted graph $G_n = (V, \boldsymbol{A}_n)$, the relation*

$$S_k^2(\boldsymbol{X}) = \mathcal{O}(\frac{1}{n})$$

*holds **AS** as $n \to \infty$ under **GC1**.*

Proof. It is easy to see that $S_k^2(P_k; \boldsymbol{X})$ is equal to the left-hand side of (15), therefore, it is of order $\mathcal{O}(1/n)$. This is also inherited to $S_k^2(\boldsymbol{X}) = \min_{P'_k \in \mathcal{P}_k} S_k^2(P'_k; \boldsymbol{X})$. □

Consequently, the addition of any kind of a Wigner-noise to a weight matrix that has a blown-up structure will not change the order of its protruding eigenvalues, and the block structure of it can be concluded from the vertex representatives of the noisy matrix, where the representation is performed by means of the corresponding eigenvectors.

Hereby, we enlist the eigenvalues and illustrate the clusters of some generalized random graphs based on the following pattern matrices.

- *Type 1*: Edges come into existence with large probability in the diagonal, and small probability in the off-diagonal blocks:

$$k = 3,\ \boldsymbol{P} = \begin{pmatrix} 0.8 & 0.1 & 0.15 \\ 0.1 & 0.75 & 0.2 \\ 0.15 & 0.2 & 0.7 \end{pmatrix},\ \frac{n_1}{n} = 0.30,\ \frac{n_2}{n} = 0.33,\ \frac{n_3}{n} = 0.37.$$

  Figures 1 and 2 show the noisy blown-up matrices with the special noise of (3) and (4), that is the adjacency matrices of generalized random graphs, where black and white squares correspond to the 1 and 0 entries of the adjacency matrix. The largest absolute value eigenvalues of the 1000×1000 normalized modularity matrix are

  $$0.64791,\ 0.50964,\ -0.21923,\ 0.21511,\ -0.20884,\ 0.20768,\ 0.19999$$

  with a gap (in absolute value) after the second one, indicating the existence of 3 underlying clusters. The vertices are sorted according to their cluster memberships, where the clusters are obtained by the k-means algorithm.

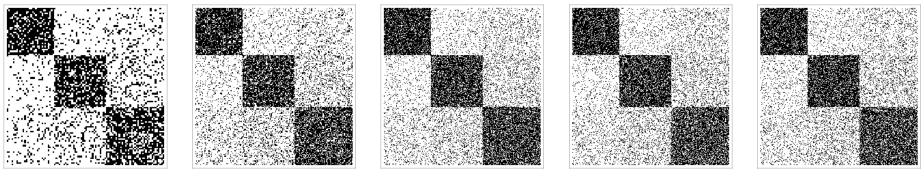

Figure 1: Adjacency matrices of sizes $n \times n$ ($n = 100, 200, 300, 400, 500$) of *Type 1* generalized random graphs, where edges (black) come into existence between blocks i and j with probability p_{ij} (entries of the pattern matrix $\boldsymbol{P}$).

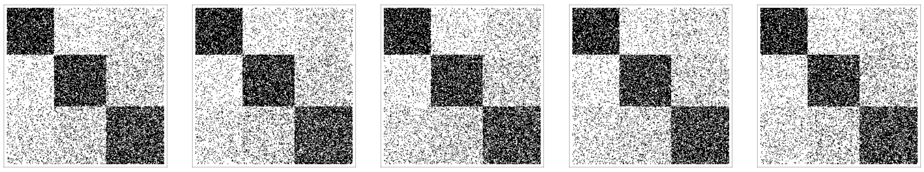

Figure 2: Adjacency matrices of sizes $n \times n$ ($n = 600, 700, 800, 900, 1000$) of *Type 1* generalized random graphs with the same pattern matrix.

- *Type 2*: Edges come into existence with small probability in the diagonal, and large probability in the off-diagonal blocks:

  $$k = 3,\ \boldsymbol{P} = \begin{pmatrix} 0.1 & 0.8 & 0.75 \\ 0.8 & 0.15 & 0.7 \\ 0.75 & 0.7 & 0.2 \end{pmatrix},\ \frac{n_1}{n} = 0.30,\ \frac{n_2}{n} = 0.33,\ \frac{n_3}{n} = 0.37.$$

  Figures 3 and 4 show the noisy blown-up matrices. The largest absolute value eigenvalues of the 1000×1000 normalized modularity matrix are

  $$-0.38458,\ -0.33551,\ 0.06739,\ 0.06672,\ -0.06653,\ -0.06591,\ 0.06502$$

  with a gap (in absolute value) after the second one, indicating the existence of 3 underlying clusters.

- *Type 3*: General pattern matrix of probabilities of inter-cluster connec-

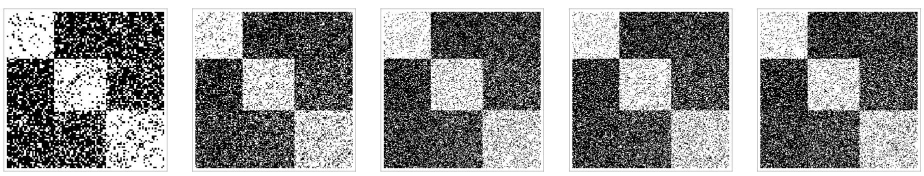

Figure 3: Adjacency matrices of sizes $n \times n$ ($n = 100, 200, 300, 400, 500$) of *Type 2* generalized random graphs, where edges (black) come into existence between blocks i and j with probability p_{ij} (entries of the pattern matrix $\boldsymbol{P}$).

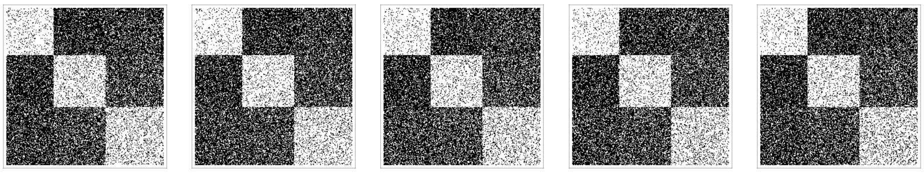

Figure 4: Adjacency matrices of sizes $n \times n$ ($n = 600, 700, 800, 900, 1000$) of *Type 2* generalized random graphs with the same pattern matrix.

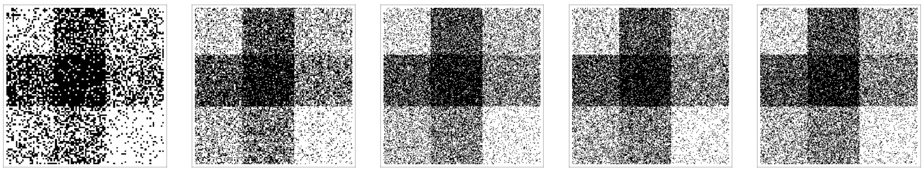

Figure 5: Adjacency matrices of sizes $n \times n$ ($n = 100, 200, 300, 400, 500$) of *Type 3* generalized random graphs, where edges (black) come into existence between blocks i and j with probability p_{ij} (entries of the pattern matrix $\boldsymbol{P}$).

tions:

$$k = 3,\ \boldsymbol{P} = \begin{pmatrix} 0.2 & 0.7 & 0.3 \\ 0.7 & 0.9 & 0.5 \\ 0.3 & 0.5 & 0.1 \end{pmatrix},\ \frac{n_1}{n} = 0.30,\ \frac{n_2}{n} = 0.33,\ \frac{n_3}{n} = 0.37.$$

Figures 5 and 6 show the noisy blown-up matrices. The largest absolute value eigenvalues of the 1000×1000 normalized modularity matrix are

$$-0.16481,\ 0.14511,\ -0.06924,\ 0.06853,\ -0.06810,\ 0.06736,\ -0.06736$$

with a gap (in absolute value) after the second one, indicating the existence of 3 underlying clusters.

3 Recognizing the structure

In the previous section we investigated how the addition of a completely random Wigner-noise influences the behavior of the outstanding, in other words, *structural eigenvalues* of the underlying matrix of a deterministic structure.

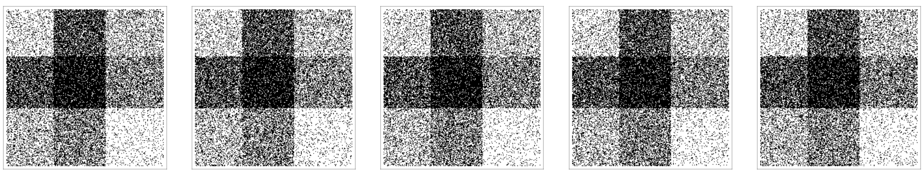

Figure 6: Adjacency matrices of sizes $n \times n$ ($n = 600, 700, 800, 900, 1000$) of *Type 3* generalized random graphs with the same pattern matrix.

Wigner-type matrices became important in quantum mechanics, whereas in case of real-life matrices they are merely landmarks of a random noise added to the underlying linear structure of the edge-weight matrix of some communication, social, or biological networks. Whatever hard is to recognize the structure concealed by the noise, in a number of models it is possible by means of spectral techniques and large deviations principles. When we perform spectral clustering, it is a crucial question how many structural eigenvalues – with corresponding eigenvectors – to retain for the vertex-representation.

Note that numerical algorithms for the SD of a matrix with size of millions are not immediately applicable, and some newly developed randomized algorithms are to be used instead, e.g. [1]. These algorithms exploit the randomness of the underlying matrix, and rely on the fact that a random noise will not change the order of magnitude of the structural eigenvalues. Sometimes, instead of depriving the matrix of the noise, rather a noise is added (by digitalizing the entries of or sparsifying the underlying matrix with an appropriate randomization) to make the matrix more easily decomposable by means of the classical methods (e.g. the Lánczos method, see [39]). These algorithms are only capable to find a low-rank approximation of our matrix, where this rank cannot exceed the number of the structural eigenvalues of the original matrix. If there are no such eigenvalues, this property is also inherited to the randomized matrix, so the worst that can happen: we get known of the sad fact that there is no linear structure in our matrix at all, but it is a noise itself. In all the other cases we obtain a good approximation for the part of the spectrum needed, exploiting the randomness in our original data.

Both the number of eigenvalues to be retained and algorithmic questions can be analyzed by means of the results of this section. Also note that Wigner-type noises over a remarkable structure are not only numerically tractable, but have significance in real-life networks too. For example, sociologist [40] shows that sometimes week ties better help people to find job than strong (historical or family) relations in which they are stuck.

We investigate the opposite question too: what kind of random matrices have a blown-up matrix as a skeleton, except of a 'small' perturbation? The following proposition states that under very general conditions an $n \times n$ random symmetric matrix with nonnegative, uniformly bounded entries (so that it can be the weight matrix of an edge-weighted graph) has at least one eigenvalue greater than of order $\sqrt{n}$.

Proposition 3. *Let $\boldsymbol{A}$ be an $n \times n$ random symmetric matrix such that $0 \le a_{ij} \le 1$ and the entries are independent for $i \le j$. Further, let us assume that there are positive constants c_1 and c_2 and $0 < \delta \le \Delta \le 1/2$ such that, with the notation $X_i = \sum_{j=1}^{n} a_{ij}$,*

$$\mathbb{E}(X_i) \ge c_1 n^{\frac{1}{2}+\delta} \quad \textit{and} \quad \mathrm{Var}(X_i) \le c_2 n^{\frac{1}{2}+\Delta}, \quad i = 1, \dots, n.$$

Then for every $0 < \varepsilon < \delta$:

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\max}(\boldsymbol{A}) \ge c_1 n^{\frac{1}{2}+\varepsilon}\right) = 1.$$

Remark that the above conditions automatically hold true if there is a constant $0 < \mu_0 < 1$ such that $\mathbb{E}(a_{ij}) \ge \mu_0$ for all i, j pairs. This is the case in the theorems of [45] and [38]. In our case there can be a lot of zero entries, we require only that in each row there are at least $c_1 n^{1/2+\delta}$ entries with expectation greater than or equal to any small fixed positive constant μ_0. As the matrix is symmetric, this also holds for its columns. Therefore, among the n^2 entries there must be at least $\Theta(n^{1+2\delta})$ ones (but not anyhow) with expectation at least a fixed $0 < \mu_0 < 1$, all the others can be zeros. Also note that by the Perron–Frobenius theory, the largest eigenvalue of $\boldsymbol{A}$ is always positive.

To prove Proposition 3, the following lemma is needed (see e.g. [42]).

Lemma 1 (Chernoff inequality for large deviations)**.** *Let $X_1, \dots, X_n$ be independent random variables, $|X_i| \le K$, $X := \sum_{i=1}^n X_i$. Then for every $a > 0$:*

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > a\right) \le e^{-\frac{a^2}{2(\mathrm{Var}(X)+Ka/3)}} .$$

Proposition *3.* With some linear algebra, $\lambda_{\max}(\boldsymbol{A}) \ge \min_i X_i$, hence

$$\mathbb{P}\left(\lambda_{\max}(\boldsymbol{A}) \ge c_1 n^{\frac{1}{2}+\varepsilon}\right) \ge \boldsymbol{P}\left(\min_i X_i \ge c_1 n^{\frac{1}{2}+\varepsilon}\right),$$

and it is enough to prove that the latter probability tends to 1 as $n \to \infty$. We will prove that the probability of the complementary event tends to 0:

$$\mathbb{P}\left(\text{for at least one } i: \; X_i < c_1 n^{\frac{1}{2}+\varepsilon}\right) \le n\boldsymbol{P}\left(\text{for a general } i: \; X_i < c_1 n^{\frac{1}{2}+\varepsilon}\right). \tag{20}$$

From now on, we will drop the suffix i and X denotes the sum of the entries in an arbitrary row of $\boldsymbol{A}$. As X is the sum of n independent random variables satisfying the conditions of Lemma 1 with $K = 1$,

$$\begin{aligned}
\mathbb{P}\left(X < c_1 n^{\frac{1}{2}+\varepsilon}\right) &= \mathbb{P}\left(\mathbb{E}(X) - X > \mathbb{E}(X) - c_1 n^{\frac{1}{2}+\varepsilon}\right) \\
&\le \mathbb{P}\left(|X - \mathbb{E}(X)| > \mathbb{E}(X) - c_1 n^{\frac{1}{2}+\varepsilon}\right) \\
&\le \mathbb{P}\left(|X - \mathbb{E}(X| > c_1 n^{\frac{1}{2}}(n^{\delta} - n^{\varepsilon})\right) \\
&\le e^{-\frac{c_1^2 n (n^{\delta}-n^{\varepsilon})^2}{2(c_2 n^{\frac{1}{2}+\Delta} + n^{\frac{1}{2}}(n^{\delta}-n^{\varepsilon})/3)}} \\
&\le e^{-c_3 n^{\frac{1}{2}} \frac{(n^{\delta}-n^{\varepsilon})^2}{n^{\Delta}}} \\
&= e^{-c_3 n^{\frac{1}{2}-\Delta}(n^{\delta}-n^{\varepsilon})^2}
\end{aligned}$$

with some positive constant c_3, in view of the inequalities $0 < \varepsilon < \delta \le \Delta \le 1/2$. Thus, the right-hand side of (20) can be estimated from above by

$$\frac{n}{e^{c_3 n^{\frac{1}{2}-\Delta}(n^{\delta}-n^{\varepsilon})^2}} \le \frac{n}{e^{c_4 n^{\gamma}}}$$

with some constants $c_4 > 0$ and $\gamma > 0$, because of $0 < \varepsilon < \delta \le \Delta \le 1/2$. The last term above tends to 0 as $n \to \infty$, that finishes the proof. □

Note that the constants δ and Δ were only responsible for the speed of the convergence. Now we will use Proposition 3 to deprive a random symmetric matrix of the noise.

Theorem 3. *Let $\boldsymbol{A}_n$ be a sequence of $n \times n$ symmetric matrices, where n tends to infinity. Assume that $\boldsymbol{A}_n$ has exactly k eigenvalues of order greater than $\sqrt{n}$ (k is fixed), and there is a k-partition of the vertices of $G_n = (V, \boldsymbol{A}_n)$ such that the k-variance of the representatives – in the representation with the corresponding eigenvectors – is $\mathcal{O}(1/n)$. Then there is an explicit construction for a symmetric blown-up matrix $\boldsymbol{B}_n$ such that $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{E}_{m\times n}$, with $\|\boldsymbol{E}_n\| = \mathcal{O}(\sqrt{n})$.*

Instead of the complete proof we only describe the construction, since the estimations are special cases of those performed more generally, for rectangular matrices soon. However, we now prove the following lemma which will frequently be used in the sequel.

Lemma 2. *Let $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{y}_1, \dots, \mathbf{y}_k \in \mathbb{R}^n$ be orthonormal sets ($k \le n$). Then another orthonormal set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ within $F = \mathrm{Span}\{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ can be found such that*

$$\sum_{i=1}^{k} \|\mathbf{x}_i - \mathbf{v}_i\|^2 \le 2 \sum_{i=1}^{k} \mathrm{dist}^2(\mathbf{x}_i, F).$$

Proof. In the proof we follow the ideas of [10]. We will give a construction for such $\mathbf{v}_i$'s which are 'close' to the individual $\mathbf{x}_i$'s, respectively. Note that $\boldsymbol{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ and $\boldsymbol{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$ are $n \times k$ suborthogonal matrices. Since the vectors $\mathbf{v}_i$'s also form an orthonormal set within F, they can be obtained by applying a rotation (within F) for the $\mathbf{y}_i$'s. That is, we are looking for a $k \times k$ orthogonal matrix $\boldsymbol{R}$ such that $(\mathbf{v}_1, \dots, \mathbf{v}_k) = \boldsymbol{Y}\boldsymbol{R}$ and, with it,

$$\sum_{i=1}^{k} \|\mathbf{x}_i - \mathbf{v}_i\|^2 = \mathrm{tr}[(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{R})^T(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{R})] \le 2 \sum_{i=1}^{k} d^2(\mathbf{x}_i, F) \tag{21}$$

holds. By the properties of the trace operator,

$$\begin{aligned} \mathrm{tr}[(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{R})^T(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{R})] &= \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}) + \mathrm{tr}(\boldsymbol{R}^T\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{R}) - 2\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{R}) \\ &= \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}) + \mathrm{tr}[(\boldsymbol{Y}^T\boldsymbol{Y})(\boldsymbol{R}\boldsymbol{R}^T)] - 2\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{R}) \\ &= 2[k - \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{R})] \end{aligned} \tag{22}$$

is obtained, where we used that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}_k$. The expression in (22) is minimal if and only if $\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{R})$ is maximal as a function of $\boldsymbol{R}$. At this point, we use a linear algebra fact, according to which $\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y})\boldsymbol{R}$ is maximal if $(\boldsymbol{X}^T\boldsymbol{Y})\boldsymbol{R}$ is symmetric, and the maximum is $\sum_{i=1}^k s_i$, with s_i's being the singular values of $\boldsymbol{X}^T\boldsymbol{Y}$. Indeed, let $\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^T$ the SVD of $\boldsymbol{X}^T\boldsymbol{Y}$, where $\boldsymbol{V}$ and $\boldsymbol{U}$ are $k \times k$ orthogonal matrices and $\boldsymbol{S}$ is $k \times k$ diagonal matrix with the singular values in its main diagonal. Note that the singular values s_i's are the cosines of the *principal (canonical) angles* between the subspaces $\mathrm{Span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ and F.

By the above SVD, $(\boldsymbol{X}^T\boldsymbol{Y})\boldsymbol{R} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^T\boldsymbol{R}$ is symmetric if $\boldsymbol{U}^T\boldsymbol{R} = \boldsymbol{V}^T$, i.e. $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{V}^T$. Consequently, the minimum that can be attained in (22) is equal to

$$2\sum_{i=1}^{k}(1 - s_i). \tag{23}$$

Eventually, the sum of the distances in (21) can also be written in terms of the singular values $s_1, \dots, s_k$. Since $\boldsymbol{Y}\boldsymbol{Y}^T$ is the matrix of the orthogonal projection onto F,

$$\sum_{i=1}^{k} d^2(\mathbf{x}_i, F) = \mathrm{tr}[(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})^T(\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})] = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}) - \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})$$
$$= k - \sum_{i=1}^{k} s_i^2 = \sum_{i=1}^{k}(1 - s_i^2), \tag{24}$$

where we also used that the $k \times k$ symmetric matrix $\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$ has eigenvalues $s_1^2, \dots, s_k^2$.

Comparing (23) and (24), it remains to show that $\sum_{i=1}^{k}(1 - s_i) \le \sum_{i=1}^{k}(1 - s_i^2)$. But s_i's are the singular values of the matrix $\boldsymbol{X}^T\boldsymbol{Y}$, therefore denoting by $s_{\max}(.)$ the maximum singular value of the matrix in the argument, we have

$$s_i \le s_{\max}(\boldsymbol{X}^T\boldsymbol{Y}) \le s_{\max}(\boldsymbol{X}) \cdot s_{\max}(\boldsymbol{Y}) = 1,$$

as all positive singular values of the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ – being suborthogonal matrices – are equal to 1. Hence, $s_i \ge s_i^2$ implies the desired relation (21). □

***Theorem** 3, the construction part.* Let $\mathbf{x}_1, \dots, \mathbf{x}_k$ denote the eigenvectors corresponding to $\lambda_1, \dots, \lambda_k$, the k largest absolute value (of order larger than $\sqrt{n}$) eigenvalues of $\boldsymbol{A}_n$. The representatives – that are row vectors of the $n \times k$ matrix $\boldsymbol{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ – by the assumption of the theorem, form k clusters in $\mathbb{R}^k$ with k-variance less than c/n with some constant c. Let $V_1, \dots, V_k$ denote the clusters (properly reordering the rows of $\boldsymbol{X}$, together they give the index set $\{1, \dots, n\}$). Let $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathbb{R}^k$ be the Euclidean representatives of the vertices (the rows of $\boldsymbol{X}$), and let $\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^k$ denote the cluster centers. Now let us choose the following representation of the vertices. The representatives are row vectors of the $n \times k$ matrix $\tilde{\boldsymbol{X}}$ such that the first n_1 rows of $\tilde{\boldsymbol{X}}$ are equal to $\bar{\mathbf{x}}^1$, the next n_2 rows of $\tilde{\boldsymbol{X}}$ are equal to $\bar{\mathbf{x}}^2$, …, and so on; the last n_k rows of $\tilde{\boldsymbol{X}}$ are equal to $\bar{\mathbf{x}}^k$. Finally, let $\mathbf{y}_1, \dots, \mathbf{y}_k \in \mathbb{R}^n$ be the column vectors of $\tilde{\boldsymbol{X}}$. By the considerations of the Proof of Theorem 2,

$$S_k^2(\boldsymbol{X}) = \sum_{i=1}^{k} \mathrm{dist}^2(\mathbf{x}_i, F),$$

where the k-dimensional subspace F is spanned by the vectors $\mathbf{y}_1, \dots, \mathbf{y}_k$; further, by the assumption of the theorem, $S_k^2(\boldsymbol{X}) < \frac{c}{n}$.

Then, in view of Lemma 2, a set $\mathbf{v}_1, \dots, \mathbf{v}_k$ of orthonormal vectors within F can be found such that

$$\sum_{i=1}^{k} \|\mathbf{x}_i - \mathbf{v}_i\|^2 \le 2\frac{c}{n}$$

holds **AS**. It is important that $\mathbf{v}_i$'s are also piecewise constant vectors over the vertex-clusters $V_1, \dots, V_k$.

Finally, for the matrix $\boldsymbol{A}_n = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^T$, the blown-up matrix $\boldsymbol{B}_n = \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ is constructed. Then the spectral norm of the error matrix $\boldsymbol{E}_n = \boldsymbol{A}_n - \boldsymbol{B}_n$ is $\mathcal{O}(\sqrt{n})$, as it will be proved more generally, in the rectangular case. □

In the models discussed in Subsection 2, a very general kind of a random noise was added to the weight matrix of some special edge-weighted graphs. We have shown that if the weight matrix has some structural eigenvalues, then a Wigner-noise cannot essentially disturb this structure: the adjacency matrix of the noisy graph will have the same number of protruding eigenvalues with corresponding eigenvectors revealing the structure of the graph. Vice versa, if – in addition – the representation with the eigenvectors corresponding to the structural eigenvalues shows well classification properties, we have shown how to find the clusters themselves. Now we will briefly discuss some other kind of contemporary random graph models from the point of view of stability. [46] investigates the possibly complex eigenvalues of not necessarily symmetric random block matrices under more special conditions on the entries.

Theoretically, for any graph on n vertices, the Regularity Lemma of Szemerédi (see [51, 61, 64, 65]) guarantees the existence of a favorable partition of the vertices with maximum number of clusters independent of n. Here we cite the Regularity Lemma together with the notion of favorable cluster pairs between which the edge-density is homogeneous.

Definition 5. We say that the disjoint pair $V_i, V_j \subset V$ $(i \neq j)$ is ε-regular, if for any $A \subset V_i$, $B \subset V_j$, with $|A| > \varepsilon|V_i|$, $|B| > \varepsilon|V_j|$,

$$|\rho(A,B) - \rho(V_i,V_j)| < \varepsilon$$

holds, where $\rho(A,B)$ denotes the edge-density between the disjoint vertex-subsets A and B. More precisely, denoting by $e(A,B)$ the number of cut-edges between A and B,

$$\rho(A,B) = \frac{e(A,B)}{|A| \cdot |B|}.$$

Theorem 4 (Szemerédi's Regularity Lemma)**.** *For every $\varepsilon > 0$ and integer $m > 0$ there are integers $P(\varepsilon,m)$, $Q(\varepsilon,m)$ with the following property: for every simple graph $G = (V,E)$ with $n > P(\varepsilon,m)$ vertices there is a partition of V into $k+1$ classes $V_0, V_1, \dots, V_k$ such that*

- $m \leq k \leq Q(\varepsilon,m)$,
- $|V_0| \leq \varepsilon n$,
- $|V_1| = |V_2| = \cdots = |V_k|$,
- *all but at most εk^2 of the pairs (V_i, V_j) are ε-regular.*

If the graph is sparse – the number of edges $e = o(n^2)$ – then $k = 1$, otherwise k can be arbitrarily large (but it does not depend on n, it merely depends on ε).

If our random graph is a generalized random graph, then $e(V_i, V_j)$ is the sum of $|V_i| \cdot |V_j|$ independent, identically distributed Bernoulli variables with parameter p_{ij} $(1 \leq i,j \leq k)$, where p_{ij}'s are entries of the pattern matrix $\boldsymbol{P}$. Hence $e(A,B)$ is a binomially distributed random variable with expectation $|A| \cdot |B| \cdot p_{ij}$ and variance $|A| \cdot |B| \cdot p_{ij}(1-p_{ij})$. Therefore, by Lemma 1 (with the choice of $K = 1$) and with $A \subset V_i$, $B \subset V_j$, $|A| > \varepsilon|V_i|$, $|B| > \varepsilon|V_j|$ we have that

$$
\begin{aligned}
\mathbb{P}\left(|\rho(A,B)-p_{ij}|>\varepsilon\right) &= \mathbb{P}\left(|e(A,B)-|A|\cdot|B|\cdot p_{ij}|>\varepsilon\cdot|A|\cdot|B|\right)\\
&\le e^{-\frac{\varepsilon^2|A|^2|B|^2}{2[|A||B|p_{ij}(1-p_{ij})+\varepsilon|A||B|/3]}}\\
&= e^{-\frac{\varepsilon^2|A||B|}{2[p_{ij}(1-p_{ij})+\varepsilon/3]}}\\
&\le e^{-\frac{\varepsilon^4|V_i||V_j|}{2[p_{ij}(1-p_{ij})+\varepsilon/3]}},
\end{aligned}
$$

that tends to 0 as $|V_i| = n_i \to \infty$ and $|V_j| = n_j \to \infty$. Hence, any pair V_i, V_j is **AS** ε-regular. We note, however, that the Regularity Lemma does not give a construction for the clusters. Provided the conditions of Theorem 3 hold, by the cluster centers a similar construction may exist. The algorithmic aspects of the Regularity Lemma are discussed in [4]. A matrix approximation theorem to derive a constructive version of the Regularity Lemma is introduced in [37]. Further aspects will be discussed in Section 5.

4 Noisy contingency tables

Contingency tables are rectangular arrays of nonnegative real entries, corresponding to standardized counts of two categorical variables, e.g. keyword–document matrices or microarrays. In this section, the results of the previous section will be extended to the stability of the SVD of large noisy contingency tables, further to two-way clustering of the row and column categories of the underlying categorical variables. As the categories may be measured in different units, sometimes a normalization is necessary. For this reason, normalized contingency tables and Correspondence Analysis techniques are used. To begin with, the notions of the previous section are adopted to rectangular arrays.

Definition 6. The $m \times n$ real matrix $\boldsymbol{W}$ is a Wigner-noise if its entries w_{ij} ($1 \le i \le m$, $1 \le j \le n$) are independent random variables, $\mathbb{E}(w_{ij}) = 0$, and the w_{ij}'s are uniformly bounded (i.e. there is a constant $K > 0$, independently of m and n, such that $|w_{ij}| \le K$, $\forall i, j$).

Though, the main results of this paper can be extended to w_{ij}'s with any light-tail distribution (especially to Gaussian distributed entries), the subsequent results will be based on the assumptions of Definition 6, sometimes also using the notation σ^2 for the common bound of the entries' variances. Akin to the quadratic case, the spectral norm of an $m \times n$ Wigner-noise $\boldsymbol{W}_{m\times n}$ is at most of order $\sqrt{m+n}$, with probability tending to 1 as $n, m \to \infty$.

Definition 7. The $m \times n$ real matrix $\boldsymbol{B}$ is a blown-up matrix if there is an $a \times b$ so-called pattern matrix $\boldsymbol{P}$ with entries $0 < p_{ij} < 1$, and there are positive integers $m_1, \dots, m_a$ with $\sum_{i=1}^{a} m_i = m$ and $n_1, \dots, n_b$ with $\sum_{i=1}^{b} n_i = n$ such that – after rearranging its rows and columns – the matrix $\boldsymbol{B}$ can be divided into $a \times b$ blocks, where the block (i, j) is an $m_i \times n_j$ matrix with entries all equal to p_{ij} ($1 \le i \le a$, $1 \le j \le b$).

Let us fix $\boldsymbol{P}$, blow it up to an $m \times n$ matrix $\boldsymbol{B}_{m\times n}$, and consider the noisy matrix $\boldsymbol{A}_{m\times n} = \boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n}$ as $m, n \to \infty$ under one or both of the subsequent growth rate conditions.

Definition 8. The following growth rate conditions for the growth of the sizes and that of the cluster sizes of an $m \times n$ rectangular array are introduced.

GC1 There exists a constant $0 < c \le \frac{1}{a}$ such that $\frac{m_i}{m} \ge c$ $(i = 1, \dots, a)$ and a constant $0 < d \le \frac{1}{b}$ such that $\frac{n_i}{n} \ge d$ $(i = 1, \dots, b)$.

GC2 There exist constants $C \ge 1$, $D \ge 1$, and $C_0 > 0$, $D_0 > 0$ such that $m \le C_0 n^C$ and $n \le D_0 m^D$ for sufficiently large m and n.

We remark the following:

- **GC1** implies that

$$c \le \frac{m_k}{m_i} \le \frac{1}{c} \quad \text{and} \quad d \le \frac{n_\ell}{n_j} \le \frac{1}{d} \tag{25}$$

hold for any pair of indices $k, i \in \{1, \dots, a\}$ and $\ell, j \in \{1, \dots, b\}$.

- **GC2** implies that for sufficiently large m and n,

$$(\frac{1}{D_0})^{\frac{1}{D}} n^{\frac{1}{D}} \le m \le C_0 n^C \quad \text{and} \quad (\frac{1}{C_0})^{\frac{1}{C}} m^{\frac{1}{C}} \le n \le D_0 m^D.$$

Therefore, **GC2** mildly regulates the relation between m and n, but they need not tend to infinity at the same rate.

While perturbing $\boldsymbol{B}_{m\times n}$ by $\boldsymbol{W}_{m\times n}$, assume that for the uniform bound of the entries of $\boldsymbol{W}_{m\times n}$ the condition

$$K \le \min\{ \min_{\substack{i\in\{1,\dots,a\}\\ j\in\{1,\dots,b\}}} p_{ij} \,,\, 1 - \max_{\substack{i\in\{1,\dots,a\}\\ j\in\{1,\dots,b\}}} p_{ij} \} \tag{26}$$

is satisfied. In this way, the entries of $\boldsymbol{A}_{m\times n}$ are in the [0,1] interval, and hence, $\boldsymbol{A}_{m\times n}$ defines a standardized contingency table. We are interested in asymptotic properties of the SVD of this expanding contingency table sequence as $n, m \to \infty$ under the growth rate conditions.

With an appropriate Wigner-noise we can guarantee that the noisy table $\boldsymbol{A}_{m\times n}$ contains 1's in the (u, v)-th block with probability p_{uv}, and 0's otherwise. Indeed, for indices $1 \le u \le a$, $1 \le v \le b$, and $i \in R_u$, $j \in C_v$ let

$$w_{ij} := \begin{cases} 1 - p_{uv} & \text{with probability} \quad p_{uv} \\ -p_{uv} & \text{with probability} \quad 1 - p_{uv} \end{cases} \tag{27}$$

be independent random variables. This $\boldsymbol{W}_{m\times n}$ satisfies the conditions of Definition 6 with uniformly bounded entries of zero expectation. Let $R_1, \dots, R_a$ and $C_1, \dots, C_b$ denote the row- and column clusters induced by the blow-up. In the random 0-1 contingency table $\boldsymbol{A}_{m\times n}$, the row and column categories of R_u and C_v are in interaction with probability p_{uv}. Such schemes are sought for in microarray analysis and they are called chess-board patterns, see [50] for details. In terms of microarrays, the above property means that genes of the same cluster R_u equally influence conditions of the same cluster C_v.

Definition 9. Let $\boldsymbol{A}_{m\times n}$ be an $m \times n$ contingency table and $\mathcal{P}_{m,n}$ a property which mostly depends on the SVD of $\boldsymbol{A}_{m\times n}$. Then $\boldsymbol{A}_{m\times n}$ can have the property $\mathcal{P}_{m,n}$ in the two following senses.

WP1 The property $\mathcal{P}_{m,n}$ holds for $\boldsymbol{A}_{m\times n}$ *with probability tending to 1* if $\lim_{m,n\to\infty} \mathbb{P}\left(\boldsymbol{A}_{m\times n} \in \mathcal{P}_{m,n}\right) = 1$.

AS The property $\mathcal{P}_{m,n}$ holds for $\boldsymbol{A}_{m\times n}$ *almost surely* if

$$\mathbb{P}\left(\exists\, m_0, n_0 \in \mathbb{N}\,\text{s.t. for}\, m \geq m_0 \,\text{and}\, n \geq n_0 \,:\, \boldsymbol{A}_{m\times n} \in \mathcal{P}_{m,n}\right) = 1.$$

Here we may assume **GC1** and/or **GC2** for the simultaneous growth of m and n. In fact, **GC2** will only be used for noisy correspondence matrices.

Analogously to the quadratic case, **AS** always implies **WP1**. Furthermore, if in addition to **WP1**

$$\sum_{m,n=1}^{\infty} \mathbb{P}\left(\boldsymbol{A}_{m\times n} \notin \mathcal{P}_{m,n}\right) < \infty$$

also holds, then, by the Borel–Cantelli lemma, $\boldsymbol{A}_{m\times n}$ has $\mathcal{P}_{m,n}$ **AS**.

To find **AS** estimate for the spectral norm, i.e. the largest singular value of $\boldsymbol{W}_{m\times n}$, the following is used. Let $\boldsymbol{W}_{m\times n}$ be a Wigner-noise with entries uniformly bounded by K. The $(m+n)\times(m+n)$ symmetric matrix

$$\widetilde{\boldsymbol{W}} = \frac{1}{K}\cdot\begin{pmatrix} \boldsymbol{0} & \boldsymbol{W}_{m\times n} \\ \boldsymbol{W}_{m\times n}^T & \boldsymbol{0} \end{pmatrix}$$

satisfies the conditions of Theorem **??**, its largest and smallest eigenvalues are

$$\lambda_i(\widetilde{\boldsymbol{W}}) = -\lambda_{n+m-i+1}(\widetilde{\boldsymbol{W}}) = \frac{1}{K}\cdot s_i(\boldsymbol{W}_{m\times n}), \quad i = 1,\ldots,\min\{m,n\},$$

while the others are zeros, where $\lambda_i(.)$ and $s_i(.)$ denote the ith largest eigenvalue and singular value of the matrix in the argument, respectively. Therefore

$$\boldsymbol{P}\left(|s_1(\boldsymbol{W}_{m\times n}) - \mathbb{E}(s_1(\boldsymbol{W}_{m\times n}))| > t\right) \leq \exp\left(-\frac{(1-o(1))t^2}{32K^2}\right). \tag{28}$$

The fact that $s_1(\boldsymbol{W}_{m\times n}) = \|\boldsymbol{W}_{m\times n}\| = \mathcal{O}(\sqrt{m+n})$ **WP1** and inequality (28) together ensure that $\mathbb{E}(\|\boldsymbol{W}\|) = \mathcal{O}(\sqrt{m+n})$. Hence, no matter how $\mathbb{E}\|\boldsymbol{W}_{m\times n}\|$ behaves when $m,n\to\infty$, the following rough estimate holds: there exist positive constants c_1 and c_2, depending merely on the common bound K of the entries of $\boldsymbol{W}_{m\times n}$, such that

$$\mathbb{P}\left(\|\boldsymbol{W}_{m\times n}\| > c_1\sqrt{m+n}\right) \leq e^{-c_2(m+n)}. \tag{29}$$

Since the right-hand side of (29) forms a convergent series, the spectral norm of a Wigner-noise $\boldsymbol{W}_{m\times n}$ is of order $\sqrt{m+n}$ **AS**. This observation will provide the base of the subsequent **AS** results, which are also **WP1** ones.

4.1 Singular values of a noisy contingency table

Proposition 4. *Under* ***GC1****, all the non-zero singular values of the blown-up contingency table* $\boldsymbol{B}_{m\times n}$ *are of order* $\sqrt{mn}$.

Proof. As there are at most a and b linearly independent rows and linearly independent columns in $\boldsymbol{B}_{m\times n}$, respectively, the rank r of the matrix $\boldsymbol{B}_{m\times n}$ cannot exceed $\min\{a,b\}$. Note that r is also the rank of the pattern matrix $\boldsymbol{P}$. Let $s_1 \geq s_2 \geq \cdots \geq s_r > 0$ be the positive singular values of $\boldsymbol{B}_{m\times n}$. Let $\mathbf{v}_k \in \mathbb{R}^m$, $\mathbf{u}_k \in \mathbb{R}^n$ be a singular vector pair corresponding to s_k, $k=1,\dots,r$. Without loss of generality, $\mathbf{v}_1,\dots,\mathbf{v}_r$ and $\mathbf{u}_1,\dots,\mathbf{u}_r$ can be chosen unit-norm, pairwise orthogonal vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

For the subsequent calculations we drop the subscript k, and $\mathbf{v}$, $\mathbf{u}$ denotes a singular vector pair corresponding to the singular value $s>0$ of the blown-up matrix $\boldsymbol{B}_{m\times n}$, $\|\mathbf{v}\| = \|\mathbf{u}\| = 1$. It is easy to see that they have piecewise constant structures: $\mathbf{v}$ has m_i coordinates equal to $v(i)$, $i=1,\dots,a$ and $\mathbf{u}$ has n_j coordinates equal to $u(j)$, $j=1,\dots,b$. With these coordinates, the singular value–singular vector equation

$$\boldsymbol{B}_{m\times n}\mathbf{u} = s\mathbf{v} \tag{30}$$

has the form

$$\sum_{j=1}^{b} n_j p_{ij} u(j) = s v(i), \quad i=1,\dots,a. \tag{31}$$

With the notation

$$\tilde{\mathbf{u}} = (u(1),\dots,u(a))^T, \qquad \tilde{\mathbf{v}} = (v(1),\dots,v(b))^T,$$

and

$$\boldsymbol{D}_a = \operatorname{diag}(m_1,\dots,m_a), \qquad \boldsymbol{D}_b = \operatorname{diag}(n_1,\dots,n_b), \tag{32}$$

the equations in (31) can be written as

$$\boldsymbol{P}\boldsymbol{D}_b\tilde{\mathbf{u}} = s\tilde{\mathbf{v}}. \tag{33}$$

Introducing the transformations

$$\mathbf{w} = \boldsymbol{D}_b^{1/2}\tilde{\mathbf{u}}, \qquad \mathbf{z} = \boldsymbol{D}_a^{1/2}\tilde{\mathbf{v}}, \tag{34}$$

the Equation 33 is equivalent to

$$\boldsymbol{D}_a^{1/2}\boldsymbol{P}\boldsymbol{D}_b^{1/2}\mathbf{w} = s\mathbf{z}. \tag{35}$$

Applying the transformation (34) for the $\tilde{\mathbf{u}}_k, \tilde{\mathbf{v}}_k$ pairs obtained from the $\mathbf{u}_k, \mathbf{v}_k$ pairs ($k=1,\dots,r$), orthonormal systems in $\mathbb{R}^a$ and $\mathbb{R}^b$ are obtained:

$$\mathbf{w}_k{}^T\mathbf{w}_\ell = \sum_{j=1}^{b} n_j u_k(j) u_\ell(j) = \delta_{k\ell} \quad \text{and} \quad \mathbf{z}_k{}^T\mathbf{z}_\ell = \sum_{i=1}^{a} m_i v_k(i) v_\ell(i) = \delta_{k\ell}.$$

Consequently, $\mathbf{z}_k$, $\mathbf{w}_k$ is a singular vector pair corresponding to the singular value s_k of the $a\times b$ matrix $\boldsymbol{D}_a^{1/2}\boldsymbol{P}\boldsymbol{D}_b^{1/2}$ ($k=1,\dots,r$). With the shrinking

$$\widetilde{\boldsymbol{D}}_a = \frac{1}{m}\boldsymbol{D}_a, \qquad \widetilde{\boldsymbol{D}}_b = \frac{1}{n}\boldsymbol{D}_b,$$

an equivalent form of (35) is yielded by

$$\widetilde{\boldsymbol{D}}_a^{1/2}\boldsymbol{P}\widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{w} = \frac{s}{\sqrt{mn}}\mathbf{z},$$

that is the $a \times b$ matrix $\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}$ has non-zero singular values $\frac{s_k}{\sqrt{mn}}$ with the same singular vector pairs $\mathbf{z}_k, \mathbf{w}_k$ $(k = 1, \dots, r)$. If the s_k's are not distinct numbers, the singular vector pairs corresponding to a multiple singular value are not unique, but still they can be obtained from the SVD of the shrunken matrix $\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}$.

Now we want to establish relations between the singular values of $\boldsymbol{P}$ and $\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}$. Since we are interested only in the first r singular values, where $r = \mathrm{rank}(\boldsymbol{B}_{m\times n}) = \mathrm{rank}(\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2})$, it suffices to consider vectors $\mathbf{x}$, for which $\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2} \mathbf{x} \neq \mathbf{0}$. Therefore, with $k \in \{1, \dots, r\}$ and an arbitrary k-dimensional subspace $H \subset \mathbb{R}^b$ one can write

$$\begin{aligned}
\min_{\mathbf{x}\in H} \frac{\|\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\mathbf{x}\|} &= \min_{\mathbf{x}\in H} \frac{\|\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|} \cdot \frac{\|\boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|} \cdot \frac{\|\widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\mathbf{x}\|} \\
&\geq \min_a s_a(\widetilde{\boldsymbol{D}}_a^{1/2}) \cdot \min_{\mathbf{x}\in H} \frac{\|\boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|} \cdot \min_b s_b(\widetilde{\boldsymbol{D}}_b^{1/2}) \\
&\geq \sqrt{cd} \cdot \min_{\mathbf{x}\in H} \frac{\|\boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|}{\|\widetilde{\boldsymbol{D}}_b^{1/2}\mathbf{x}\|},
\end{aligned}$$

with c, d of **GC1**. Now taking the maximum for all possible k-dimensional subspace H, we obtain that $s_k(\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}) \geq \sqrt{cd} \cdot s_k(\boldsymbol{P}) > 0$. On the other hand,

$$s_k(\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}) \leq \|\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2}\| \leq \|\widetilde{\boldsymbol{D}}_a^{1/2}\| \cdot \|\boldsymbol{P}\| \cdot \|\widetilde{\boldsymbol{D}}_b^{1/2}\| \leq \|\boldsymbol{P}\| \leq \sqrt{ab}.$$

These inequalities imply that $s_k(\widetilde{\boldsymbol{D}}_a^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_b^{1/2})$ is bounded from below and from above with a positive constant that does not depend on m and n, and because of $s_k(\widetilde{\boldsymbol{D}}_m^{1/2} \boldsymbol{P} \widetilde{\boldsymbol{D}}_n^{1/2}) = \frac{s_k}{\sqrt{mn}}$, we obtain that $s_1, \dots, s_r = \Theta(\sqrt{mn})$. □

Theorem 5. *Let $\boldsymbol{A}_{m\times n} = \boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n}$ be an $m \times n$ random matrix, where $\boldsymbol{B}_{m\times n}$ is a blown-up matrix with positive singular values $s_1, \dots, s_r$ and $\boldsymbol{W}_{m\times n}$ is a Wigner-noise. Then, under **GC1**, the matrix $\boldsymbol{A}_{m\times n}$ has r singular values $z_1, \dots, z_r$, such that*

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

and for the other singular values

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

*holds, **AS**.*

Proof. The statement follows from the analog of the Weyl's perturbation theorem for the singular values of rectangular matrices: if $s_i(\boldsymbol{A}_{m\times n})$ and $s_i(\boldsymbol{B}_{m\times n})$ denote the ith largest singular values of the matrix in the argument then for the difference of the corresponding pairs

$$|s_i(\boldsymbol{A}_{m\times n}) - s_i(\boldsymbol{B}_{m\times n})| \leq \|\boldsymbol{W}\|, \quad i = 1, \dots, \min\{m, n\}.$$

As a consequence of (29), $\|\boldsymbol{W}\|$ is of order $\sqrt{m+n}$ **AS**, that finishes the proof. □

Summarizing, with the notation

$$\varepsilon := \|\boldsymbol{W}_{m\times n}\| = \mathcal{O}(\sqrt{m+n}) \quad \text{and} \quad \Delta := \min_{1\le i\le r} s_i(\boldsymbol{B}_{m\times n}) = \Theta(\sqrt{mn}) \tag{36}$$

there is a spectral gap of size $\Delta - 2\varepsilon$ between the r largest and the other singular values of the perturbed matrix $\boldsymbol{A}_{m\times n}$, and this gap is significantly larger than ε.

4.2 Clustering the rows and columns via singular vector pairs

Here perturbation results for the corresponding singular vector pairs are established. To this end, with the help of Theorem 5, we estimate the distances between the corresponding right- and left-hand side eigenspaces of the matrices $\boldsymbol{B}_{m\times n}$ and $\boldsymbol{A}_{m\times n} = \boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n}$.

With the notation of the Proof of Proposition 4, $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathbb{R}^m$ and $\mathbf{u}_1, \dots, \mathbf{u}_n \in \mathbb{R}^n$ are orthonormal left- and right-hand side singular vectors of $\boldsymbol{B}_{m\times n}$. They are piecewise constant vectors over the row and column clusters determined by the blow-up; further, they satisfy

$$\boldsymbol{B}_{m\times n}\mathbf{u}_i = s_i\mathbf{v}_i, \quad i = 1, \dots, r \quad \text{and} \quad \boldsymbol{B}_{m\times n}\mathbf{u}_j = 0, \quad j = r+1, \dots, n.$$

Let us also denote the unit-norm, pairwise orthogonal left- and right-hand side singular vectors corresponding to the r outstanding singular values $z_1, \dots, z_r$ of $\boldsymbol{A}_{m\times n}$ by $\mathbf{y}_1, \dots, \mathbf{y}_r \in \mathbb{R}^m$ and $\mathbf{x}_1, \dots, \mathbf{x}_r \in \mathbb{R}^n$, respectively. Then, for them $\boldsymbol{A}_{m\times n}\mathbf{x}_i = z_i\mathbf{y}_i$ holds ($i = 1, \dots, r$). Let

$$F := \mathrm{Span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\} \quad \text{and} \quad G := \mathrm{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$$

denote the spanned linear subspaces of piecewise constant vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

Proposition 5. *With the above notation, under* ***GC1****, the following estimate holds* ***AS****:*

$$\sum_{i=1}^{r} \mathrm{dist}^2(\mathbf{y}_i, F) \le r\frac{\varepsilon^2}{(\Delta-\varepsilon)^2} = \mathcal{O}\left(\frac{m+n}{mn}\right), \tag{37}$$

and analogously,

$$\sum_{i=1}^{r} \mathrm{dist}^2(\mathbf{x}_i, G) \le r\frac{\varepsilon^2}{(\Delta-\varepsilon)^2} = \mathcal{O}\left(\frac{m+n}{mn}\right). \tag{38}$$

Proof. Let us choose one of the right-hand side singular vectors $\mathbf{x}_1, \dots, \mathbf{x}_r$ of $\boldsymbol{A}_{m\times n} = \boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n}$ and denote it simply by $\mathbf{x}$ with corresponding singular value z. We will estimate the distance between $\mathbf{y} = \frac{1}{z}\boldsymbol{A}\mathbf{x}$ and F. For this purpose, we expand $\mathbf{x}$ and $\mathbf{y}$ in the orthonormal bases $\mathbf{u}_1, \dots, \mathbf{u}_n$ and $\mathbf{v}_1, \dots, \mathbf{v}_m$, respectively:

$$\mathbf{x} = \sum_{i=1}^{n} t_i\mathbf{u}_i \quad \text{and} \quad \mathbf{y} = \sum_{i=1}^{m} l_i\mathbf{v}_i.$$

Then on the one hand,

$$\boldsymbol{A}_{m\times n}\mathbf{x} = (\boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n})\mathbf{x} = \sum_{i=1}^{r} t_i s_i \mathbf{v}_i + \boldsymbol{W}_{m\times n}\mathbf{x} \tag{39}$$

and, on the other hand,

$$\boldsymbol{A}_{m\times n}\mathbf{x} = z\mathbf{y} = \sum_{i=1}^{m} zl_i\mathbf{v}_i. \tag{40}$$

Equating the right-hand sides of (39) and (40), we obtain

$$\sum_{i=1}^{r}(zl_i - t_i s_i)\mathbf{v}_i + \sum_{i=r+1}^{m} zl_i\mathbf{v}_i = \boldsymbol{W}_{m\times n}\mathbf{x}.$$

In view of the Pythagorean theorem,

$$\sum_{i=1}^{r}(zl_i - t_i s_i)^2 + z^2 \sum_{i=r+1}^{m} l_i^2 = \|\boldsymbol{W}_{m\times n}\mathbf{x}\|^2 \le \varepsilon^2, \tag{41}$$

because $\|\mathbf{x}\| = 1$ and $\|\boldsymbol{W}\| = \varepsilon$.

As $z \ge \Delta - \varepsilon$ holds **AS** by Theorem 5,

$$\operatorname{dist}^2(\mathbf{y}, F) = \sum_{i=r+1}^{m} l_i^2 \le \frac{\varepsilon^2}{z^2} \le \frac{\varepsilon^2}{(\Delta-\varepsilon)^2}.$$

The order of the above estimate follows from the order of ε and Δ of (36):

$$\operatorname{dist}^2(\mathbf{y}, F) = \mathcal{O}(\frac{m+n}{mn}), \tag{42}$$

AS. Applying (42) for the left-hand side singular vectors $\mathbf{y}_1, \dots, \mathbf{y}_r$, by Definition 4,

$$\mathbb{P}\left(\exists m_{0i}, n_{0i} \in \mathbb{N}\,\text{s.t. for}\, m \ge m_{0i}\, n \ge n_{0i} \colon \operatorname{dist}^2(\mathbf{y}_i, F) \le \frac{\varepsilon^2}{(\Delta-\varepsilon)^2}\right) = 1,$$

for $i = 1, \dots, r$. Hence,

$$\mathbb{P}\left(\exists m_0, n_0 \in \mathbb{N}\,\text{s.t. for}\, m \ge m_0\, n \ge n_0 \colon \operatorname{dist}^2(\mathbf{y}_i, F) \le \frac{\varepsilon^2}{(\Delta-\varepsilon)^2},\, i = 1, \dots, r\right) = 1,$$

consequently,

$$\mathbb{P}\left(\exists m_0, n_0 \in \mathbb{N}\,\text{s. t. for}\, m \ge m_0\, n \ge n_0 \colon \sum_{i=1}^{r} \operatorname{dist}^2(\mathbf{y}_i, F) \le r\frac{\varepsilon^2}{(\Delta-\varepsilon)^2}\right) = 1$$

also holds, which finishes the proof of the first statement.

The estimate for the squared distance between G and a right-hand side singular vector $\mathbf{x}$ of $\boldsymbol{A}_{m\times n}$ follows in the same way starting with $\boldsymbol{A}_{m\times n}^T\mathbf{y} = z\mathbf{x}$ and using the fact that $\boldsymbol{A}_{m\times n}^T$ has the same singular values as $\boldsymbol{A}_{m\times n}$. □

By Proposition 5, the individual distances between the original and the perturbed subspaces and also the sum of these distances tend to zero **AS** as $m, n \to \infty$ under **GC1**.

Now let $\boldsymbol{A}_{m\times n}$ be a microarray on m genes and n conditions, with a_{ij} denoting the expression level of gene i under condition j. We assume that $\boldsymbol{A}_{m\times n}$ is a noisy random matrix obtained by adding a Wigner-noise $\boldsymbol{W}_{m\times n}$ to the blown-up matrix $\boldsymbol{B}_{m\times n}$. Let us denote by $R_1,\dots,R_a$ the partition of the genes and by $C_1,\dots,C_b$ the partition of the conditions with respect to the blow-up (they can also be thought of as clusters of the genes and conditions).

Proposition 5 also implies the well-clustering property of the representatives of the genes and conditions in the following representation. Let $\boldsymbol{Y}$ be the $m\times r$ matrix containing the left-hand side singular vectors $\mathbf{y}_1,\dots,\mathbf{y}_r$ of $\boldsymbol{A}_{m\times n}$ in its columns. Likewise, let $\boldsymbol{X}$ be the $n\times r$ matrix containing the right-hand side singular vectors $\mathbf{x}_1,\dots,\mathbf{x}_r$ of $\boldsymbol{A}_{m\times n}$ in its columns. Let the r-dimensional representatives of the genes be the row vectors of $\boldsymbol{Y}$: $\mathbf{y}^1,\dots,\mathbf{y}^m\in\mathbb{R}^r$, while the r-dimensional representatives of the conditions be the row vectors of $\boldsymbol{X}$: $\mathbf{x}^1,\dots,\mathbf{x}^n\in\mathbb{R}^r$. Let $S_a^2(\boldsymbol{Y})$ denote the a-variance, introduced in (**??**) of the genes' representatives:

$$S_a^2(\boldsymbol{Y})=\min_{\{R'_1,\dots,R'_a\}}\sum_{i=1}^{a}\sum_{j\in R'_i}\|\mathbf{y}^j-\bar{\mathbf{y}}^i\|^2,\quad\text{where}\quad\bar{\mathbf{y}}^i=\frac{1}{m_i}\sum_{j\in R'_i}\mathbf{y}^j,$$

while $S_b^2(\boldsymbol{X})$ denotes the b-variance of the conditions' representatives:

$$S_b^2(\boldsymbol{X})=\min_{\{C'_1,\dots,C'_b\}}\sum_{i=1}^{b}\sum_{j\in C'_i}\|\mathbf{x}^j-\bar{\mathbf{x}}^i\|^2,\quad\text{where}\quad\bar{\mathbf{x}}^i=\frac{1}{n_i}\sum_{j\in C'_i}\mathbf{x}^j,$$

when the partitions $\{R'_1,\dots,R'_a\}$ and $\{C'_1,\dots,C'_b\}$ vary over all a- and b-partitions of the genes and conditions, respectively.

Theorem 6. *With the above notation, under **GC1**, for the a- and b-variances of the representation of the microarray $\boldsymbol{A}_{m\times n}$ the relations*

$$S_a^2(\boldsymbol{Y})=\mathcal{O}\left(\frac{m+n}{mn}\right)\quad\textit{and}\quad S_b^2(\boldsymbol{X})=\mathcal{O}\left(\frac{m+n}{mn}\right)$$

*hold **AS**.*

Proof. Since $S_a^2(\boldsymbol{Y})\le\sum_{i=1}^{a}\sum_{j\in R_i}\|\mathbf{y}^j-\bar{\mathbf{y}}^i\|^2$ and $S_b^2(\boldsymbol{X})\le\sum_{i=1}^{b}\sum_{j\in C_i}\|\mathbf{x}^j-\bar{\mathbf{x}}^i\|^2$, by the considerations of the Proof of Proposition **??**, the right-hand sides are equal to the left-hand sides of (37) and (38), respectively. Therefore, they are also of order $\frac{m+n}{mn}$. □

Hence, the addition of any kind of a Wigner-noise to a rectangular matrix that has a blown-up structure $\boldsymbol{B}_{m\times n}$, will not change the order of the outstanding singular values, and the block structure of $\boldsymbol{B}_{m\times n}$ can be reconstructed from the representatives of the row and column items of the noisy matrix $\boldsymbol{A}_{m\times n}$. So far, we have only used **GC1**, and no restriction for the relation between m and n has been made. For noisy correspondence matrices, **GC2** will also be used.

5 Discrepancy and spectra

Here under clustering we understand partition of the vertex-set into subsets of similar vertices, i.e., members of a cluster behave similarity toward members of each (other or own) cluster. We will generalize the Laplacian and modularity based spectral clustering methods to recover so-called regular cluster pairs such that the information flow between the pairs and within the clusters is as homogeneous as possible. The notion of volume-regularity is also extended to contingency tables. For this purpose, we take into account both ends of the normalized Laplacian spectrum, i.e., large absolute value, so-called structural eigenvalues of the normalized modularity matrix, or the largest singular values of the normalized contingency table. First we introduce the notion of multiway discrepancy for rectangular arrays of nonnegative entries, of which the quadratic edge-weight matrices are special cases.

5.1 Estimating the singular values of normalized contingency tables by the multiway discrepancy

Definition 10. The multiway discrepancy of the rectangular array $\boldsymbol{C}$ of nonnegative entries in the proper k-partition $R_1, \dots, R_k$ of its rows and $C_1, \dots, C_k$ of its columns is

$$\mathrm{md}(\boldsymbol{C}; R_1, \dots, R_k, C_1, \dots, C_k) = \max_{\substack{1 \le a,b \le k \\ X \subset R_a,\, Y \subset C_b}} \frac{|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)|}{\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}}, \tag{43}$$

where $c(X,Y)$, $\mathrm{Vol}(X)$, and $\mathrm{Vol}(Y)$ are volumes, whereas $\rho(R_a, C_b) = \frac{c(R_a,C_b)}{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$ denotes the relative density between R_a and C_b. The minimum k-way discrepancy of $\boldsymbol{C}$ itself is

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

We will also extend this notion to an edge-weighted graph G and denote it by $\mathrm{md}_k(G)$. In that setup, $\boldsymbol{C}$ plays the role of the edge-weight matrix: symmetric in the undirected; quadratic, but usually not symmetric in the directed case; and it is the adjacency matrix if G is a simple graph.

Note that the division by $\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}$ ensures that the multiway discrepancy is not affected by the scaling of the entries of $\boldsymbol{C}$, akin to the normalized table $\boldsymbol{C}_D$. Therefore, without loss of generality, $\sum_{i=1}^n \sum_{j=1}^m c_{ij} = 1$ will be assumed.

Observe that $\mathrm{md}(\boldsymbol{C}; R_1, \dots, R_k, C_1, \dots, C_k)$ is the smallest α such that for every R_a, C_b pair and for every $X \subset R_a$, $Y \subset C_b$,

$$|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)| \le \alpha \sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)} \tag{44}$$

holds. Therefore, in the k-partitions of the rows and columns, giving the minimum k-way discrepancy (say, α^*) of $\boldsymbol{C}$, every R_a, C_b pair is α^*-regular in terms of the volumes, and α^* is the smallest possible discrepancy that can be attained with proper k-partitions. It resembles the notion of ϵ-regular pairs in the Szemerédi Regularity Lemma [64], albeit with given number of vertex-clusters, which are usually not equitable; further, with volumes, instead of cardinalities.

Historically, the notion of discrepancy together with the expander mixing lemma was introduced for simple, regular graphs, see e.g., Alon, Spencer, Hoory, Linial, Widgerson [5, 44], and extended to Hermitian matrices by Bollobás, Nikiforov [22]. In Chung, Graham, Wilson [27], the authors use the term quasirandom for simple graphs that satisfy any of some equivalent properties, some of them closely related to discrepancy and eigenvalue separation. Chung and Graham [29] prove that for simple graphs 'small' discrepancy $\mathrm{disc}(G)$ (with our notation, $\mathrm{md}_1(G)$) is caused by eigenvalue 'separation': the second largest singular value (which is also the second largest absolute value eigenvalue), s_1, of the normalized adjacency matrix is 'small', i.e., separated from the trivial singular value $s_0 = 1$, which is the edge of the spectrum. More exactly, they prove $\mathrm{disc}(G) \le s_1$, hence giving some kind of generalization of the *expander mixing lemma for irregular graphs*.

In the other direction, for Hermitian matrices, Bollobás and Nikiforov [22] estimate the second largest singular value of an $n \times n$ Hermitian matrix $\boldsymbol{A}$ by $C\mathrm{disc}(\boldsymbol{A}) \log n$ (where C is an absolute constant), and show that this is best possible up to a multiplicative constant. Bilu and Linial [9] prove the converse of the expander mixing lemma for simple regular graphs, but their key Lemma 3.3, producing this statement, goes beyond regular graphs. In Alon et al. [6], the authors relax the notion of eigenvalue separation to essential eigenvalue separation (by introducing a parameter for it, and requiring the separation only for the eigenvalues of a relatively large part of the graph). Then they prove relations between the constants of this kind of eigenvalue separation and the discrepancy.

For a general rectangular array $\boldsymbol{C}$ of nonnegative entries, Butler [24] proves the following forward and backward statement in the $k = 1$ case:

$$\mathrm{disc}(\boldsymbol{C}) \le s_1 \le 150\mathrm{disc}(\boldsymbol{C})(1 - 8 \ln \mathrm{disc}(\boldsymbol{C})), \tag{45}$$

where his $\mathrm{disc}(\boldsymbol{C})$ is our $\mathrm{md}_1(\boldsymbol{C})$ and, with our notation, s_1 is the largest nontrivial singular value of $\boldsymbol{C}_D$ (he denotes is with σ_2). Since $s_1 < 1$, the upper estimate makes sense for very small discrepancy, in particular, for $\mathrm{disc}(\boldsymbol{C}) \le 8.868 \times 10^{-5}$. The lower estimate of (45) further generalizes the expander mixing lemma to rectangular matrices, but it can be proved with the same tools as in the quadratic case (see the forthcoming Proposition 6 in Section 5.3).

The above papers consider the overall discrepancy in the sense that $\mathrm{disc}(\boldsymbol{C})$ or $\mathrm{disc}(G)$ measure the largest possible deviation between the actual and expected connectedness of arbitrary (sometimes disjoint) subsets X, Y, where under expected the hypothesis of independence is understood (which corresponds to the rank 1 approximation of the normalized matrix). Our purpose is, in the multicluster scenario, to find similar relations between the minimum k-way discrepancy and the SVD of the normalized matrix, for given k. In one direction, we are able to prove the following.

Theorem 7 ([16])**.** *For every non-degenerate real matrix $\boldsymbol{C}$ of nonnegative entries and integer $1 \le k \le \mathrm{rank}(\boldsymbol{C})$,*

$$s_k \le 9\mathrm{md}_k(\boldsymbol{C})(k + 2 - 9k \ln \mathrm{md}_k(\boldsymbol{C})) \tag{46}$$

holds, provided $0 < \mathrm{md}_k(\boldsymbol{C}) < 1$, where s_k is the k-th largest nontrivial singular value of the normalized matrix $\boldsymbol{C}_D$ of $\boldsymbol{C}$.

Note that $\mathrm{md}_k(\boldsymbol{C}) = 0$ if $\boldsymbol{C}$ has a block structure with k row- and column-blocks, in which case $s_k = 0$ also holds. Likewise, $\mathrm{md}_k(\boldsymbol{C}) < 1$ is not a peculiar requirement, since in view of $s_k < 1$, the upper bound of the theorem has relevance only for $\mathrm{md}_k(\boldsymbol{C})$ much smaller than 1; for example, for $\mathrm{md}_1(\boldsymbol{C}) \le 1.866 \times 10^{-3}$, $\mathrm{md}_2(\boldsymbol{C}) \le 8.459 \times 10^{-4}$, $\mathrm{md}_3(\boldsymbol{C}) \le 5.329 \times 10^{-4}$, etc.

Before proving the theorem, we encounter some lemmas of other authors that will be used, possibly with some modifications.

Lemma 3 of Bollobás and Nikiforov [22] is the key to prove their main result. This lemma states that to every $0 < \varepsilon < 1$ and vector $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\| = 1$, there exists a vector $\mathbf{y} \in \mathbb{C}^n$ such that its coordinates take no more than $\lceil \frac{8\pi}{\varepsilon} \rceil \lceil \frac{4}{\varepsilon} \log \frac{2n}{\varepsilon} \rceil$ distinct values and $\|\mathbf{x} - \mathbf{y}\| \le \varepsilon$. We will rather use the construction of the following lemma, which is indeed a consequence of Lemma 3 of [22].

Lemma 3 (Lemma 3 of Butler [24])**.** *To any vector $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\| = 1$ and diagonal matrix $\boldsymbol{D}$ of positive real diagonal entries, one can construct a step-vector $\mathbf{y} \in \mathbb{C}^n$ such that $\|\mathbf{x} - \boldsymbol{D}\mathbf{y}\| \le \frac{1}{3}$, $\|\boldsymbol{D}\mathbf{y}\| \le 1$, and the nonzero entries of $\mathbf{y}$ are of the form $\left(\frac{4}{5}\right)^j e^{\frac{\ell}{29}2\pi i}$ with appropriate integers j (taking on $\mathcal{O}(\log n)$ distinct values) and ℓ $(0 \le \ell \le 28)$.*

Note that starting with an $\mathbf{x}$ of real coordinates, we do not need all the 29 values of ℓ, only two of them will show up, as it follows from a better understanding of the construction of [24]. In fact, by the idea of [22], j's come from dividing the coordinates of $\boldsymbol{D}^{-1}\mathbf{x}/\|\boldsymbol{D}^{-1}\mathbf{x}\|$ in decreasing absolute values into groups, where the cut-points are powers of $\frac{4}{5}$. With the notation $\mathbf{x} = (x_s))_{s=1}^n$, if x_s is in the j-th group, then the corresponding coordinate of the approximating complex vector $\mathbf{y} = (y_s)_{s=1}^n$ is as follows. If $x_s = 0$, then $y_s = 0$, otherwise $y_s = \left(\frac{4}{5}\right)^j e^{\left(\lfloor \frac{29\theta}{2\pi} \rfloor / 29\right)2\pi i}$, where θ is the argument of x_s, $0 \le \theta < 2\pi$, and therefore, $\ell = \lfloor \frac{29\theta}{2\pi} \rfloor$ is an integer between 0 and 28. However, when the coordinates of $\mathbf{x}$ are real numbers, then only the values 0 and 14 of ℓ can occur, since θ can take only one of the values 0 or π, depending on whether x_s is positive or negative. We will intensively use this observation in our proof.

Lemma 4 (Lemma 4 of Butler [24])**.** *Let $\boldsymbol{M}$ be a matrix with largest singular value σ and corresponding unit-norm singular vector pair $\mathbf{v}, \mathbf{u}$. If $\mathbf{x}$ and $\mathbf{y}$ are vectors such that $\|\mathbf{x}\| \le 1$, $\|\mathbf{y}\| \le 1$, $\|\mathbf{v} - \mathbf{x}\| \le \frac{1}{3}$, $\|\mathbf{u} - \mathbf{y}\| \le \frac{1}{3}$, then $\sigma \le \frac{9}{2}\langle \mathbf{x}, \boldsymbol{M}\mathbf{y} \rangle$.*

Lemma 5 (Thompson)**.** *Let the $n \times n$ matrix $\boldsymbol{A}$ have singular values $\alpha_1 \ge \cdots \ge \alpha_n$ and $1 \le k \le n$ be a fixed integer. Then an $n \times n$ matrix $\mathbf{X}$ exists with $\mathrm{rank}(\mathbf{X}) \le k$ such that $\boldsymbol{B} = \boldsymbol{A} + \mathbf{X}$ has singular values $\beta_1 \ge \cdots \ge \beta_n$ if and only if*

$$\alpha_{i+k} \le \beta_i \le \alpha_{i-k}, \qquad i = 1, \dots, n$$

with the understanding that $\alpha_j = +\infty$ if $j \le 0$ and $\alpha_j = 0$ if $j \ge n$.

Proof of Theorem 7. Assume that $\alpha^* = \mathrm{md}_k(\boldsymbol{C}) \in (0,1)$ and it is attained with the proper k-partition $R_1, \dots, R_k$ of the rows and $C_1, \dots, C_k$ of the columns of $\boldsymbol{C}$; i.e., for every R_a, C_b pair and $X \subset R_a$, $Y \subset C_b$ we have

$$|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)| \le \alpha^* \sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}. \tag{47}$$

Our purpose is to put Inequality (47) in a matrix form by using indicator vectors and introducing the $m \times n$ auxiliary matrix

$$\boldsymbol{F} = \boldsymbol{C} - \boldsymbol{D}_{row}\boldsymbol{R}\boldsymbol{D}_{col}, \tag{48}$$

where $\boldsymbol{R} = (\rho(R_a, C_b))$ is the $m \times n$ block-matrix of $k \times k$ blocks with entries equal to $\rho(R_a, C_b)$ over the block $R_a \times C_b$. With the indicator vectors $\mathbf{1}_X$ and $\mathbf{1}_Y$ of $X \subset R_a$ and $Y \subset C_b$, Inequality (47) has the following equivalent form:

$$|\langle \mathbf{1}_X, \boldsymbol{F}\mathbf{1}_Y \rangle| \le \alpha^* \sqrt{\langle \mathbf{1}_X, \boldsymbol{C}\mathbf{1}_n \rangle \langle \mathbf{1}_m, \boldsymbol{C}\mathbf{1}_Y \rangle}, \tag{49}$$

where $\mathbf{1}_n$ denotes the all 1's vector of size n. At the same time, Equation (48) yields

$$\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2} = \boldsymbol{D}_{row}^{-1/2} \boldsymbol{C} \boldsymbol{D}_{col}^{-1/2} - \boldsymbol{D}_{row}^{1/2} \boldsymbol{R} \boldsymbol{D}_{col}^{1/2} = \boldsymbol{C}_D - \boldsymbol{D}_{row}^{1/2} \boldsymbol{R} \boldsymbol{D}_{col}^{1/2}.$$

Since the rank of the matrix $\boldsymbol{D}_{row}^{1/2} \boldsymbol{R} \boldsymbol{D}_{col}^{1/2}$ is at most k, by the upper estimate of Lemma 5 (with the rolecast $\boldsymbol{A} = \boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}$, $\boldsymbol{B} = \boldsymbol{C}_D$, $\mathbf{X} = \boldsymbol{D}_{row}^{1/2} \boldsymbol{R} \boldsymbol{D}_{col}^{1/2}$, and $i = k+1$) [1] we obtain the following upper estimate for s_k, that is the $(k+1)$-th largest (including the trivial 1) singular value of $\boldsymbol{C}_D$:

$$s_k \le s_{max}(\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}) = \|\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}\|,$$

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

Let $\mathbf{v} \in \mathbb{R}^m$ be the left and $\mathbf{u} \in \mathbb{R}^n$ be the right unit-norm singular vector corresponding to the maximal singular value of $\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}$, i.e.,

$$|\langle \mathbf{v}, (\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2})\mathbf{u} \rangle| = \|\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}\|.$$

In view of Lemma 3, there are step-vectors $\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{y} \in \mathbb{C}^n$ such that $\|\mathbf{v} - \boldsymbol{D}_{row}^{1/2}\mathbf{x}\| \le \frac{1}{3}$ and $\|\mathbf{u} - \boldsymbol{D}_{col}^{1/2}\mathbf{y}\| \le \frac{1}{3}$; further, $\|\boldsymbol{D}_{row}^{1/2}\mathbf{x}\| \le 1$ and $\|\boldsymbol{D}_{col}^{1/2}\mathbf{y}\| \le 1$. Then Lemma 4 yields

$$\|\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2}\| \le \frac{9}{2} \left| \langle (\boldsymbol{D}_{row}^{1/2}\mathbf{x}), (\boldsymbol{D}_{row}^{-1/2} \boldsymbol{F} \boldsymbol{D}_{col}^{-1/2})(\boldsymbol{D}_{col}^{1/2}\mathbf{y}) \rangle \right| = \frac{9}{2} |\langle \mathbf{x}, \boldsymbol{F}\mathbf{y} \rangle|.$$

Now we will use the construction of the proof of the Lemma 3 in the special case when the vectors $\mathbf{v} = (v_s))_{s=1}^m$ and $\mathbf{u} = (u_s))_{s=1}^n$, to be approximated, have real coordinates. Therefore, only the following three types of coordinates of the approximating complex vectors $\mathbf{x} = (x_s))_{s=1}^m$ and $\mathbf{y} = (y_s)_{s=1}^n$ will appear. If $v_s = 0$, then $x_s = 0$; if $v_s > 0$, then $x_s = (\frac{4}{5})^j$ with some integer j; if $v_s < 0$, then $x_s = (\frac{4}{5})^j e^{\frac{28}{29}\pi i}$ with some integer j. Likewise, if $u_s = 0$, then $y_s = 0$; if $u_s > 0$, then $y_s = (\frac{4}{5})^\ell$ with some integer ℓ; if $u_s < 0$, then $y_s = (\frac{4}{5})^\ell e^{\frac{28}{29}\pi i}$ with some integer ℓ. With these observations, the step-vectors $\mathbf{x}$ and $\mathbf{y}$ can be written as the following finite sums with respect to the integers j and ℓ:

$$\mathbf{x} = \sum_j (\frac{4}{5})^j \mathbf{x}^{(j)}, \quad \mathbf{x}^{(j)} = \sum_{a=1}^k (\mathbf{1}_{\mathcal{X}_{ja1}} + e^{\frac{28}{29}\pi i} \mathbf{1}_{\mathcal{X}_{ja2}}), \quad \text{where}$$

$$\mathcal{X}_{ja1} = \{s : \, v_s > 0, \, s \in R_a\} \quad \text{and} \quad \mathcal{X}_{ja2} = \{s : \, v_s < 0, \, s \in R_a\};$$

[1] Actually, Lemma 5 is about square matrices, but in the possession of a rectangular one, we can supplement it with zero rows or columns to make it quadratic; further, the nonzero singular values of the so obtained square matrix are the same as those of the rectangular one, supplemented with additional zero singular values that will not alter the shifted interlacing facts.

likewise,

$$\mathbf{y} = \sum_{\ell} (\frac{4}{5})^{\ell} \mathbf{y}^{(\ell)}, \quad \mathbf{y}^{(\ell)} = \sum_{b=1}^{k} (\mathbf{1}_{\mathcal{Y}_{\ell b1}} + e^{\frac{28}{29}\pi i} \mathbf{1}_{\mathcal{Y}_{\ell b2}}), \quad \text{where}$$

$$\mathcal{Y}_{\ell b1} = \{s :\, u_s > 0,\, s \in C_b\} \quad \text{and} \quad \mathcal{Y}_{\ell b2} = \{s :\, u_s < 0,\, s \in C_b\}.$$

It is important that the $2k$ indicator vectors appearing in the decomposition of any $\mathbf{x}^{(j)}$ or $\mathbf{y}^{(\ell)}$ are disjointly supported, and so, all the coordinates of these vectors are of absolute value 1. These considerations give rise to the following estimation.

$$\begin{aligned}
|\langle \mathbf{x}^{(j)}, \boldsymbol{F} \mathbf{y}^{(\ell)} \rangle| &\le \sum_{a=1}^{k} \sum_{p=1}^{2} \sum_{b=1}^{k} \sum_{q=1}^{2} \left| \langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{F} \mathbf{1}_{\mathcal{Y}_{\ell bq}} \rangle \right| \\
&\overset{(49)}{\le} \sum_{a=1}^{k} \sum_{p=1}^{2} \sum_{b=1}^{k} \sum_{q=1}^{2} \alpha^* \sqrt{\langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{C} \mathbf{1}_n \rangle \langle \mathbf{1}_m, \boldsymbol{C} \mathbf{1}_{\mathcal{Y}_{\ell bq}} \rangle} \\
&\le \alpha^* 2k \sqrt{\sum_{a=1}^{k} \sum_{p=1}^{2} \sum_{b=1}^{k} \sum_{q=1}^{2} \langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{C} \mathbf{1}_n \rangle \langle \mathbf{1}_m, \boldsymbol{C} \mathbf{1}_{\mathcal{Y}_{\ell bq}} \rangle} \qquad (50) \\
&= 2k\alpha^* \sqrt{\langle \sum_{a=1}^{k} \sum_{p=1}^{2} \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{C} \mathbf{1}_n \rangle \langle \mathbf{1}_m, \boldsymbol{C} \sum_{b=1}^{k} \sum_{q=1}^{2} \mathbf{1}_{\mathcal{Y}_{\ell bq}} \rangle} \\
&= 2k\alpha^* \sqrt{\langle |\mathbf{x}^{(j)}|, \boldsymbol{C} \mathbf{1}_n \rangle \langle \mathbf{1}_m, \boldsymbol{C} |\mathbf{y}^{(\ell)}| \rangle},
\end{aligned}$$

where in the first inequality we used the triangle inequality and $|e^{\frac{28}{29}\pi i}| = 1$, in the second one we used (49), while in the third one, the Cauchy–Schwarz inequality with $4k^2$ terms.

In the last step we exploited that the indicator vectors composing $\mathbf{x}^{(j)}$ and $\mathbf{y}^{(\ell)}$ are disjointly supported. We also introduced the notation $|\mathbf{z}| = (|z_s|)_{s=1}^n$ for the real vector, the coordinates of which are the absolute values of the corresponding coordinates of the (possibly complex) vector $\mathbf{z}$. (Note that the so introduced $|\mathbf{z}|$ is a vector, unlike $\|\mathbf{z}\| = (\sum_{s=1}^n |z_s|^2)^{1/2}$.) In the same spirit, let $|\boldsymbol{M}|$ denote the matrix whose entries are the absolute values of the corresponding entries of $\boldsymbol{M}$ (we will use this only for real matrices). With this formalism, this is the right moment to prove the following inequalities that will be used soon to finish the proof:

$$\sum_{\ell} |\langle \mathbf{x}^{(j)}, \boldsymbol{F} \mathbf{y}^{(\ell)} \rangle| \le 2 \langle |\mathbf{x}^{(j)}|, \boldsymbol{C} \mathbf{1}_n \rangle, \quad \sum_{j} |\langle \mathbf{x}^{(j)}, \boldsymbol{F} \mathbf{y}^{(\ell)} \rangle| \le 2 \langle \mathbf{1}_m, \boldsymbol{C} |\mathbf{y}^{(\ell)}| \rangle. \quad (51)$$

Since the two inequalities are of the same flavor, it suffices to prove only the first one. Note that it is here, where we use the exact definition of $\boldsymbol{F}$ as follows.

$$\begin{aligned}
\sum_{\ell} |\langle \mathbf{x}^{(j)}, \boldsymbol{F} \mathbf{y}^{(\ell)} \rangle| &\le \langle |\mathbf{x}^{(j)}|, |\boldsymbol{F}| \sum_{\ell} |\mathbf{y}^{(\ell)}| \rangle \\
&\le \langle |\mathbf{x}^{(j)}|, (\boldsymbol{C} + \boldsymbol{D}_{row} \boldsymbol{R} \boldsymbol{D}_{col}) \mathbf{1}_n \rangle | = 2 \langle |\mathbf{x}^{(j)}|, \boldsymbol{C} \mathbf{1}_n \rangle
\end{aligned}$$

because $|\mathbf{y}^{(\ell)}|$ is a 0-1 vector and $\boldsymbol{C} + \boldsymbol{D}_{row} \boldsymbol{R} \boldsymbol{D}_{col}$ is a (real) matrix of non-negative entries. We also used that the i-th coordinate of the vector $(\boldsymbol{C} +$

$\boldsymbol{D}_{row}\boldsymbol{R}\boldsymbol{D}_{col})\mathbf{1}_n$ for $i \in R_a$ is

$$d_{row,i}\left(1+\sum_{b=1}^{k}\rho(R_a,C_b)\mathrm{Vol}(C_b)\right)=2d_{row,i}$$

(here we utilized that the sum of the entries of $\boldsymbol{C}$ is 1), and therefore,

$$(\boldsymbol{C}+\boldsymbol{D}_{row}\boldsymbol{R}\boldsymbol{D}_{col})\mathbf{1}_n=2\boldsymbol{C}\mathbf{1}_n.$$

Finally, we will finish the proof with similar considerations as in [24]. Let us further estimate

$$\langle \mathbf{x},\boldsymbol{F}\mathbf{y}\rangle=\sum_j\sum_\ell\langle(\frac{4}{5})^j\mathbf{x}^{(j)},\boldsymbol{F}(\frac{4}{5})^\ell\mathbf{y}^{(\ell)}\rangle.$$

Put $\gamma:=\log_{4/5}\alpha^*$; in view of $\alpha^*<1$, $\gamma>0$ holds. Then we divide the above summation into three parts as follows.

$$|\langle \mathbf{x},\boldsymbol{F}\mathbf{y}\rangle|\le\sum_j\sum_\ell(\frac{4}{5})^{j+\ell}|\langle\mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle|$$
$$=\underbrace{\sum_{|j-\ell|\le\gamma}(\frac{4}{5})^{j+\ell}|\langle\mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle|}_{\text{(a)}}+\underbrace{\sum_{j-\ell>\gamma}(\frac{4}{5})^{j+\ell}|\langle\mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle|}_{\text{(b)}}+\underbrace{\sum_{j-\ell<-\gamma}(\frac{4}{5})^{j+\ell}|\langle\mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle|}_{\text{(c)}}.$$

The three terms are estimated separately. Term (a) can be bounded from above as follows:

$$\begin{aligned}\sum_{|j-\ell|\le\gamma}(\frac{4}{5})^{j+\ell}|\langle\mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle| &\overset{(50)}{\le} 2k\alpha^*\sum_{|j-\ell|\le\gamma}\sqrt{(\frac{4}{5})^{2j}\langle|\mathbf{x}^{(j)}|,\boldsymbol{C}\mathbf{1}_n\rangle(\frac{4}{5})^{2\ell}\langle\mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle}\\ &\overset{(*)}{\le} k\alpha^*\sum_{|j-\ell|\le\gamma}\left[(\frac{4}{5})^{2j}\langle|\mathbf{x}^{(j)}|,\boldsymbol{C}\mathbf{1}_n\rangle+(\frac{4}{5})^{2\ell}\langle\mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle\right]\\ &\overset{(**)}{\le} k\alpha^*(2\gamma+1)\left[\sum_j(\frac{4}{5})^{2j}\langle|\mathbf{x}^{(j)}|,\boldsymbol{C}\mathbf{1}_n\rangle+\sum_\ell(\frac{4}{5})^{2\ell}\langle\mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle\right],\\ &\overset{(***)}{\le} 2k\alpha^*(2\gamma+1),\end{aligned}$$

where in the first inequality, the estimate of (50), and in (*), the geometric-arithmetic mean inequality were used; (**) comes from the fact that in the second line, the first term depends merely on j, while the second one merely on ℓ, and so, for fixed j or ℓ, any term can show up at most $2\gamma+1$ times; (***) is due to the easy observation that

$$\sum_j(\frac{4}{5})^{2j}\langle|\mathbf{x}^{(j)}|,\boldsymbol{C}\mathbf{1}_n\rangle=\|\boldsymbol{D}_{row}^{1/2}\mathbf{x}\|^2\le 1,\quad \sum_\ell(\frac{4}{5})^{2\ell}\langle\mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle=\|\boldsymbol{D}_{col}^{1/2}\mathbf{y}\|^2\le 1. \tag{52}$$

Terms (b) and (c) are of similar appearance (the role of j and ℓ is symmetric in them), therefore, we will estimate only (b). Here $j-\ell>\gamma$, yielding $j+\ell>2\ell+\gamma$.

Therefore,

$$\sum_{j-\ell>\gamma}(\frac{4}{5})^{j+\ell}|\langle \mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle| \le \sum_{\ell}(\frac{4}{5})^{2\ell+\gamma}\sum_{j}|\langle \mathbf{x}^{(j)},\boldsymbol{F}\mathbf{y}^{(\ell)}\rangle| \overset{(51)}{\le} \sum_{\ell}(\frac{4}{5})^{2\ell+\gamma}2\langle \mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle$$
$$= 2(\frac{4}{5})^{\gamma}\sum_{\ell}(\frac{4}{5})^{2\ell}\langle \mathbf{1}_m,\boldsymbol{C}|\mathbf{y}^{(\ell)}|\rangle \overset{(52)}{\le} 2(\frac{4}{5})^{\gamma}$$

where, in the second and third inequalities, (51) and (52) were used. Consequently, (c) can also be estimated from above with $2(\frac{4}{5})^{\gamma}$.

Collecting the so obtained estimates together, we get

$$s_k \le \frac{9}{2}|\langle \mathbf{x},\boldsymbol{F}\mathbf{y}\rangle| \le \frac{9}{2}\left[2k\alpha^*(2\gamma+1)+4(\frac{4}{5})^{\gamma}\right] = 9\alpha^*\left[2k\frac{\ln\alpha^*}{\ln\frac{4}{5}}+k+2\right]$$
$$\le 9\alpha^*[2k(-4.5)\ln\alpha^*+k+2] = 9\alpha^*(k+2-9k\ln\alpha^*),$$

that was to be proved. □

Note that for $k=1$, our upper bound is tighter than that of (45), see Theorem 2 of [24].

Observe that for small discrepancies, the right-hand side of (46) is a strictly increasing function of $\mathrm{md}_k(\boldsymbol{C})$ when it is 'small'. Actually, the same function of $\mathrm{md}(\boldsymbol{C};R_1,\dots,R_k,C_1,\dots,C_k)$ is also a valid upper estimate for s_k whenever the row-partitions $R_1,\dots,R_k$ and the column-partitions $C_1,\dots,C_k$ are such that $\mathrm{md}(\boldsymbol{C};R_1,\dots,R_k,C_1,\dots,C_k)<1$ holds. Since the function $f(x)=9x(k+2-9k\ln x)$ is strictly increasing near zero, $\mathrm{md}_k(\boldsymbol{C})$ is the best upper estimate.

5.2 Estimating the multiway discrepancy of contingency tables by the singular values and subspace deviations of the normalized table

In the forward direction, we did not manage to estimate the k-way discrepancy from above merely by means of the k-th largest non-trivial singular value of the normalized table, but had to use the k-variances of the optimal $(k-1)$-dimensional row- and column-representatives too. In the proof, we applied a bit different notion of the multiway discrepancy, but at the end, we will discuss the relation between it and that of Definition 10. Actually, we used a notion similar to that of the volume regularity introduced in Alon and coauthors [6], where the authors also give an algorithm that computes a regular partition of a given (possibly sparse) graph in polynomial time giving some kind of construction for the Szemerédi Regularity Lemma.

Definition 11. The row–column cluster pair $R\subset Row$, $C\subset Col$ of the contingency table $\boldsymbol{C}$ of total volume 1 is α-volume regular if for every $X\subset R$ and $Y\subset C$ the relation

$$|c(X,Y)-\rho(R,C)\mathrm{Vol}(X)\mathrm{Vol}(Y)| \le \alpha\sqrt{\mathrm{Vol}(R)\mathrm{Vol}(C)} \tag{53}$$

holds, where $\rho(R,C)$ is the relative inter-cluster density of the row–column pair R,C, introduced in Definition 10.

Theorem 8. *Let $\boldsymbol{C}$ be a non-degenerate contingency table of m rows and n columns, with row- and column sums $d_{row,1},\dots,d_{row,m}$ and $d_{col,1},\dots,d_{col,n}$, respectively. Assume that $\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}=1$ and there are no dominant rows and columns: $d_{row,i}=\Theta(\frac{1}{m})$, $i=1,\dots,m$ and $d_{col,j}=\Theta(\frac{1}{n})$, $j=1,\dots,n$ as $m,n\to\infty$. Let the singular values of $\boldsymbol{C}_D$ be*

$$1=s_0>s_1\geq\cdots\geq s_{k-1}>\varepsilon\geq s_i,\quad i\geq k.$$

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 $\tilde{S}_k^2(\boldsymbol{X})$ and $\tilde{S}_k^2(\boldsymbol{Y})$ of the optimal row and column representatives collected in $\boldsymbol{X}$ and $\boldsymbol{Y}$. Assume 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 $\mathcal{O}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})\tilde{S}_k(\boldsymbol{Y}))+\varepsilon)$-volume regular $(i,j=1,\dots,k)$.

For the proof, we need the definition of the cut-norm and the relation between it and the spectral norm (see also [37]).

Definition 12. The cut-norm of the real matrix $\boldsymbol{A}$ with row-set *Row* and column-set *Col* is

$$\|\boldsymbol{A}\|_\square=\max_{R\subset Row,\,C\subset Col}\left|\sum_{i\in R}\sum_{j\in C}a_{ij}\right|.$$

Lemma 6. *For the $m\times n$ real matrix $\boldsymbol{A}$,*

$$\|\boldsymbol{A}\|_\square\leq\sqrt{mn}\|\boldsymbol{A}\|,$$

where the right hand side contains the spectral norm, i.e., the largest singular value of $\boldsymbol{A}$.

Proof of Lemma 6.

$$\|\boldsymbol{A}\|_\square=\max_{\mathbf{x}\in\{0,1\}^m,\,\mathbf{y}\in\{0,1\}^n}|\mathbf{x}^T\boldsymbol{A}\mathbf{y}|=\max_{\mathbf{x}\in\{0,1\}^m,\,\mathbf{y}\in\{0,1\}^n}\left|(\frac{\mathbf{x}}{\|\mathbf{x}\|})^T\boldsymbol{A}(\frac{\mathbf{y}}{\|\mathbf{y}\|})\right|\cdot\|\mathbf{x}\|\cdot\|\mathbf{y}\||$$
$$\leq\sqrt{mn}\max_{\|\mathbf{x}\|=1,\,\|\mathbf{y}\|=1}|\mathbf{x}^T\boldsymbol{A}\mathbf{y}|=\sqrt{mn}\|\boldsymbol{A}\|,$$

since for $\mathbf{x}\in\{0,1\}^m$, $\|\mathbf{x}\|\leq\sqrt{m}$, and for $\mathbf{y}\in\{0,1\}^n$, $\|\mathbf{y}\|\leq\sqrt{n}$, that finishes the proof. □

The definition of the cut-norm and the result of the above lemma naturally extends to symmetric matrices with $m=n$. Note that B. Szegedy estimates the cut-norm of a graphon from above by the spectral norm of the corresponding compact operator. Since our normalization is for matrices and not for graphons, the estimate of Lemma 6 does contain the size of the matrix.

Proof of Theorem 8. Let $\boldsymbol{C}_D=\sum_{i=0}^{r-1}s_i\mathbf{v}_i\mathbf{u}_i^T$ be SVD, where $r=\text{rank}(\boldsymbol{C})=\text{rank}(\boldsymbol{C}_D)$. Recall that provided $\boldsymbol{C}$ is non-degenerate, the largest singular value $s_0=1$ of $\boldsymbol{C}_D$ is single with corresponding singular vector pair $\mathbf{v}_0=\boldsymbol{D}_{row}^{1/2}\mathbf{1}_m$ and $\mathbf{u}_0=\boldsymbol{D}_{col}^{1/2}\mathbf{1}_n$, respectively. The optimal k-dimensional representatives of the rows and columns are row vectors of the matrices $\boldsymbol{X}=(\mathbf{x}_0,\dots,\mathbf{x}_{k-1})$ and $\boldsymbol{Y}=(\mathbf{y}_0,\dots,\mathbf{y}_{k-1})$, where $\mathbf{x}_i=\boldsymbol{D}_{row}^{-1/2}\mathbf{v}_i$ and $\mathbf{y}_i=\boldsymbol{D}_{col}^{-1/2}\mathbf{u}_i$, respectively

($i = 0, \dots, k-1$). (Note that the first columns of equal coordinates can as well be omitted.) Assume that the minimum weighted k-variance is attained at the k-partition $(R_1, \dots, R_k)$ of the rows and $(C_1, \dots, C_k)$ of the columns, respectively. By the usual analysis of variance argument, it follows that

$$\tilde{S}_k^2(\boldsymbol{X}) = \sum_{i=0}^{k-1} \mathrm{dist}^2(\mathbf{v}_i, F), \quad \tilde{S}_k^2(\boldsymbol{Y}) = \sum_{i=0}^{k-1} \mathrm{dist}^2(\mathbf{u}_i, G),$$

where $F = \mathrm{Span}\{\boldsymbol{D}_{row}^{1/2}\mathbf{w}_1, \dots, \boldsymbol{D}_{row}^{1/2}\mathbf{w}_k\}$ and $G = \mathrm{Span}\{\boldsymbol{D}_{col}^{1/2}\mathbf{z}_1, \dots, \boldsymbol{D}_{col}^{1/2}\mathbf{z}_k\}$ with the so-called normalized row partition vectors $\mathbf{w}_1, \dots, \mathbf{w}_k$ of coordinates $w_{ji} = \frac{1}{\sqrt{\mathrm{Vol}(R_i)}}$ if $j \in R_i$ and 0, otherwise; and column partition vectors $\mathbf{z}_1, \dots, \mathbf{z}_k$ of coordinates $z_{ji} = \frac{1}{\sqrt{\mathrm{Vol}(C_i)}}$ if $j \in C_i$ and 0, otherwise ($i = 1, \dots, k$). Note that the vectors $\boldsymbol{D}_{row}^{1/2}\mathbf{w}_1, \dots, \boldsymbol{D}_{row}^{1/2}\mathbf{w}_k$ and $\boldsymbol{D}_{col}^{1/2}\mathbf{z}_1, \dots, \boldsymbol{D}_{col}^{1/2}\mathbf{z}_k$ form orthonormal systems in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively (but they are, usually, not complete). By Lemma 2, we can find orthonormal systems $\tilde{\mathbf{v}}_0, \dots, \tilde{\mathbf{v}}_{k-1} \in F$ and $\tilde{\mathbf{u}}_0, \dots, \tilde{\mathbf{u}}_{k-1} \in G$ such that

$$\tilde{S}_k^2(\boldsymbol{X}) \le \sum_{i=0}^{k-1} \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\|^2 \le 2\tilde{S}_k^2(\boldsymbol{X}), \quad \tilde{S}_k^2(\boldsymbol{Y}) \le \sum_{i=0}^{k-1} \|\mathbf{u}_i - \tilde{\mathbf{u}}_i\|^2 \le 2\tilde{S}_k^2(\boldsymbol{Y}). \quad (54)$$

We approximate $\boldsymbol{C}_D$ by the rank k matrix $\sum_{i=0}^{k-1} s_i \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T$ with the following accuracy (in spectral norm):

$$\left\| \sum_{i=0}^{r-1} s_i \mathbf{v}_i \mathbf{u}_i^T - \sum_{i=0}^{k-1} s_i \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T \right\| \le \sum_{i=0}^{k-1} s_i \left\| \mathbf{v}_i \mathbf{u}_i^T - \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T \right\| + \left\| \sum_{i=k}^{r-1} s_i \mathbf{v}_i \mathbf{u}_i^T \right\|, \quad (55)$$

where the spectral norm of the last term is at most ε, and the individual terms of the first one are estimated from above in the following way.

$$\begin{aligned} s_i \|\mathbf{v}_i \mathbf{u}_i^T - \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T\| &\le \|(\mathbf{v}_i \mathbf{u}_i^T - \tilde{\mathbf{v}}_i \mathbf{u}_i^T) + (\tilde{\mathbf{v}}_i \mathbf{u}_i^T - \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T)\| \\ &\le \|(\mathbf{v}_i - \tilde{\mathbf{v}}_i)\mathbf{u}_i^T\| + \|\tilde{\mathbf{v}}_i (\mathbf{u}_i - \tilde{\mathbf{u}}_i)^T\| \\ &= \sqrt{\|(\mathbf{v}_i - \tilde{\mathbf{v}}_i)\mathbf{u}_i^T \mathbf{u}_i (\mathbf{v}_i - \tilde{\mathbf{v}}_i)^T\|} + \sqrt{\|(\mathbf{u}_i - \tilde{\mathbf{u}}_i)\tilde{\mathbf{v}}_i^T \tilde{\mathbf{v}}_i (\mathbf{u}_i - \tilde{\mathbf{u}}_i)^T\|} \\ &= \sqrt{(\mathbf{v}_i - \tilde{\mathbf{v}}_i)^T (\mathbf{v}_i - \tilde{\mathbf{v}}_i)} + \sqrt{(\mathbf{u}_i - \tilde{\mathbf{u}}_i)^T (\mathbf{u}_i - \tilde{\mathbf{u}}_i)} \\ &= \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\| + \|\mathbf{u}_i - \tilde{\mathbf{u}}_i\| \end{aligned}$$

where we exploited that the spectral norm (i.e., the largest singular value) of an $m \times n$ matrix $\boldsymbol{A}$ is equal to either the squareroot of the largest eigenvalue of the matrix $\boldsymbol{A}\boldsymbol{A}^T$ or equivalently, that of $\boldsymbol{A}^T\boldsymbol{A}$. In the above calculations all of these matrices are of rank 1, hence, the largest eigenvalue of the symmetric, positive semidefinite matrix under the squareroot is the only non-zero eigenvalue of it, therefore, it is equal to its trace; finally, we used the commutativity of the trace, and in the last line we have the usual vector norm.

Therefore, the first term in (55) can be estimated from above with

$$\begin{aligned} \sum_{i=0}^{k-1} \|\mathbf{v}_i \mathbf{u}_i^T - \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T\| &\le \sqrt{k} \sqrt{\sum_{i=0}^{k-1} \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\|^2} + \sqrt{k} \sqrt{\sum_{i=0}^{k-1} \|\mathbf{u}_i - \tilde{\mathbf{u}}_i\|^2} \\ &\le \sqrt{k}\left(\sqrt{2\tilde{S}_k^2(\boldsymbol{X})} + \sqrt{2\tilde{S}_k^2(\boldsymbol{Y})}\right) = \sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})), \end{aligned}$$

where we also used the upper estimate of (54).

Based on these considerations and relation between the cut-norm and the spectral norm, the densities to be estimated in the defining formula (53) of volume regularity can be written in terms of step-vectors in the following way. The vectors $\hat{\mathbf{v}}_i := \boldsymbol{D}_{row}^{-1/2}\tilde{\mathbf{v}}_i$ are stepwise constant on the partition $(R_1,\dots,R_k)$ of the rows, whereas the vectors $\hat{\mathbf{u}}_i := \boldsymbol{D}_{col}^{-1/2}\tilde{\mathbf{u}}_i$ are stepwise constant on the partition $(C_1,\dots,C_k)$ of the columns, $i=0,\dots,k-1$. The matrix

$$\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T$$

is therefore an $n\times m$ block-matrix on $k\times k$ blocks corresponding to the above partition of the rows and columns. Let $\hat{c}_{ab}$ denote its entries in the ab block $(a,b=1,\dots,k)$. Using (55), the rank k approximation of the matrix $\boldsymbol{C}$ is performed with the following accuracy of the perturbation $\boldsymbol{E}$ in spectral norm:

$$\|\boldsymbol{E}\| = \left\|\boldsymbol{C} - \boldsymbol{D}_{row}(\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T)\boldsymbol{D}_{col}\right\| = \left\|\boldsymbol{D}_{row}^{1/2}(\boldsymbol{C}_D - \sum_{i=0}^{k-1} s_i \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T)\boldsymbol{D}_{col}^{1/2}\right\|.$$

Therefore, the entries of $\boldsymbol{C}$ can be decomposed as

$$c_{ij} = d_{row,i} d_{col,j} \hat{c}_{ab} + \eta_{ij} \qquad (i\in R_a, \quad j\in C_b)$$

where the cut-norm of the $n\times m$ error matrix $\boldsymbol{E}=(\eta_{ij})$ restricted to $R_a\times C_b$ (otherwise it contains entries all zeros) and denoted by $\boldsymbol{E}_{ab}$, is estimated as follows. Making use of Lemma 6,

$$\begin{aligned}
\|\boldsymbol{E}_{ab}\|_\square &\le \sqrt{mn}\|\boldsymbol{E}_{ab}\| \le \sqrt{nm}\cdot\|\boldsymbol{D}_{row,a}^{1/2}\|\cdot(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})+\tilde{S}_k(\boldsymbol{Y}))+\varepsilon)\cdot\|\boldsymbol{D}_{col,b}^{1/2}\| \\
&\le \sqrt{nm}\sqrt{c_1\frac{\mathrm{Vol}(R_a)}{|R_a|}}\cdot\sqrt{c_2\frac{\mathrm{Vol}(C_b)}{|C_b|}}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})+\tilde{S}_k(\boldsymbol{Y}))+\varepsilon) \\
&= \sqrt{c_1c_2}\cdot\sqrt{\frac{n}{|R_a|}}\cdot\sqrt{\frac{m}{|C_b|}}\cdot\sqrt{\mathrm{Vol}(R_a)}\sqrt{\mathrm{Vol}(C_b)}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})+\tilde{S}_k(\boldsymbol{Y}))+\varepsilon) \\
&\le \sqrt{\frac{c_1c_2}{K_1K_2}}\sqrt{\mathrm{Vol}(R_a)}\sqrt{\mathrm{Vol}(C_b)}(\sqrt{2k}s+\varepsilon) \\
&= c\sqrt{\mathrm{Vol}(R_a)}\sqrt{\mathrm{Vol}(C_b)}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})+\tilde{S}_k(\boldsymbol{Y}))+\varepsilon)
\end{aligned}$$

where the $n\times n$ diagonal matrix $\boldsymbol{D}_{row,a}$ inherits $\boldsymbol{D}_{row}$'s diagonal entries over R_a, whereas the $m\times m$ diagonal matrix $\boldsymbol{D}_{col,b}$ inherits $\boldsymbol{D}_{col}$'s diagonal entries over C_b, otherwise they are zeros. Further, the constants c_1, c_2 are due to the fact that there are no dominant rows and columns, while K_1, K_2 are from the cluster size balancing conditions. Hence,

$$\|\boldsymbol{E}_{ab}\|_\square \le c\sqrt{\mathrm{Vol}(R_a)}\sqrt{\mathrm{Vol}(C_b)}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X})+\tilde{S}_k(\boldsymbol{Y}))+\varepsilon)$$

where the constant c does not depend on n and m. Consequently, for $a,b=$

$1, \dots, k$ and $X \subset R_a$, $Y \subset C_b$,

$$|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)| =$$

$$\left| \sum_{i\in X} \sum_{j \in Y} (d_{row,i} d_{col,j} \hat{c}_{ab} + \eta_{ij}) - \frac{\mathrm{Vol}(X)\mathrm{Vol}(Y)}{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)} \sum_{i \in R_a} \sum_{j \in C_b} (d_{row,i} d_{col,j} \hat{c}_{ab} + \eta_{ij}) \right| =$$

$$\left| \sum_{i\in X} \sum_{j \in Y} \eta_{ij} - \frac{\mathrm{Vol}(X)\mathrm{Vol}(Y)}{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)} \sum_{i \in R_a} \sum_{j \in C_b} \eta_{ij} \right| \le 2\|\boldsymbol{E}_{ab}\|_\square$$

$$\le 2c(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + \varepsilon)\sqrt{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$$

that gives the required statement for $a, b = 1, \dots, k$, and finishes the proof. □

So we managed to prove the following. Given the $m \times n$ contingency table $\boldsymbol{C}$, consider the spectral clusters $R_1, \dots, R_k$ of its rows and $C_1, \dots, C_k$ of its columns, obtained by applying the weighted k-means algorithm to the $(k-1)$-dimensional row- and column representatives, defined as the row vectors of the matrices $(\boldsymbol{D}_{row}^{-1/2}\mathbf{v}_1, \dots, \boldsymbol{D}_{row}^{-1/2}\mathbf{v}_{k-1})$ and $(\boldsymbol{D}_{col}^{-1/2}\mathbf{u}_1, \dots, \boldsymbol{D}_{col}^{-1/2}\mathbf{u}_{k-1})$, respectively, where $\mathbf{v}_i, \mathbf{u}_i$ is the unit norm singular vector pair corresponding to s_i $(i = 1, \dots, k-1)$. In fact, these partitions minimize the weighted k-variances $\tilde{S}_k^2(\boldsymbol{X})$ and $\tilde{S}_k^2(\boldsymbol{Y})$ of these row- and column-representatives. Then, under some balancing conditions for $d_{row,i}$'s and $d_{col,j}$'s (there are no dominant rows and columns) and for the cluster sizes, we proved that $\mathrm{md}'_k(\boldsymbol{C}) = \mathcal{O}(\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + s_k)$, where $\mathrm{md}'_k(\boldsymbol{C})$ is a somewhat modified version of the k-way discrepancy; the only difference is that in the definition of $\mathrm{md}'_k(\boldsymbol{C})$ we substitute $\sqrt{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$ for $\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}$ in the denominator of (43). In accordance with the original definition of the discrepancy in the Szemerédi Regularity Lemma [64] for simple graphs, in (43), we may take the maximum over subsets $X \subset V_a$, $Y \subset V_b$ such that $\mathrm{Vol}(X) \ge \epsilon \mathrm{Vol}(V_a)$ and $\mathrm{Vol}(Y) \ge \epsilon \mathrm{Vol}(V_b)$ with some fixed $\epsilon > 0$. If we impose similar conditions on the row- and column-subsets, our result also implies that $\mathrm{md}_k(\boldsymbol{C})$ is of order $\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + s_k$.

The message of Theorems 7 and 8 is that the k-way discrepancy, when it is 'small' enough, suppresses s_k. Conversely, s_k together with 'small' enough $\tilde{S}_k(\boldsymbol{X})$ and $\tilde{S}_k(\boldsymbol{Y})$ also suppresses the k-way discrepancy. By using perturbation theory of spectral subspaces, (in the framework of edge-weighted graphs), we also discuss that a 'large' gap between s_{k-1} and s_k suppresses $\tilde{S}_k(\boldsymbol{X})$ and $\tilde{S}_k(\boldsymbol{Y})$. Therefore, if we want to find row–column cluster pairs of small discrepancy, we must select a k such that there is a remarkable gap between s_{k-1} and s_k; further s_k is small enough. Moreover, by using this k and the construction in the proof of the forward statement of Theorem 8, we are able to find these clusters with spectral clustering tools. It makes sense, for example, when we want to find clusters of genes and conditions simultaneously in microarrays so that genes of the same row-cluster would 'equally' influence conditions of the same column-cluster.

We also remark the following. When we perform correspondence analysis on a large $m \times n$ contingency table and consider the rank k approximation of it, the entries of this matrix will not necessarily be positive at all. Nonetheless, the entries $\hat{c}_{ij}$'s of the block-matrix constructed in the proof of Theorem 8 will already be positive provided the weighted k-variances $\tilde{S}_k(\boldsymbol{X})$ and $\tilde{S}_k(\boldsymbol{Y})$ are 'small' enough. Let us discuss this issue more precisely.

In accord with the notation used in the proof, denote by ab in the lower index if the matrix is restricted to the $R_a \times C_b$ block (otherwise it has zero entries). Then for the squared Frobenius norm of the rank k approximation of $\boldsymbol{D}_{row}^{-1}\boldsymbol{C}\boldsymbol{D}_{col}^{-1}$, restricted to the ab block, we have that

$$\begin{aligned} &\left\| \boldsymbol{D}_{row,a}^{-1}\boldsymbol{C}_{ab}\boldsymbol{D}_{col,b}^{-1} - (\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T)_{ab} \right\|_2^2 = \sum_{i\in R_a}\sum_{j\in C_b} (\frac{c_{ij}}{d_{row,i}d_{col,j}} - \hat{c}_{ab})^2 \\ &= \sum_{i\in R_a}\sum_{j\in C_b} (\frac{c_{ij}}{d_{row,i}d_{col,j}} - \bar{c}_{ab})^2 + |R_a||C_b|(\bar{c}_{ab} - \hat{c}_{ab})^2 \end{aligned} \tag{56}$$

where we used the Steiner equality with the average $\bar{c}_{ab}$ of the entries of $\boldsymbol{D}_{row}^{-1}\boldsymbol{C}\boldsymbol{D}_{col}^{-1}$ in the ab block. Now we estimate the above Frobenius norm by a constant multiple of the spectral norm, where for the spectral norm

$$\begin{aligned} &\left\| \boldsymbol{D}_{row,a}^{-1}\boldsymbol{C}_{ab}\boldsymbol{D}_{col,b}^{-1} - (\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T)_{ab} \right\| = \left\| \boldsymbol{D}_{row,a}^{-1/2}(\boldsymbol{C}_{corr} - \sum_{i=0}^{k-1} s_i \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T)_{ab}\boldsymbol{D}_{col,b}^{-1/2} \right\| \\ &\le \max_{i\in R_a} \frac{1}{\sqrt{d_{row,i}}} \cdot \max_{j\in C_b} \frac{1}{\sqrt{d_{col,j}}} \cdot [\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + \varepsilon] \end{aligned}$$

holds. Therefore,

$$\begin{aligned} &\left\| \boldsymbol{D}_{row,a}^{-1}\boldsymbol{C}_{ab}\boldsymbol{D}_{col,b}^{-1} - (\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T)_{ab} \right\|_2^2 \\ &\le \min\{|R_a|, |C_b|\} \cdot \max_{i\in R_a} \frac{1}{d_{row,i}} \cdot \max_{j\in C_b} \frac{1}{d_{col,j}} \cdot [\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + \varepsilon]^2. \end{aligned}$$

Consequently, in view of (56),

$$(\bar{c}_{ab} - \hat{c}_{ab})^2 \le \frac{1}{\max\{|R_a|, |C_b|\}} \cdot \max_{i\in R_a} \frac{1}{d_{row,i}} \cdot \max_{j\in C_b} \frac{1}{d_{col,j}} \cdot [\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + \varepsilon]^2.$$

But using the conditions on the block sizes and the row- and column-sums of Theorem 8, provided

$$\sqrt{2k}(\tilde{S}_k(\boldsymbol{X}) + \tilde{S}_k(\boldsymbol{Y})) + \varepsilon) = \mathcal{O}\left(\frac{1}{(\min\{m,n\})^{\frac{1}{2}+\tau}} \right)$$

holds with some 'small' $\tau > 0$, the relation $\bar{c}_{ab} - \hat{c}_{ab} \to 0$ also holds as $n, m \to \infty$. Therefore, both $\hat{c}_{ab}$ and $\hat{c}_{ab} d_{row,i} d_{col,j}$ are positive over such blocks that are not constantly zero in the original table if m and n are large enough.

5.3 Multiway discrepancy of undirected graphs

The notion of multiway discrepancy naturally extends to edge-weighted graphs.

Definition 13. The multiway discrepancy of the undirected, edge-weighted graph $G = (V, \boldsymbol{W})$ in the proper k-partition $(V_1, \dots, V_k)$ of its vertices is

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

The minimum k-way discrepancy of the undirected edge-weighted graph $G = (V, \boldsymbol{W})$ is

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

A result, analogous to that of Theorem 7 can now be proved in terms of the normalized modularity matrix of G.

Theorem 9 ([16])**.** *Let $G = (V, \boldsymbol{W})$ be an edge-weighted, undirected graph, $\boldsymbol{W}$ is irreducible. Then for any integer $1 \le k < \mathrm{rank}(\boldsymbol{W})$,*

$$|\mu_k| \le 9\mathrm{md}_k(G)(k + 2 - 9k \ln \mathrm{md}_k(G)) \tag{57}$$

holds, provided $0 < \mathrm{md}_k(G) < 1$, where μ_k is the k-th largest absolute value eigenvalue of the normalized modularity matrix $\boldsymbol{M}_D$ of G.

Proof (of Theorem 9). The proof follows the same considerations as the proof of Theorem 7 with the difference that here we use symmetric matrices. In particular, $\boldsymbol{R} = (\rho(V_a, V_b))$ is an $n \times n$ symmetric block-matrix of $k \times k$ blocks corresponding to the partition $V_1, \dots, V_k$ of the vertices for which $\alpha^* = \mathrm{md}_k(G) = \mathrm{md}(G; V_1, \dots, V_k)$; consequently, the matrix $\boldsymbol{F} = \boldsymbol{W} - \boldsymbol{DRD}$ is also symmetric. Therefore, in accord with (49) and Definition 13: for every V_a, V_b pair and $X \subset V_a$, $Y \subset V_b$ $(1 \le a \le b \le k)$ we have

$$|\langle \mathbf{1}_X, \boldsymbol{F}\mathbf{1}_Y \rangle| \le \alpha^* \sqrt{\langle \mathbf{1}_X, \boldsymbol{W}\mathbf{1}_n \rangle \langle \mathbf{1}_n, \boldsymbol{W}\mathbf{1}_Y \rangle}. \tag{58}$$

The left and right singular vectors ($\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$) corresponding to the maximal singular value of the real symmetric matrix $\boldsymbol{D}^{-1/2}\boldsymbol{F}\boldsymbol{D}^{-1/2}$ satisfy $\mathbf{u} = \pm\mathbf{v}$ (the sign is the same as the sign of the eigenvalue of the maximal absolute value). If $\mathbf{u} = \mathbf{v}$, then $\mathcal{Y}_{\ell bq} = \mathcal{X}_{\ell bq}$ for every ℓ, $b = 1, \dots, k$, and $q = 1, 2$. If $\mathbf{u} = -\mathbf{v}$, then $\mathcal{Y}_{\ell b1} = \mathcal{X}_{\ell b2}$ and $\mathcal{Y}_{\ell b2} = \mathcal{X}_{\ell b1}$ for every ℓ and $b = 1, \dots, k$. Consequently, in the estimates of (50), when we use the absolute values of the coordinates of the vectors $\mathbf{x}^{(j)}$ and $\mathbf{y}^{(\ell)}$, constructed in the rectangular case, the inequalities remain valid. Namely,

$$|\langle \mathbf{x}^{(j)}, \boldsymbol{F}\mathbf{y}^{(\ell))} \rangle| \le \sum_{a=1}^{k} \sum_{p=1}^{2} \sum_{b=1}^{k} \sum_{q=1}^{2} \left| \langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{F}\mathbf{1}_{\mathcal{X}_{\ell bq}} \rangle \right| .$$

Here the summation is for every $1 \le a, b \le k$ pair. However, if $a \le b$, then by (58) we get

$$\left| \langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{F}\mathbf{1}_{\mathcal{X}_{\ell bq}} \rangle \right| \le \alpha^* \sqrt{\langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{W}\mathbf{1}_n \rangle \langle \mathbf{1}_n, \boldsymbol{W}\mathbf{1}_{\mathcal{X}_{\ell bq}} \rangle};$$

whereas, if $a > b$, then by the symmetry of $\boldsymbol{F}$:

$$\begin{aligned} \left| \langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{F}\mathbf{1}_{\mathcal{X}_{\ell bq}} \rangle \right| &= \left| \langle \mathbf{1}_{\mathcal{X}_{\ell bq}}, \boldsymbol{F}\mathbf{1}_{\mathcal{X}_{jap}} \rangle \right| \le \alpha^* \sqrt{\langle \mathbf{1}_{\mathcal{X}_{lbq}}, \boldsymbol{W}\mathbf{1}_n \rangle \langle \mathbf{1}_n, \boldsymbol{W}\mathbf{1}_{\mathcal{X}_{jap}} \rangle} \\ &= \sqrt{\langle \mathbf{1}_{\mathcal{X}_{jap}}, \boldsymbol{W}\mathbf{1}_n \rangle \langle \mathbf{1}_n, \boldsymbol{W}\mathbf{1}_{\mathcal{X}_{\ell bq}} \rangle}. \end{aligned}$$

Therefore, for $a \ne b$, the same term appears twice, and all the subsequent estimates remain valid by substituting $\boldsymbol{W}$ for $\boldsymbol{C}$ and $\boldsymbol{D}$ for both $\boldsymbol{D}_{row}$ and $\boldsymbol{D}_{col}$. This completes the proof. □

Recall that Bilu and Linial [9] proved the following converse of the expander mixing lemma for simple d-regular graphs on n vertices. Assume that for any disjoint vertex-subsets S, T: $|e(S,T) - \frac{|S||T|d}{n}| \le \alpha\sqrt{|S||T|}$. Then all but the largest adjacency eigenvalue of G are bounded (in absolute value) by $O(\alpha(1 + \log\frac{d}{\alpha}))$. Note that for a d-regular graph the adjacency eigenvalues are d times larger than the normalized adjacency ones, and the deviation between $e(S,T)$ and the one what is expected in a random d-regular graph, is also proportional to our (1-way) discrepancy in terms of the volumes (note that $\mathrm{Vol}(S)$ is also proportional to $|S|$). Though they use disjoint subsets S, T, their upper estimate for the absolute value of the second largest (in absolute value) eigenvalue with the (1-way) discrepancy α is $C\alpha(1 - A\log\alpha)$ with some absolute constants A, C. Hence, the upper estimate of (45) or that of (57) in the $k = 1$ case are reminiscent of this.

In the other direction, for the $k = 1$ case, a straightforward generalization of the *expander mixing lemma for irregular graphs* is the following.

Proposition 6 ([14]).

$$\mathrm{disc}(G) = \mathrm{md}_1(G) \le \|\boldsymbol{M}_D\| = s_1 = |\mu_1|,$$

where $\|\boldsymbol{M}_D\|$ *is the spectral norm of the normalized modularity matrix of* G.

Though, with different notation sometimes even a stronger version of this statement is proved in [**?**, 24, 29], we give another short proof here.

Proof. Via separation theorems for singular values, $s_1 = |\mu_1|$ is the maximum of the bilinear form $\mathbf{v}^T\boldsymbol{M}_D\mathbf{u}$ over the unit sphere. Let $X, Y \subset V$ be arbitrary, and denote by $\mathbf{1}_X, \mathbf{1}_Y \in \mathbb{R}^n$ the indicator vectors of them. Then

$$\|\boldsymbol{M}_D\| = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} |\mathbf{v}^T\boldsymbol{M}_D\mathbf{u}| \ge \left| \left(\frac{\boldsymbol{D}^{1/2}\mathbf{1}_X}{\|\boldsymbol{D}^{1/2}\mathbf{1}_X\|} \right)^T \boldsymbol{M}_D \left(\frac{\boldsymbol{D}^{1/2}\mathbf{1}_Y}{\|\boldsymbol{D}^{1/2}\mathbf{1}_Y\|} \right) \right|$$
$$= \frac{|\mathbf{1}_X^T\boldsymbol{M}\mathbf{1}_Y|}{\|\boldsymbol{D}^{1/2}\mathbf{1}_X\| \cdot \|\boldsymbol{D}^{1/2}\mathbf{1}_Y\|} = \frac{|w(X,Y) - \mathrm{Vol}(X)\mathrm{Vol}(Y)|}{\sqrt{\mathrm{Vol}(X)}\sqrt{\mathrm{Vol}(Y)}}.$$

Taking the maxima on the right-hand side over subsets $X, Y \subset V$, the desired statement follows. Note that the estimate is also valid if we take maxima over disjoint X, Y pairs only. □

For an arbitrary integer k, in the range $1 \le k < \mathrm{rank}(\boldsymbol{W})$, the following analogue of Theorem 8 for undirected, edge-weighted graph.

Theorem 10. *Let* $G = (V, \boldsymbol{W})$ *be an edge-weighted graph on* n *vertices, with generalized degrees* $d_1, \dots, d_n$ *and degree-matrix* $\boldsymbol{D}$*. Assume that* G *is connected,* $\mathrm{Vol}(V) = 1$*, and there are no dominant vertices:* $d_i = \Theta(1/n)$*,* $i = 1, \dots, n$ *as* $n \to \infty$*. Let the eigenvalues of the normalized modularity matrix* $\boldsymbol{M}_D$ *of* G*, enumerated in decreasing absolute values, be*

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

The partition $(V_1, \dots, V_k)$ *of* V *is defined so that it minimizes the weighted* k*-variance* $\tilde{S}_k^2(\boldsymbol{X}^*)$ *of the optimal vertex representatives obtained as row vectors of the* $n \times (k-1)$ *matrix* $\boldsymbol{X}^*$ *of column vectors* $\boldsymbol{D}^{-1/2}\mathbf{u}_i$*, where* $\mathbf{u}_i$ *is the unit-norm eigenvector corresponding to* μ_i $(i = 1, \dots, k-1)$*. Assume that there is* *a constant* $0 < K \le \frac{1}{k}$ *such that* $|V_i| \ge Kn$, $i = 1, \dots, k$. *With the notation* $\sigma_k = \sqrt{\tilde{S}_k^2(\boldsymbol{X}^*)}$, *the* (V_i, V_j) *pairs are* $\mathcal{O}(\sqrt{2k}\sigma_k + \varepsilon)$*-volume regular* $(i \ne j)$ *and for the clusters* V_i $(i = 1, \dots, k)$ *the following holds: for all* $X, Y \subset V_i$,

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

where $\rho(V_i) = \frac{w(V_i,V_i)}{\mathrm{Vol}^2(V_i)}$ *is the relative intra-cluster density of* V_i.

In fact, inspired by Alon et al. [6], in [**?**] we used a bit different notation and concept of α-volume regular pairs, namely, for every $X \subseteq V_a$, $Y \subseteq V_b$ we required

$$|w(X,Y) - \rho(V_a,V_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)| \le \alpha\sqrt{\mathrm{Vol}(V_a)\mathrm{Vol}(V_b)}.$$

In the above formula, the right had side contains the squareroots of the volumes of the clusters, unlike (44), which contains the squareroots of the volumes of X and Y. However, with the same argument as in the rectangular case: in the spirit of the Szemerédi Regularity Lemma [64], if we require (44) to hold only for X, Y's satisfying $\mathrm{Vol}(X) \ge \epsilon\mathrm{Vol}(V_i)$, $\mathrm{Vol}(Y) \ge \epsilon\mathrm{Vol}(V_j)$ with some fixed ϵ, then the so modified k-way discrepancy is $\mathcal{O}(\sqrt{2k}\sigma_k + |\mu_k|)$, and so does $\mathrm{md}_k(G)$.

In view of subspace perturbation theorems, the larger the gap between $|\mu_{k-1}|$ and $|\mu_k|$, the smaller σ_k is. So the message is, that here the eigenvectors corresponding to the largest absolute value eigenvalues have to be used, unlike usual spectral clustering methods which automatically use the bottom eigenvalues of the Laplacian or normalized Laplacian matrix (latter one is just $\boldsymbol{I} - \boldsymbol{W}_D$). The clusters or cluster-pairs of small discrepancy behave like expanders or bipartite expanders. In another context, they resemble the generalized random or quasirandom graphs of Lovász, Sós, Simonovits [55, 61].

Note that in the special two-cluster case, the 2-variance of the optimal one-dimensional representatives can be directly estimated from above by the gap between the two largest absolute value eigenvalues of $\boldsymbol{M}_D$, and hence, the statement of Theorem 10 simplifies as follows. The optimal pair (V_1, V_2) based on minimizing the weighted 2-variance of the coordinates of $\mathbf{u}_1$ is $\mathcal{O}(\sqrt{\frac{1-\delta}{1-\varepsilon}})$-volume regular, where $\delta = |\mu_1|$ and $\varepsilon = |\mu_2|$, provided $\boldsymbol{W}$ is non-degenerate (the underlying graph is connected, but not bipartite). With other methods, the same estimate is obtained in [14], where we treat the case when the two largest absolute value eigenvalues of the normalized modularity matrix are positive, though it can be adopted to the other situations too; see, for example, the dual Cheeger inequality of Luca Trevisan [**?**].

In some special cases, $\sigma_k = 0$, and then, $\mathrm{md}_k(G) \le B|\mu_k| = Bs_k$ follows from the above results. In particular, $\sigma_k = 0$ whenever the vectors $\boldsymbol{D}^{-1/2}\mathbf{u}_1, \dots, \boldsymbol{D}^{-1/2}\mathbf{u}_{k-1}$ are step-vectors over the same proper k-partition of the vertices. Some examples:

- If $k = 1$, then the unit-norm eigenvector corresponding to $\mu_0 = 1$ is $\mathbf{u}_0 = \sqrt{\mathbf{d}}$, and $\boldsymbol{D}^{-1/2}\mathbf{u}_0 = \mathbf{1}$ is the all 1's vector. Consequently, the variance of its coordinates is $\sigma_1 = 0$. But in this case, by Proposition 6, we already know that $\mathrm{disc}(G)$ can be estimated from above merely by $|\mu_1| = s_1$.

- If $k = 2$ and G is bipartite, then $\mu_1 = -1$, $s_1 = 1$, and σ_2^2, i.e., the 2-variance of the coordinates of the transformed eigenvector corresponding to μ_1 can be small if $|\mu_2|$ is separated from $|\mu_1| = 1$ (like the bipartite expanders of Alon [3]).

- Let $k = 2$ and G be *bipartite, biregular* on the independent vertex-subsets V_1, V_2. That is, all the edge-weights within V_1 or V_2 are zeros, and the 0-1 weights between vertices of V_1 and V_2 are such that $d_i = k_1$ if $i \in V_1$ and $d_i = k_2$ if $i \in V_2$ with the understanding that $|V_1|k_1 = |V_2|k_2$ (both are the total number of edges in G). It is easy to see that the unit-norm eigenvector corresponding to the eigenvalue $\mu_1 = -1$ is $\mathbf{u}_1 = \boldsymbol{D}^{1/2}\mathbf{1}_{V_1} - \boldsymbol{D}^{1/2}\mathbf{1}_{V_2}$, and $\boldsymbol{D}^{-1/2}\mathbf{u}_1 = \mathbf{1}_{V_1} - \mathbf{1}_{V_2}$. Therefore, the representatives of vertices of V_1 are all 1's, and those of V_2 are -1's, so $\sigma_2^2 = 0$. Consequently, $\mathrm{md}_2(G) \le B|\mu_2|$, with some absolute constant B. Up to the constant, this was another proof of Lemma 3.2 of Evra et al. [?]. They call their result expander mixing lemma for bipartite graphs, and use cardinalities instead of volumes, but in this special case, these cardinalities are proportional to the volumes both within V_1 and V_2.

- Let G_n be a generalized random graph (see Definition ??) over the symmetric $k \times k$ probability matrix $\boldsymbol{P} = (p_{ab})$, i.e., there is a proper k-partition, $V_1, \dots, V_k$, of its vertices such that $|V_a| = n_a$ $(a = 1, \dots, k)$, $\sum_{a=1}^k n_a = n$, and for any $1 \le a \le b \le k$, vertices $i \in V_a$ and $j \in V_b$ are connected independently, with the same probability p_{ab}. This is the k-cluster generalization of the classical Erdős–Rényi random graph, see also [55] for their generalized quasirandom counterparts. In [12] (see also Chapter ??) we characterize the adjacency and normalized Laplacian spectra of such graphs, that extends to their normalized modularity spectra as follows: both $|\mu_k| = s_k$ and σ_k^2 tend to zero almost surely when $n \to \infty$, under some balancing conditions for the cluster sizes ($\frac{n_a}{n} \ge c$ with some constant c, for $a = 1, \dots, k$). By Theorem 10, it also holds for the k-way discrepancy in the clustering $V_1, \dots, V_k$. However, this is not surprising, since this almost sure limit for the k-way discrepancy is easily obtained with large deviation principles too.

Summarizing, in the $k = 1$ case: when the second singular value $|\mu_1| = s_1$ is small (much smaller than $s_0 = 1$), then the overall discrepancy is small. However, for $k > 1$, a small s_k is necessary but not sufficient for a small k-way discrepancy. In addition, the weighted k-variance σ_k^2 should be small too. With subspace perturbation theorems, it is small if s_k is much smaller than s_{k-1}. Hence, a gap in the normalized modularity spectrum may be an indication for the number of clusters. The two directions together may give a hint about the optimal choice of k if a practitioner wants to find a k-clustering of the rows and columns (or just of the vertices of a graph) with small pairwise discrepancies. If there does not exist a fairly 'small' k with this property, then in the worst case scenario, the Szemerédi Regularity Lemma [64] with an enormously large number of clusters (which number only depends on the maximum pairwise discrepancy to be attained, and does not depend on n) comes into existence. Weak versions of this lemma (where $V_1, \dots, V_k$ are not necessarily equitable) are also available. Eventually, note that B. Szegedy and T. Tao (in his blog https://terrytao.wordpress.com/ 2012/12/03/the-spectral-proof-of-the-szemeredi-regularity-lemma/) also use the normalized adjacency eigenvalues in their decreasing absolute values to give a matrix proof of the Szemerédi Regularity Lemma.

5.4 Multiway discrepancy of directed graphs

A directed edge-weighted graph $G = (V, \boldsymbol{W})$ is given by its quadratic, but usually not symmetric edge-weight matrix $\boldsymbol{W} = (w_{ij})$ of zero diagonal, where w_{ij} is the nonnegative weight of the $i \to j$ edge ($i \neq j$). The row-sums $d_{out,i} = \sum_{j=1}^{n} w_{ij}$ and column-sums $d_{in,j} = \sum_{i=1}^{n} w_{ij}$ of $\boldsymbol{W}$ are the *out- and in-degrees*, while $\boldsymbol{D}_{out} = \mathrm{diag}(d_{out,1}, \dots, d_{out,n})$ and $\boldsymbol{D}_{in} = \mathrm{diag}(d_{in,1}, \dots, d_{in,n})$ are the diagonal *out- and in-degree matrices*, respectively. Now Definition 10 can be formulated as follows.

Definition 14. The multiway discrepancy of the directed, weighted graph $G = (V, \boldsymbol{W})$ in the in-clustering $(V_{in,1}, \dots, V_{in,k})$ and out-clustering $(V_{out,1}, \dots, V_{out,k})$ of its vertices is

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

where $w(X,Y)$ is the sum of the weights of the $X \to Y$ edges, whereas $\mathrm{Vol}_{out}(X) = \sum_{i \in X} d_{out,i}$ and $\mathrm{Vol}_{in}(Y) = \sum_{j \in Y} d_{in,j}$ are the out- and in-volumes, respectively. The minimum k-way discrepancy of the directed edge-weighted graph $G = (V, \boldsymbol{W})$ is

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

Butler [**?**] and Chung and Kenter [28] treat the $k = 1$ case, and for a general k, Theorem 7 implies the following.

Theorem 11 ([**?**])**.** *Let $G = (V, \boldsymbol{W})$ be directed edge-weighted graph, $\boldsymbol{W}$ is non-degenerate. Then for any integer $1 \le k \le \mathrm{rank}(\boldsymbol{W})$,*

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

holds, provided $0 < \mathrm{md}_k(G) < 1$, where s_k is the k-th largest nontrivial singular value of the normalized edge-weight matrix $\boldsymbol{W}_D = \boldsymbol{D}_{out}^{-1/2} \boldsymbol{W} \boldsymbol{D}_{in}^{-1/2}$ of G.

Together with BSM students, in 2012 we applied this spectral method to find migration patterns in the set of 75 countries, and found 3 underlying immigration and emigration trait clusters. The SVD-based algorithm is the same as the one introduced in the construction for rectangular matrices of nonnegative entries.

References

[1] Achlioptas D and McSherry F 2007 Fast computation of low-rank matrix approximations. *J. ACM* **54** (2), Article 9.

[2] Almaas E, Kovács B, Vicsek T, Oltvai ZN and Barabási A-L 2004 Global Organization of the Metabolic Fluxes in the Bacterium Escherichia Coli. *Nature* **427**, 839–843.

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

[4] Alon N, Duke RA, Lefmann H, Rödl V and Yuster R 1994 The algotithmic aspects of the Regularity Lemma. *J. Algorithms* **16**, 80–109.

[5] Alon N and Spencer JH 2000 *The Probabilistic Method*. Wiley.

[6] Alon N, Coja-Oghlan A, Han H, Kang M, Rödl V and Schacht M 2010 Quasi-randomness and algorithmic regularity for graphs with general degree distributions, *Siam J. Comput.* **39** (6), 2336–2362.

[7] Barabási A-L and Albert R 1999 Emergence of Scaling in Random Networks. *Science* **286**, 509–512.

[8] 2009Bickel Bickel PJ and Chen A 2009 A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** (50), 21068–21073.

[9] Bilu Y and Linial N 2006 Lifts, discrepancy and nearly optimal spectral gap. *Combinatorica* **26** (5), 495–519.

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

[11] Bolla M 2005 Recognizing Linear Structure in Noisy Matrices. *Linear Algebra Appl.* **402**, 228–244.

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

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

[14] Bolla M 2011 Beyond the expanders. *Int. J. Comb.*, 787596.

[15] Bolla M 2011 Spectra and structure of weighted graphs. *Electronic Notes in Discrete Mathematics* **38**, 149–154.

[16] Bolla, M., Relating multiway discrepancy and singular values of nonnegative rectangular matrices, *Discrete Applied Mathematics* **203** (2016), 26-34.

[17] Bolla, M., Elbanna, A., Matrix and discrepancy view of generalized random and quasirandom graphs graphs, *Special Matrices*, Volume 4, Issue 1, Pages 31–45, ISSN (Online) 2300-7451, De Gruyter Open (2016).

[18] Bolla, M., Generalized quasirandom properties of expanding graph sequences, *Nonlinearity* **33** No.4 (2020), 1405-1424.

[19] Bollobás B 2001 *Random Graphs*, 2nd edn. Cambridge Univ. Press, Cambridge.

[20] Bollobás B, Riordan O, Spencer J and Tusnády G 2001 The degree sequence of a scale-free random graph process. *Random Struct. Algorithms* **18** (1), 279–290.

[21] Bollobás B and Riordan O 2003 Robustness and vulnerability of scale-free random graphs. *Internet Math.* **1** (1), 1–35.

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

[23] Bollobás B, Janson S and Riordan O 2003 The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms* **31**, 3–122.?

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

[25] Chaudhuri K, Chung F and Tsiatas A 2012 Spectral clustering of graphs with general degrees in the extended planted partition model. In *JMLR Workshop and Conference Proceedings. Proc. 25th Annual Conference on Learning Theory (COLT 2012)* (Mannor S, Srebro N and Williamson RC eds), Vol. 23, pp. 35.1–35.23. Edinbourgh, Scotland.

[26] Choi DS, Wolfe PJ and Airoldi EM 2012 Stochastic blockmodels with growing number of classes. *Biometrika* **99** (2), 273–284.

[27] Chung F, Graham RL and Wilson RK 1989 Quasi-random graphs, *Combinatorica* **9**, 345–362.

[28] Chung F, Lu L and Vu V 2003 Eigenvalues of random power law graphs. *Ann. Comb.* **7**, 21–33.

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

[30] Coja-Oghlan A and Lanka A 2009 The spectral gap of random graphs with given expected degrees. *Electron. J. Comb.* **16** (1), R138.

[31] Coja-Oghlan A and Lanka A 2009 Finding planted partitions in random graphs with general degree distributions. *J. Discret. Math.* **23** (4), 1682–1714.

[32] Coja-Oghlan A 2010 Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** (2), 227–284.

[33] Dhillon IS 2001 Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. ACM Int. Conf. Knowledge Disc. Data Mining (KDD 2001)* (Provost FJ and Srikant R eds), pp. 269-274. Association for Computer Machinery, New York.

[34] Drineas P, Frieze A, Kannan R, Vempala S and Vinay V 2004 Clustering large graphs via the singular value decomposition. *Mach. Learn.* **56**, 9–33.

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

[36] Farkas IJ, Derényi I, Barabási, A-L. and Vicsek T 2001 Spectra of 'Real-World' Graphs: Beyond the Semi-Circle Law. *Phys. Rev. E* **64**, 026704.

[37] Frieze A and Kannan R 1999 Quick approximation to matrices and applications. *Combinatorica* **19** (2), 175–220.

[38] Füredi Z and Komlós J 1981 The eigenvalues of random symmetric matrices. *Combinatorica* **1**, 233–241.

[39] Golub GH and Van Loan CF 1996 *Matrix Computations*, 3rd edn. Johns Hopkins University Press, Baltimore, MD.

[40] Granovetter MS 1973 The Strength of Weak Ties. *Amer. J. Sociology* **78**, 1360–1380.

[41] Gundert A and Wagner U 2012 On Laplacians of random complexes. In *Proc. 28th Annual ACM Symposium on Computational Geometry, SoCG* (Dey TK and Whitesides S eds), pp. 151-160.

[42] Hoeffding W 1963 Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **58** (301), 13–30.

[43] Holland P, Laskey KB and Leinhardt S 1983 Stochastic blockmodels: some first steps. *Social Networks* **5**, 109–137.

[44] Hoory S, Linial N and Widgerson A 2006 Expander graphs and their applications. *Bull. Amer. Math. Soc. (N. S.)* **43** (4), 439–561.

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

[46] Juhász F 1996 On the structural eigenvalues of block random matrices. *Linear Algebra Appl.* **246**, 225–231.

[47] Kannan R, Salmasian H and Vempala S 2005 The spectral method for general mixture models. In *Proc. 18th Annual Conference on Learning Theory* (Auer P and Meir R eds), pp. 444–457. Bertinoro, Italy.

[48] Karrer B and Newman MEJ 2011 Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.

[49] Karrer B Levina E and Newman MEJ 2008 Robustness of community structure in networks. *Phys. Rev. E* **77**, 046119.

[50] Kluger Y, Basri R, Chang JT and Gerstein M 2003 Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Rerearch* **13**, 703–716.

[51] Komlós J, Shokoufanden A, Simonovits M and Szemerédi E 2002 Szemerédi's Regularity Lemma and Its Applications in Graph Theory. In *Lecture Notes in Computer Science*, Vol 2292, pp. 84–112. Springer, Berlin.

[52] Leskovec J, Lang KI, Dasgupta A and Mahoney MW 2008 Statistical properties of community structure in large social and information networks. In *Proc. 17th International Conference on World Wide Web, Bejing, China*, pp. 695–704. ACM New York NY.

[53] Leskovec J, Lang KI, Dasgupta A and Mahoney MW 2009 Community structure in large networks: natural cluster sizes and the absence of well-defined clusters. *Internet Math.* **6** (1), 29–123.

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

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

[56] Martella F and Vichi M Clustering microarray data using model-based double K-means. *J. Appl. Stat.*, **39** (9), 1853–1869.

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

[58] Móri TF 2005 The Maximum Degree of the Barabási–Albert Random Tree. *Combin. Probab. Comput.* **14** (3), 339-348.

[59] Reichardt J and Bornholdt S 2007 Partitioning and modularity of graphs with arbitrary degree distribution. *Physical Review E* **76**, 015102(R).

[60] Rohe K, Chatterjee S and Yu B 2011 Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Stat.* **39** (4), 1878–1915.

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

[62] Söderberg B 2003 Properties of random graphs with hidden color. *Phys. Rev. E* **68**, 026107.

[63] Stewart GW and Sun J-g. 1990 *Matrix Pertubation Theory.* Academic Press.

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

[65] Tao T 2006 Szemerédi's regularity lemma revisited. *Contrib. Discret. Math.* **1**, 8–28.

[66] Von Luxburg U Belkin M and Bousquet O 2008 Consistency of spectral clustering. *Ann. Stat.* **36**, 555–586.

[67] Wasserman S and Faust K 1994 *Social Network Analysis: Methods and Applications.* Cambridge Univ. Press, Cambridge.