SPECTRAL CLUSTERING, Lesson 5.
Testable graph and contingency table parameters

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

November 5, 2020

In this lesson, the theory of convergent graph sequences, elaborated by graph theorists (see e.g. [5, 6, 17, 25]), will be used for vertex- and edge-weighted graphs and contingency tables.Roughly speaking, graphs and contingency tables of a convergent sequence become more and more similar to each other in small details, which fact is made exact in terms of the convergence of homomorphism densities when 'small' simple graphs or binary tables are mapped into the 'large' networks of the sequence. The convergence can as well be formulated in terms of the cut-distance, and limit objects are defined. This cut-distance also makes it possible to classify graphs and contingency tables, or to assign them to given prototypes.

Testable parameters are, in fact, nonparametric statistics defined on graphs and contingency tables that can be consistently estimated based on a smaller sample, selected randomly from the underlying huge network. Real-world graphs or rectangular arrays are sometimes considered as samples from a large network, and we want to conclude the network's parameters from the same parameters of its smaller parts. The theory guarantees that this can be done if the investigated parameter is testable. Certain maximum and balanced minimum multiway cut densities are indeed testable.

1 Convergent graph sequences

Let $G = G_n$ be a weighted graph on the vertex set $V(G) = \{1, \dots, n\} = [n]$ and edge set $E(G)$. Both the edges and vertices have weights: the edge-weights are pairwise similarities $\beta_{ij} = \beta_{ji} \in [0,1]$, $i, j \in [n]$ between the vertices, while the vertex-weights $\alpha_i > 0$ ($i \in [n]$) indicate relative significance of the vertices. For example, in a social network, the edge-weights are pairwise associations between people, while the vertex-weights can be individual abilities in the same context in which relations between them exist; e.g. in the actors' network, relative frequencies of costarring of actors are the pairwise similarities, whereas individual abilities of the actors are their weights. In strategic interaction networks, the edge-weights represent the mutual effect of the pairs of individuals on each other's strategies, while vertex-weights are actions of the individuals that they would follow themselves, without being aware of actions of their neighbors (neighbors are persons to whom an individual is connected with positive weights), see e.g. [2].

It is important that the edge-weights are nonnegative ($\beta_{ij} = 0$ means that vertices i and j are not connected at all), the normalization into the [0,1] interval is for the sake of treating them later as probabilities for random sampling. A simple graph has edge-weights 0 or 1, and vertex-weights all 1. Note that the edge-weights resemble the w_{ij}'s used in the previous lessons, but here loops may also be present, namely, if $\beta_{ii} > 0$ for some i. If all the edge-weights are positive, our graph is called *soft-core*. Soft-core graphs are dense, and among the simple graphs they are the complete graphs.

Let $\mathcal{G}$ denote the set of all such edge- and vertex-weighted graphs. The *volume* of $G \in \mathcal{G}$ is defined by $\alpha_G = \sum_{i=1}^n \alpha_i$, while that of the vertex-subset S by $\alpha_S = \sum_{i\in S} \alpha_i$. Note that this notion of volume coincides with that of the previous lessons only if the vertex-weights are the generalized degrees. Further,

$$e_G(S,T) = \sum_{s\in S} \sum_{t\in T} \alpha_s \alpha_t \beta_{st}$$

denotes the *vertex- and edge-weighted cut* between the (not necessarily disjoint) vertex-subsets S and T, which is equal to $w(S,T)$ of the previous lessons only if the vertex-weights are all 1. However, for brevity, we will also call $e_G(S,T)$ weighted cut throughout this chapter.

[6] define the *homomorphism density* between the simple graph F (on vertex set $V(F) = [k]$) and the above weighted graph G as

$$t(F,G) = \frac{1}{(\alpha_G)^k} \sum_{\Phi: V(F)\to V(G)} \prod_{i=1}^k \alpha_{\Phi(i)} \prod_{ij\in E(F)} \beta_{\Phi(i)\Phi(j)}.$$

Note that under homomorphism, an edge-preserving map is understood in the following sense. If F is a simple and G is an edge-weighted graph, then $\Phi : V(F) \to V(G)$ is a homomorphism when for every $ij \in E(F)$, $\beta_{\Phi(i)\Phi(j)} > 0$. Therefore, in the above summation, a zero term will correspond to a Φ which is not a homomorphism. For 'large' n, [6] relate this quantity to the probability that the following sampling results in F: k vertices of G are selected with replacement out of the n ones, with respective probabilities α_i/α_G $(i = 1, \dots, n)$; given the vertex-subset $\{\Phi(1), \dots, \Phi(k)\}$, the edges come into existence conditionally independently, with probabilities of the edge-weights. Such a *random simple graph is denoted by* $\xi(k,G)$. In the sequel only the $k \ll n$ case – when $t(F,G)$ is very close to to the probability that the above sampling from G results in F – makes sense, and this is the situation we need: k is kept fixed, while n tends to infinity. (For a more precise formulation with induced and injective homomorphisms, see [6]).

Definition 1 *The weighted graph sequence* (G_n) *is (left-)convergent if the sequence* $t(F,G_n)$ *converges for any simple graph* F *as* $n \to \infty$.

As other kinds of convergence are not discussed here, in the sequel, the word left will be omitted, and we simply use convergence. Note that [8] define other kinds of graph convergence too, together with equivalences and implications between them.

[25] also construct the limit object that is a symmetric, bounded, measurable function $W : [0,1] \times [0,1] \to \mathbb{R}$, called *graphon.* Let $\mathcal{W}$ denote the set of these functions. The interval [0,1] corresponds to the vertices and the values $W(x,y) = W(y,x)$ to the edge-weights. In view of the conditions imposed on the edge-weights, the range is also the [0,1] interval. The set of symmetric, measurable functions $W : [0,1] \times [0,1] \to [0,1]$ is denoted by $\mathcal{W}_{[0,1]}$. The stepfunction graphon $W_G \in \mathcal{W}_{[0,1]}$ is assigned to the weighted graph $G \in \mathcal{G}$ in the following way: the sides of the unit square are divided into intervals $I_1, \dots, I_n$ of lengths $\alpha_1/\alpha_G, \dots, \alpha_n/\alpha_G$, and over the rectangle $I_i \times I_j$ the stepfunction takes on the value β_{ij}.

The so-called *cut distance* between the graphons W and U is

$$\delta_\square(W,U) = \inf_\nu \|W - U^\nu\|_\square \tag{1}$$

where the *cut-norm* of the graphon $W \in \mathcal{W}$ is defined by

$$\|W\|_\square = \sup_{S,T \subset [0,1]} \left| \iint_{S \times T} W(x,y)\,dx\,dy \right|,$$

and the infimum in (1) is taken over all measure-preserving bijections $\nu : [0,1] \to [0,1]$, while U^ν denotes the transformed U after performing the same measure-preserving bijection ν on both sides of the unit square. (You can also think of ν such that for a uniformly distributed random variable ξ over $(0,1)$, $\nu(\xi)$ has the same distribution.) An equivalence relation is defined over the set of graphons: two graphons belong to the same class if they can be transformed into each other by a measure-preserving bijection, i.e. their $\delta_\square$ distance is zero. In the sequel, we consider graphons modulo measure preserving maps, and under graphon we understand the whole equivalence class. By Theorem 5.1 of [26], the classes of $\mathcal{W}_{[0,1]}$ form a compact metric space with the $\delta_\square$ metric. Based on this fact, the authors give an analytic proof for the weak version of the Szemerédi's Regularity Lemma.

We will intensively use the following reversible relation between convergent weighted graph sequences and graphons (see Corollary 3.9 of [6]).

Fact 1 *For any convergent sequence (G_n) of weighted graphs with uniformly bounded edge-weights there exists a graphon such that $\delta_\square(W_{G_n}, W) \to 0$. Conversely, any graphon W can be obtained as the limit of a sequence of weighted graphs with uniformly bounded edge-weights. The limit of a convergent graph sequence is essentially unique: if $G_n \to W$, then also $G_n \to W'$ for precisely those graphons W' for which $\delta_\square(W, W') = 0$.*

Authors of [6, 7] define the $\delta_\square$ distance of two weighted graphs and that of a graphon and a graph in the following way. For the weighted graphs G, G', and for the graphon W

$$\delta_\square(G,G') = \delta_\square(W_G, W_{G'}) \quad \text{and} \quad \delta_\square(W,G) = \delta_\square(W, W_G).$$

Theorem 2.6 of [6] states that a sequence of weighted graphs with uniformly bounded edge-weights is convergent if and only if it is a Cauchy sequence in the metric $\delta_\square$.

A simple graph on k vertices can be sampled based on W in the following way: k uniform random numbers, $X_1, \dots, X_k$ are generated on [0,1] independently. Then we connect the vertices corresponding to X_i and X_j with probability $W(X_i, X_j)$. For the so obtained simple graph $\xi(k, W)$ the following large deviation result is proved (see Theorem 4.7 of [6]).

Fact 2 *Let k be a positive integer and $W \in \mathcal{W}_{[0,1]}$ be a graphon. Then with probability at least $1 - e^{-k^2/(2\log_2 k)}$, we have*

$$\delta_\Box(W, \xi(k, W)) \leq \frac{10}{\sqrt{\log_2 k}}. \tag{2}$$

Fixing k, Inequality (2) holds uniformly for any graphon $W \in \mathcal{W}_{[0,1]}$, especially for W_G. Further, the sampling from W_G is identical to the previously defined sampling with replacement from G, i.e. $\xi(k, G)$ and $\xi(k, W_G)$ are identically distributed. In fact, this argument is relevant in the $k \leq |V(G)|$ case.

2 Testability of weighted graph parameters

A function $f : G \to \mathbb{R}$ is called a *graph parameter* if it is invariant under isomorphism. In fact, a graph parameter is a statistic evaluated on the graph, and hence, we are interested in weighted graph parameters that are not sensitive to minor changes in the weights of the graph. By the definition of [6], the simple graph parameter f is testable if for every $\varepsilon > 0$ there is a positive integer k such that for every simple graph G on at least k vertices,

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

where $\xi(k, G)$ is a random simple graph on k vertices selected randomly from G as described above. Then they prove equivalent statements of testability for simple graphs.

These results remain valid if we consider weighted graph sequences (G_n) with *no dominant vertex-weights*, i.e. $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0$ as $n \to \infty$. To use this condition imposed on the vertex-weights, [4] slightly modified the definition of a testable graph parameter for weighted graphs.

Definition 2 *A weighted graph parameter f is testable if for every $\varepsilon > 0$ there is a positive integer k such that if $G \in \mathcal{G}$ satisfies*

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \leq \frac{1}{k}, \tag{3}$$

then

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

where $\xi(k, G)$ is a random simple graph on k vertices selected randomly from G as described in Section 1.

Note that for simple G, Inequality (3) implies that $|V(G)| \geq k$, and we get back the definition applicable to simple graphs..

By the above definition, a testable graph parameter can be consistently estimated based on a fairly large sample. As the randomization depends only on the $\alpha_i(G)/\alpha_G$ ratios, it is not able to distinguish between weighted graphs whose vertex-weights differ only in a constant factor. Thus, a testable weighted graph parameter is invariant under scaling the vertex-weights.

As a straightforward generalization of Theorem 6.1 in [6], [4] introduced the following equivalent statements of the testability for weighted graphs. We state this theorem without proof as it uses the ideas of the proof in [6], where some details for such a generalization are also elaborated.

Theorem 1 *For the weighted graph parameter f the following are equivalent:*

(a) f is testable.

(b) For every $\varepsilon > 0$ there is a positive integer k such that for every weighted graph $G \in \mathcal{G}$ satisfying the node-condition $\max_i \alpha_i(G)/\alpha_G \le 1/k$,

$$|f(G) - \boldsymbol{E}(f(\xi(k,G)))| \le \varepsilon.$$

(c) For every convergent weighted graph sequence (G_n) with $\max_i \alpha_i(G_n)/\alpha_{G_n} \to 0$*, $f(G_n)$ is also convergent $(n \to \infty)$.*

(d) f can (essentially uniquely) be extended to graphons such that the graphon functional $\tilde{f}$ is continuous in the cut-norm and $\tilde{f}(W_{G_n}) - f(G_n) \to 0$, whenever $\max_i \alpha_i(G_n)/\alpha_{G_n} \to 0$ *$(n \to \infty)$.*

(e) For every $\varepsilon > 0$ there is an $\varepsilon_0 > 0$ real and an $n_0 > 0$ integer such that if G_1, G_2 are weighted graphs satisfying $\max_i \alpha_i(G_1)/\alpha_{G_1} \le 1/n_0$, $\max_i \alpha_i(G_2)/\alpha_{G_2} \le 1/n_0$*, and $\delta_\square(G_1, G_2) < \varepsilon_0$, then $|f(G_1) - f(G_2)| < \varepsilon$ also holds.*

This theorem indicates that a testable parameter depends continuously on the whole graph, and hence, it is not sensitive to minor changes in the edge-weights. Some of these equivalences will be used in the proofs of the next section.

3 Convergence of the spectra and spectral subspaces

In [4], we proved the testability of some normalized balanced multiway cut densities such that we imposed balancing conditions on the cluster volumes. Under similar conditions, for fixed number of clusters k, the unnormalized and normalized multiway modularities are also testable, provided our edge-weighted graph has no dominant vertices. The proofs rely on statistical physics notions of [6], utilizing the fact that the graph convergence implies the convergence of the ground state energy. [31] showed that the Newman-Girvan modularity is an energy function (Hamiltonian), and hence, testability of the maximum normalized modularities, under appropriate balancing conditions, can be shown analogously. Here we rather discuss the testability of spectra and k-variances, because in spectral clustering methods these provide us with polynomial time algorithms, though only approximate solutions are expected via the spectral relaxation.

In Theorem 6.6 of [8], the authors prove that the normalized spectrum of a convergent graph sequence converges in the following sense. Let W be a graphon and (G_n) be a sequence of weighted graphs – with uniformly bounded edge-weights – tending to W. (For simplicity, we assume that $|V(G_n)| = n$). Let $|\lambda_{n,1}| \geq |\lambda_{n,2}| \geq \cdots \geq |\lambda_{n,n}|$ be the adjacency eigenvalues of G_n indexed by their decreasing absolute values, and let $\mu_{n,i} = \lambda_{n,i}/n$ $(i = 1, \dots, n)$ be the normalized eigenvalues. Further, let T_W be the $L^2[0,1] \to L^2[0,1]$ integral operator corresponding to W:

$$(T_W f)(x) = \int_0^1 W(x,y) f(y)\, dy.$$

It is well-known that his operator is self-adjoint and compact, and hence, it has a discrete real spectrum, whose only possible point of accumulation is the 0 (see the background material). Let $\mu_i(W)$ denote the ith largest absolute value eigenvalue of T_W. Then for every $i \geq 1$, $\mu_{n,i} \to \mu_i(W)$ as $n \to \infty$. In fact, the authors prove a bit more (see Theorem 6.7 of [8]): if a sequence W_n of uniformly bounded graphons converges to a graphon W, then for every $i \geq 1$, $\mu_i(W_n) \to \mu_i(W)$ as $n \to \infty$. Note that the spectrum of W_G is the normalized spectrum of G, together with countably infinitely many 0's. Therefore, the convergence of the spectrum of (G_n) is the consequence of that of (W_{G_n}).

We will prove that in the absence of dominant vertices, the normalized modularity spectrum is testable. To this end, both the modularity matrix and the graphon are related to kernels of special integral operators, discussed in Lesson 2. We recall the most important facts herein. Let (ξ, ξ') be a pair of identically distributed real-valued random variables defined over the product space $\mathcal{X} \times \mathcal{X}$ having a symmetric joint distribution $\mathbb{W}$ with equal margins $\mathbb{P}$. Suppose that the dependence between ξ and ξ' is regular, i.e., their joint distribution $\mathbb{W}$ is absolutely continuous with respect to the product measure $\mathbb{P} \times \mathbb{P}$, and let w denote its Radon–Nikodym derivative. Let $H = L^2(\xi)$ and $H' = L^2(\xi')$ be the Hilbert spaces of random variables which are functions of ξ and ξ' and have zero expectation and finite variance with respect to $\mathbb{P}$. Observe that H and H' are isomorphic Hilbert spaces with the covariance as inner product; further, they are embedded as subspaces into the L^2-space defined similarly over the product space. (Here H and H' are also isomorphic in the sense that for any $\psi \in H$ there exists a $\psi' \in H'$ and vice versa, such that ψ and ψ' are identically distributed.)

Consider the linear operator taking conditional expectation between H' and H with respect to the joint distribution. It is an integral operator and will be denoted by $P_{\mathbb{W}} : H' \to H$ as it is a projection restricted to H' and projects onto H. To $\psi' \in H'$ the operator $P_{\mathbb{W}}$ assigns $\psi \in H$ such that $\psi = \mathbb{E}_{\mathbb{W}}(\psi' \,|\, \xi)$, i.e.

$$\psi(x) = \int_{\mathcal{Y}} w(x,y) \psi'(y)\, \mathbb{P}(dy), \quad x \in \mathcal{X}.$$

If

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

then $P_{\mathbb{W}}$ is a Hilbert–Schmidt operator, therefore it is compact and has SD

$$P_{\mathbb{W}} = \sum_{i=1}^{\infty} \lambda_i \langle ., \psi_i' \rangle_{H'} \psi_i$$

where for the eigenvalues $|\lambda_i| \le 1$ holds and the eigenvalue–eigenfunction equation looks like

$$P_{\mathbb{W}} \psi'_i = \lambda_i \psi_i \quad (i = 1, 2, \dots)$$

where ψ_i and ψ'_i are identically distributed, whereas their joint distribution is $\mathbb{W}$. It is easy to see that $P_{\mathbb{W}}$ is self-adjoint and it takes the constantly 1 random variable of H' into the constantly 1 random variable of H; however, the $\psi_0 = 1, \psi'_0 = 1$ pair is not regarded as a function pair with eigenvalue $\lambda_0 = 1$, since they have no zero expectation. More precisely, the kernel is reduced to $w(x,y) - 1$.

Theorem 2 *Let $G_n = (V_n, \boldsymbol{W}_n)$ be the general entry of a convergent sequence of connected edge-weighted graphs whose edge-weights are in [0,1] and the vertex-weights are the generalized degrees. Assume that there are no dominant vertices. Let W denote the limit graphon of the sequence (G_n), and let*

$$1 \ge |\mu_{n,1}| \ge |\mu_{n,2}| \ge \dots \ge |\mu_{n,n}| = 0$$

be the normalized modularity spectrum of G_n (the eigenvalues are indexed by their decreasing absolute values). Further, let $\mu_i(P_{\mathbb{W}})$ is the ith largest absolute value eigenvalue of the integral operator $P_{\mathbb{W}} : L^2(\xi') \to L^2(\xi)$ taking conditional expectation with respect to the joint measure $\mathbb{W}$ embodied by the normalized limit graphon W, and ξ, ξ' are identically distributed random variables with the marginal distribution of their symmetric joint distribution $\mathbb{W}$. Then for every $i \ge 1$,

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

Proof 1 *In case of a finite $\mathcal{X}$ (vertex set) we have a weighted graph, and we will show that the operator taking conditional expectation with respect to the joint distribution, determined by the edge-weights, corresponds to its normalized modularity matrix.*

Indeed, let $\mathcal{X} = V$, $|V| = n$, and $G_n = (V, \boldsymbol{W})$ be an edge-weighted graph on the $n \times n$ weight matrix of the edges $\boldsymbol{W}$ with entries W_{ij}'s; now, they do not necessarily sum up to 1. (For the time being, n is kept fixed, so – for the sake of simplicity – we do not denote the dependence of $\boldsymbol{W}$ on n). Let the vertices be also weighted with special weights $\alpha_i(G_n) := \sum_{j=1}^n W_{ij}$, $i = 1, \dots, n$. Then the step-function graphon W_{G_n} is such that $W_{G_n}(x,y) = W_{ij}$ whenever $x \in I_i$ and $y \in I_j$, where the (not necessarily contiguous) intervals $I_1, \dots, I_n$ form a partition of [0,1] such that the length of I_i is $\alpha_i(G_n)/\alpha_{G_n}$ $(i = 1, \dots, n)$.

Let us transform $\boldsymbol{W}$ into a symmetric joint distribution $\mathbb{W}_n$ over $V \times V$. The entries $w_{ij} = W_{ij}/\alpha_{G_n}$ $(i,j = 1, \dots, n)$ embody this discrete joint distribution of random variables ξ and ξ' which are identically distributed with marginal distribution $d_1, \dots, d_n$, where $d_i = \alpha_i(G_n)/\alpha_{G_n}$ $(i = 1, \dots, n)$. With the previous notation $H = L^2(\xi)$, $H' = L^2(\xi')$, and the operator $P_{\mathbb{W}_n} : H' \to H$ taking conditional expectation is an integral operator with now discrete kernel $K_{ij} = \frac{w_{ij}}{d_i d_j}$. The fact that ψ, ψ' is an eigenfunction pair of $P_{\mathbb{W}_n}$ with eigenvalue λ means that

$$\frac{1}{d_i} \sum_{j=1}^n w_{ij} \psi'(j) = \sum_{j=1}^n \frac{w_{ij}}{d_i d_j} \psi'(j) d_j = \lambda \psi(i), \tag{4}$$

where $\psi(j) = \psi'(j)$ denotes the value of ψ or ψ' taken on with probability d_i (recall that ψ and ψ' are identically distributed). The above equation is equivalent to

$$\sum_{j=1}^{n} \frac{w_{ij}}{\sqrt{d_i}\sqrt{d_j}} \sqrt{d_j}\psi(j) = \lambda\sqrt{d_i}\psi(i).$$

Therefore, the vector of coordinates $\sqrt{d_i}\psi(i)$ $(i = 1, \dots, n)$ is a unit-norm eigenvector of the normalized modularity matrix with eigenvalue λ (note that the normalized modularity spectrum does not depend on the scale of the edge-weights, it is the same whether we use W_{ij}'s or w_{ij}'s as edge-weights). Consequently, the eigenvalues of the conditional expectation operator are the same as the eigenvalues of the normalized modularity matrix, and the possible values taken on by the eigenfunctions of the conditional expectation operator are the same as the coordinates of the transformed eigenvectors of the normalized modularity matrix forming the column vectors of the matrix $\mathbf{X}^*$ of the optimal $(k-1)$-dimensional representatives.

Let f be a stepwise constant function on [0,1], taking on value $\psi(i)$ on I_i. Then $\mathrm{Var}\psi = 1$ is equivalent to $\int_0^1 f^2(x)\,dx = 1$. Let K_{G_n} be the stepwise constant graphon defined as $K_{G_n}(x,y) = K_{ij}$ for $x \in I_i$ and $y \in I_j$. With this, the eigenvalue–eigenvector equation (4) looks like

$$\lambda f(x) = \int_0^1 K_{G_n}(x,y) f(y)\,dy.$$

The spectrum of K_{G_n} is the normalized modularity spectrum of G_n together with countably infinitely many 0's (it is of finite rank, and therefore, trivially compact), and because of the convergence of the weighted graph sequence G_n, in lack of dominant vertices, the sequence of graphons K_{G_n} also converges. Indeed, the $W_{G_n} \to W$ convergence in the cut metric means the convergence of the induced discrete distributions $\mathbb{W}_n$'s to the continuous $\mathbb{W}$. Since K_{G_n} and K are so-called copula transformations (see [?]) of those distributions; in lack of dominant vertices (this causes the convergence of the margins) they also converge, which in turn implies the $K_{G_n} \to K$ convergence in the cut metric.

Let K denote the limit graphon of K_{G_n} $(n \to \infty)$. This will be the kernel of the integral operator taking conditional expectation with respect to the joint distribution $\mathbb{W}$. It is easy to see that this operator is also a Hilbert–Schmidt operator, and therefore, compact. With these considerations the remainder of the proof is analogous to the proof of Theorem 6.7 of [8], where the authors prove that if the sequence (W_{G_n}) of graphons converges to the limit graphon W, then both ends of the spectra of the integral operators, induced by W_{G_n}'s as kernels, converge to the ends of the spectrum of the integral operator induced by W as kernel. We apply this argument for the spectra of the kernels induced by K_{G_n}'s and K.

Note that in [28], kernel operators are also discussed, but not with our normalization.

Remark 1 *By using Theorem 1 (c), provided there are no dominant vertices, Theorem 2 implies that for any fixed positive integer k, the $(k-1)$-tuple of the largest absolute value eigenvalues of the normalized modularity matrix is testable.*

Theorem 3 *Suppose, there are constants $0 < \varepsilon < \delta \le 1$ such that the normalized modularity spectrum (with decreasing absolute values) of any G_n satisfies*

$$1 \ge |\mu_{n,1}| \ge \cdots \ge |\mu_{n,k-1}| \ge \delta > \varepsilon \ge |\mu_{n,k}| \ge \cdots \ge |\mu_{n,n}| = 0.$$

With the notions of Theorem 2, and assuming that there are no dominant vertices of G_n's, the subspace spanned by the transformed eigenvectors $\boldsymbol{D}^{-1/2}\mathbf{u}_1$, ...,$\boldsymbol{D}^{-1/2}\mathbf{u}_{k-1}$ belonging to the $k-1$ largest absolute value eigenvalues of the normalized modularity matrix of G_n also converges to the corresponding $(k-1)$-dimensional subspace of $P_{\mathbb{W}}$. More precisely, if $\boldsymbol{P}_{n,k-1}$ denotes the projection onto the subspace spanned by the transformed eigenvectors belonging to $k-1$ largest absolute value eigenvalues of the normalized modularity matrix of G_n, and $\boldsymbol{P}_{k-1}$ denotes the projection onto the corresponding eigen-subspace of $P_{\mathbb{W}}$, then $\|\boldsymbol{P}_{n,k-1} - \boldsymbol{P}_{k-1}\| \to 0$ as $n \to \infty$ (in spectral norm).

Proof 2 *If we apply the convergence fact $\mu_{n,i} \to \mu_i(P_{\mathbb{W}})$ for indices $i = k-1$ and k, we get that there will be a gap of order $\delta - \varepsilon - o(1)$ between $|\mu_{k-1}(P_{\mathbb{W}})|$ and $|\mu_k(P_{\mathbb{W}})|$ too.*

Let $P_{\mathbb{W},n}$ denote the n-rank approximation of $P_{\mathbb{W}}$ (keeping its n largest absolute value eigenvalues, together with the corresponding eigenfunctions) in spectral norm. The projection $\boldsymbol{P}_{k-1}$ ($k < n$) operates on the eigen-subspace spanned by the eigenfunctions belonging to the $k-1$ largest absolute value eigenvalues of $P_{\mathbb{W},n}$ in the same way as on the corresponding $(k-1)$-dimensional subspace determined by $P_{\mathbb{W}}$. With these considerations, we apply the perturbation theory of eigen-subspaces with the following unitary invariant norm: the Schatten norm (or trace norm) of the Hilbert–Schmidt operator A is $\|A\|_4 = (\sum_{i=1}^{\infty} \lambda_i^4(A))^{1/4}$ (see e.g. [6] and the Linear Algebra basics) for matrices). Our argument with the finite $(k-1)$ rank projections is the following. Denoting by $P_{\mathbb{W}_n}$ the integral operator belonging to the normalized modularity matrix of G_n (with kernel K_{G_n} introduced in the proof of Theorem 2),

$$\begin{aligned}\|\boldsymbol{P}_{n,k-1} - \boldsymbol{P}_{k-1}\| = \|\boldsymbol{P}_{n,k-1}^{\perp}\boldsymbol{P}_{k-1}\| &\le \|\boldsymbol{P}_{n,k-1}^{\perp}\boldsymbol{P}_{k-1}\|_4 \\ \le& \frac{c}{\delta - \varepsilon - o(1)} \|P_{\mathbb{W}_n} - P_{\mathbb{W},n}\|_4\end{aligned}$$

with constant c that is at most $\pi/2$ (see the Linear Algebra Basics). But

$$\|P_{\mathbb{W}_n} - P_{\mathbb{W},n}\|_4 \le \|P_{\mathbb{W}_n} - P_{\mathbb{W}}\|_4 + \|P_{\mathbb{W}} - P_{\mathbb{W},n}\|_4,$$

where the last term tends to 0 as $n \to \infty$, since the tail of the spectrum (taking the fourth power of the eigenvalues) of a Hilbert–Schmidt operator converges. For the convergence of the first term we use Lemma 7.1 of [6], which states that the trace-norm of an integral operator can be estimated from above by four times the cut-norm of the corresponding kernel. But the convergence in the cut distance of the corresponding kernels to zero follows from the considerations made in the proof of Theorem 2. This finishes the proof.

Remark 2 *As the k-variance depends continuously on the above subspaces, Theorem 3 implies the testability of the k-variance as well.*

The above results suggest that in the absence of dominant vertices, even the normalized modularity matrix of a smaller part of the underlying weighted graph, selected at random with an appropriate procedure, is able to reveal its cluster structure. Hence, the gain regarding the computational time of this spectral clustering algorithm is twofold: we only use a smaller part of the graph and the spectral decomposition of its normalized modularity matrix runs in polynomial time in the reduced number of the vertices. Under the vertex- and cluster-balance conditions this method can give quite good approximations for the multiway cuts and helps us to find the number of clusters and identify the cluster structure.

Even if the spectrum of convergent graph sequences converges, the spectrum itself does not carry enough information for the cluster structure of the graph as stated in [27]. However, together with the eigenvectors it must carry the sufficient information; moreover, it suffices to consider the structural eigenvalues with their corresponding spectral subspaces.

4 Noisy graph sequences

Now, we use the above theory for perturbations, showing that special noisy weighted graph sequences converge in the sense of Section 1. If not stated otherwise, the vertex-weights are equal (say, equal to 1), and a weighted graph G on n vertices is determined by its $n \times n$ symmetric weight matrix $\boldsymbol{A}$. Let $G_{\boldsymbol{A}}$ denote the weighted graph with unit vertex-weights and edge-weights that are entries of $\boldsymbol{A}$. We will use the notion of a Wigner-noise and blown-up matrix of Lesson 4.

Let us fix the $q \times q$ symmetric pattern matrix $\boldsymbol{P}$ of entries $0 < p_{ij} < 1$, and blow it up to an $n \times n$ blown-up matrix $\boldsymbol{B}_n$ of blow-up sizes $n_1, \dots, n_q$ (note that $\boldsymbol{B}_n$ is a soft-core graph). Consider the noisy matrix $\boldsymbol{A}_n = \boldsymbol{B}_n + \boldsymbol{W}_n$ as $n_1, \dots, n_q \to \infty$ at the same rate, where $\boldsymbol{W}_n$ is an $n \times n$ Wigner-noise. While perturbing $\boldsymbol{B}_n$ by $\boldsymbol{W}_n$, assume that for the uniform bound of the entries of $\boldsymbol{W}_n$ the condition

$$K \leq \min\{ \min_{i,j \in [q]} p_{ij} \, , \, 1 - \max_{i,j \in [q]} p_{ij} \} \tag{5}$$

is satisfied. In this way, the entries of $\boldsymbol{A}_n$ are in the [0,1] interval, and hence, $G_{\boldsymbol{A}_n} \in \mathcal{G}$. We remark that $G_{\boldsymbol{W}_n} \notin \mathcal{G}$, but $W_{G_{\boldsymbol{W}_n}} \in \mathcal{W}$ and the theory of bounded graphons applies to it. In Lesson 4, we showed that by adding an appropriate Wigner-noise to $\boldsymbol{B}_n$, we can achieve that $\boldsymbol{A}_n$ becomes a 0-1 matrix: its entries are equal to 1 with probability p_{ij} and 0 otherwise within the block of size $n_i \times n_j$ (after rearranging its rows and columns). In this case, the corresponding noisy graph $G_{\boldsymbol{A}_n}$ is a generalized random graph.

By routine large deviation techniques we are able to prove that the cut-norm of the stepfunction graphon assigned to a Wigner-noise tends to zero with probability 1 as $n \to \infty$.

Proposition 1 *For any sequence $\boldsymbol{W}_n$ of Wigner-noises*

$$\lim_{n \to \infty} \| W_{G_{\boldsymbol{W}_n}} \|_{\square} = 0$$

almost surely.

The main idea of the proof is that the definition of the cut-norm of a stepfunction graphon and formulas (7.2), (7.3) of [6] yield

$$\|W_{G_{\boldsymbol{W}_n}}\|_\square = \frac{1}{n^2} \max_{U,T\subset[n]} \left| \sum_{i\in U} \sum_{j\in T} w_{ij} \right| \le 6 \max_{U\subset[n]} \frac{1}{n^2} \left| \sum_{i\in U} \sum_{j\in[n]\setminus U} w_{ij} \right| ,$$

where the entries behind the latter double summation are independent random variables. Hence, the Azuma's inequality (see [10]) is applicable, and the statement follows by the Borel–Cantelli lemma.

Let $\boldsymbol{A}_n := \boldsymbol{B}_n + \boldsymbol{W}_n$ and $n_1, \dots, n_q \to \infty$ in such a way that $\lim_{n\to\infty} \frac{n_i}{n} = r_i$ $(i = 1, \dots, q)$, $n = \sum_{i=1}^q n_i$; further, for the uniform bound K of the entries of the matrix $\boldsymbol{W}_n$ the condition (5) is assumed. Under these conditions, Proposition 1 implies that the noisy graph sequence $(G_{\boldsymbol{A}_n}) \subset \mathcal{G}$ converges almost surely in the $\delta_\square$ metric. It is easy to see that the almost sure limit is the stepfunction W_H, where the vertex- and edge-weights of the weighted graph H are

$$\alpha_i(H) = r_i \quad (i \in [q]), \qquad \beta_{ij}(H) = p_{ij} \quad (i, j \in [q]).$$

By adding a special Wigner-noise, the noisy graph sequence $(G_{\boldsymbol{A}_n})$ becomes a generalized graph sequence on the model graph H, under which the following is understood. Let $V = [n]$ be the desired vertex-set. A vertex is put into the vertex-subset V_i with probability $\alpha_i(H)$; then vertices of V_i and V_j are connected with probability $\beta_{ij}(H)$, $i, j = 1, \dots, q$ (see also the generalized random graphs of Lesson 6).

The deterministic counterparts of the generalized random graphs are the *generalized quasirandom graph*s introduced by [27] in the following way. We have a model graph graph H on q vertices with vertex-weights $r_1, \dots, r_q$ and edge-weights $p_{ij} = p_{ji}$, $i, j = 1, \dots, q$. Then (G_n) is H-quasirandom if $G_n \to H$ as $n \to \infty$ in the sense of Definition 1. Authors of [27] also prove that the vertex set V of a generalized quasirandom graph G_n can be partitioned into clusters $V_1, \dots, V_q$ in such a way that $\frac{|V_i|}{|V|} \to r_i$ $(i = 1, \dots, q)$ and the subgraph of G_n induced by V_i is the general term of a quasirandom graph sequence with edge-density tending to p_{ii} $(i = 1, \dots, q)$, whereas the bipartite subgraph between V_i and V_j is the general term of a quasirandom bipartite graph sequence with edge-density tending to p_{ij} $(i \ne j)$ as $n \to \infty$. Consequently, for any fixed finite graph F, the number of copies of F is asymptotically the same in the above generalized random and generalized quasirandom graphs on the same model graph and number of vertices.

5 Testability of minimum balanced multiway cuts

The testability of the maximum cut density is stated in [5] based on earlier algorithmic results of [1, 19, 20]. We are rather interested in the minimum cut density, which is somewhat different. We will show that it trivially tends to zero as the number of the graph's vertices tends to infinity, whereas the normalized version of it (cuts are penalized by the volumes of the clusters they connect) is not testable. For example, if a single vertex is loosely connected to a dense part, the minimum cut density of the whole graph is 'small', however, randomizing a smaller sample, with high probability, it comes from the dense part with a 'large' minimum normalized cut density. Nonetheless, if we impose conditions on the cluster volumes in anticipation, the so obtained balanced minimum cut densities are testable. Balanced multiway cuts are frequently looked for in contemporary cluster analysis when we want to find groups of a large network's vertices with sparse inter-cluster connections, where the clusters do not differ significantly in sizes.

For the proofs of the testability results we use Theorem 1 and some notion of statistical physics in the same way as in [8].

Let $G \in \mathcal{G}$ be a weighted graph on n vertices with vertex-weights $\alpha_1, \dots, \alpha_n$ and edge-weights β_{ij}'s. Let $q \le n$ be a fixed positive integer, and $\mathcal{P}_q$ denote the set of q-partitions $P = (V_1, \dots, V_q)$ of the vertex set V. The non-empty, disjoint vertex-subsets sometimes are referred to as clusters or states. The *factor graph* or *q-quotient* of G with respect to the q-partition P is denoted by G/P and it is defined as the weighted graph on q vertices with vertex- and edge-weights

$$\alpha_i(G/P) = \frac{\alpha_{V_i}}{\alpha_G} \quad (i \in [q]) \quad \text{and} \quad \beta_{ij}(G/P) = \frac{e_G(V_i, V_j)}{\alpha_{V_i}\alpha_{V_j}} \quad (i, j \in [q]),$$

respectively.

In terms of the factor graph, the following weak version of the Szemerédi's Regularity Lemma of [6].

Lemma 1 (Weak Regularity Lemma) *For every $\varepsilon > 0$, every weighted graph G has a partition P into at most $4^{\frac{1}{\varepsilon^2}}$ clusters such that*

$$\delta_\square(G, G/P) \le \varepsilon \|G\|_2,$$

where

$$\|G\|_2 = \left(\sum_{i,j} \frac{\alpha_i \alpha_j}{\alpha_G^2} \beta_{ij}^2 \right)^{1/2} .$$

Moreover, [24] gives an algorithm to compute a weak Szemerédi partition in a huge graph. The way of presenting the output of the algorithm for a large graph was formerly proposed by [19].

Let $\hat{\mathcal{S}}_q(G)$ denote the set of all q-quotients of G. The *Hausdorff distance* between $\hat{\mathcal{S}}_q(G)$ and $\hat{\mathcal{S}}_q(G')$ is defined by

$$\begin{aligned} &d^{\mathrm{Hf}}(\hat{\mathcal{S}}_q(G), \hat{\mathcal{S}}_q(G')) \\ &= \max\{ \sup_{H \in \hat{\mathcal{S}}_q(G)} \inf_{H' \in \hat{\mathcal{S}}_q(G')} d_1(H, H') \, , \sup_{H' \in \hat{\mathcal{S}}_q(G')} \inf_{H \in \hat{\mathcal{S}}_q(G)} d_1(H, H') \} \end{aligned}$$

where

$$\begin{aligned} d_1(H, H') = &\sum_{i,j \in [q]} \left| \frac{\alpha_i(H)\alpha_j(H)\beta_{ij}(H)}{\alpha_H^2} - \frac{\alpha_i(H')\alpha_j(H')\beta_{ij}(H')}{\alpha_{H'}^2} \right| \\ &+ \sum_{i \in [q]} \left| \frac{\alpha_i(H)}{\alpha_H} - \frac{\alpha_i(H')}{\alpha_{H'}} \right| \end{aligned}$$

is the l^1-distance between two weighted graphs H and H' on the same number of vertices. (If especially, H and H' are factor graphs, then $\alpha_H = \alpha_{H'} = 1$.)

Given the real symmetric $q \times q$ matrix $\boldsymbol{J}$ and the vector $\mathbf{h} \in \mathbb{R}^q$, the partitions $P \in \mathcal{P}_q$ also define a spin system on the weighted graph G. The so-called *ground state energy* (Hamiltonian) of such a spin configuration is

$$\hat{\mathcal{E}}_q(G, \boldsymbol{J}, \mathbf{h}) = - \max_{P \in \mathcal{P}_q} \left(\sum_{i \in [q]} \alpha_i(G/P) h_i + \sum_{i,j \in [q]} \alpha_i(G/P) \alpha_j(G/P) \beta_{ij}(G/P) J_{ij} \right)$$

where $\boldsymbol{J}$ is the so-called *coupling-constant matrix* with J_{ij} representing the strength of interaction between states i and j, and $\mathbf{h}$ is the magnetic field. They carry physical meaning. We will use only special $\boldsymbol{J}$ and $\mathbf{h}$.

Sometimes, we need balanced q-partitions to regulate the proportion of the cluster volumes. A slight balancing between the cluster volumes is achieved by fixing a positive real number c $(c \leq 1/q)$. Let $\mathcal{P}_q^c$ denote the set of q-partitions of V such that $\frac{\alpha_{V_i}}{\alpha_G} \geq c \quad (i \in [q])$, or equivalently, $c \leq \frac{\alpha_{V_i}}{\alpha_{V_j}} \leq \frac{1}{c}$ $(i \neq j)$. A more accurate balancing is defined by fixing a probability vector $\mathbf{a} = (a_1, \dots, a_q)$ with components forming a probability distribution over $[q]$: $a_i > 0$ $(i \in [q])$, $\sum_{i=1}^q a_i = 1$. Let $\mathcal{P}_q^{\mathbf{a}}$ denote the set of q-partitions of V such that $\left(\frac{\alpha_{V_1}}{\alpha_G}, \dots, \frac{\alpha_{V_q}}{\alpha_G} \right)$ is approximately $\mathbf{a}$-distributed, that is $\left| \frac{\alpha_{V_i}}{\alpha_G} - a_i \right| \leq \frac{\alpha_{\max}(G)}{\alpha_G} \quad (i = 1, \dots, q)$. Observe that the above difference tends to 0 as $|V(G)| \to \infty$ for weighted graphs with no dominant vertex-weights.

The *microcanonical ground state energy* of G given $\mathbf{a}$ and $\boldsymbol{J}$ $(\mathbf{h} = \mathbf{0})$ is

$$\hat{\mathcal{E}}_q^{\mathbf{a}}(G, \boldsymbol{J}) = - \max_{P \in \mathcal{P}_q^{\mathbf{a}}} \sum_{i,j \in [q]} \alpha_i(G/P) \alpha_j(G/P) \beta_{ij}(G/P) J_{ij}.$$

Theorem 2.14 and 2.15 of [8] state the following important facts.

Fact 3 *The convergence of the weighted graph sequence (G_n) with no dominant vertex-weights is equivalent to the convergence of its microcanonical ground state energies for any q, $\mathbf{a}$, and $\boldsymbol{J}$. Also, it is equivalent to the convergence of its q-quotients in Hausdorff distance for any q.*

Fact 4 *Under the same conditions, the convergence of the above (G_n) implies the convergence of its ground state energies for any q, $\boldsymbol{J}$, and $\mathbf{h}$.*

Using these facts, we investigate the testability of some special multiway cut densities defined in the forthcoming definitions.

Definition 3 *The minimum q-way cut density of G is*

$$f_q(G) = \min_{P \in \mathcal{P}_q} \frac{1}{\alpha_G^2} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} e_G(V_i, V_j),$$

the minimum c-balanced q-way cut density of G is

$$f_q^c(G) = \min_{P \in \mathcal{P}_q^c} \frac{1}{\alpha_G^2} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} e_G(V_i, V_j),$$

and the minimum $\mathbf{a}$-balanced q-way cut density of G is

$$f_q^{\mathbf{a}}(G) = \min_{P \in \mathcal{P}_q^{\mathbf{a}}} \frac{1}{\alpha_G^2} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} e_G(V_i, V_j).$$

Occasionally, we want to penalize cluster volumes that significantly differ. We therefore introduce the notions of minimum normalized cut densities.

Definition 4 *The minimum normalized q-way cut density of G is*

$$\mu_q(G) = \min_{P \in \mathcal{P}_q} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \frac{1}{\alpha_{V_i} \cdot \alpha_{V_j}} \cdot e_G(V_i, V_j),$$

the minimum normalized c-balanced q-way cut density of G is

$$\mu_q^c(G) = \min_{P \in \mathcal{P}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \frac{1}{\alpha_{V_i} \cdot \alpha_{V_j}} \cdot e_G(V_i, V_j),$$

and the minimum normalized **a**-*balanced q-way cut density of G is*

$$\mu_q^{\mathbf{a}}(G) = \min_{P \in \mathcal{P}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \frac{1}{\alpha_{V_i} \cdot \alpha_{V_j}} \cdot e_G(V_i, V_j).$$

Proposition 2 *$f_q(G)$ is testable for any $q \le |V(G)|$.*

However, this statement is not of much use, since $f_q(G_n) \to 0$ as $n \to \infty$, in the lack of dominant vertex-weights. Indeed, the minimum q-way cut density is trivially estimated from above by

$$f_q(G_n) \le (q-1)\frac{\alpha_{max}(G_n)}{\alpha_{G_n}} + \binom{q-1}{2}\left(\frac{\alpha_{max}(G_n)}{\alpha_{G_n}}\right)^2$$

that tends to 0 provided $\alpha_{\max}(G_n)/\alpha_{G_n} \to 0$ as $n \to \infty$.

Proposition 3 *$f_q^{\mathbf{a}}(G)$ is testable for any $q \le |V(G)|$ and probability vector* **a** *over $[q]$.*

Proposition 3 and Fact 3 together imply the following less obvious statement.

Proposition 4 *$f_q^c(G)$ is testable for any $q \le |V(G)|$ and $c \le 1/q$.*

Now consider the normalized density $\mu_q(G) = \min_{P \in \mathcal{P}_q} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij}(G/P)$. It is not testable as the following example shows: let $q = 2$ and G_n be a simple graph on n vertices such that about $\sqrt{n}$ vertices are connected with a single edge to the remaining vertices that form a complete graph. Then $\mu_2(G_n) \to 0$, but randomizing a sufficiently large part of the graph, with high probability, it will be a subgraph of the complete graph, whose minimum normalized 2-way cut density is of constant order. In the $q = 2$, $\alpha_i = 1$ $(\forall i)$ special case, $\mu_2(G)$ of a regular graph G is its normalized 2-way cut; consequently, the normalized cut is not testable either.

However, balanced versions of the minimum normalized q-way cut density are testable.

Proposition 5 *$\mu_q^{\mathbf{a}}(G)$ is testable for any $q \le |V(G)|$ and probability vector* **a** *over $[q]$.*

Proposition 6 *$\mu_q^c(G)$ is testable for any $q \le |V(G)|$ and $c \le 1/q$.*

6 Convergence of contingency tables

[This section can be skipped]

Now, we will extend the above theory to rectangular arrays with nonnegative, bounded entries. A statistic, defined on a contingency table, is testable if it can be consistently estimated based on a smaller, but still sufficiently large table which is selected randomly from the original one in an appropriate manner. By the above randomization, classical multivariate methods can be carried out on a smaller part of the array. This fact becomes important when our task is to discover the structure of large and evolving arrays, such as microarrays, social, and communication networks. Special block structures behind large tables are also discussed from the point of view of stability and spectra. In order to recover the structure of large rectangular arrays, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes. For the above causes, convergence and distance of general normalized arrays is introduced.

Let $\boldsymbol{C} = \boldsymbol{C}_{m\times n}$ be a contingency table on row set $Row_C = \{1,\dots,m\}$ and column set $Col_C = \{1,\dots,n\}$. The nonnegative, real entries c_{ij}'s are thought of as

associations between the rows and columns, and they are normalized such that $0 \le c_{ij} \le 1$. Sometimes we have *binary* tables of entries 0 or 1. We may assign positive weights $\alpha_1,\dots,\alpha_m$ to the rows and $\beta_1,\dots,\beta_n$ to the columns expressing individual importance of the categories embodied by the rows and columns. (In correspondence analysis, these are the row- and column-sums.) A contingency table is called *simple* if all the row- and column-weights are equal to 1. Assume that C does not contain identically zero rows or columns, moreover C is dense in the sense that the number of nonzero entries is comparable with mn. Let $\mathcal{C}$ denote the set of such tables (with any natural numbers m and n).

Consider a simple binary table $\boldsymbol{F}_{a\times b}$ and maps $\Phi : Row_F \to Row_C$, $\Psi : Col_F \to Col_C$; further

$$\alpha_\Phi := \prod_{i=1}^{a} \alpha_{\Phi(i)}, \quad \beta_\Psi := \prod_{j=1}^{b} \beta_{\Psi(j)}, \quad \alpha_C := \sum_{i=1}^{m} \alpha_i, \quad \beta_C := \sum_{j=1}^{n} \beta_j.$$

Definition 5 *The $\boldsymbol{F} \to \boldsymbol{C}$ homomorphism density is*

$$t(\boldsymbol{F},\boldsymbol{C}) = \frac{1}{(\alpha_C)^a(\beta_C)^b} \sum_{\Phi,\Psi} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}.$$

If $\boldsymbol{C}$ is simple, then $t(\boldsymbol{F},\boldsymbol{C}) = \frac{1}{m^a n^b} \sum_{\Phi,\Psi} \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}$. If, in addition, C is binary too, then $t(\boldsymbol{F},\boldsymbol{C})$ is the probability that a random map $\boldsymbol{F} \to \boldsymbol{C}$ is a homomorphism (preserves the 1's). The maps Φ and Ψ correspond to sampling a rows and b columns out of Row_C and Col_C with replacement, respectively. In case of simple $\boldsymbol{C}$ it means uniform sampling, otherwise the rows and columns are selected with probabilities proportional to their weights.

The following simple binary random table $\xi(a\times b, \boldsymbol{C})$ will play an important role in the definition of testable contingency table parameters. Select a rows and b columns of $\boldsymbol{C}$ with replacement, with probabilities α_i/α_C $(i = 1,\dots,m)$ and β_j/β_C $(j = 1,\dots,n)$, respectively. If the ith row and jth column of C are selected, they will be connected by 1 with probability c_{ij} and 0, otherwise, independently of the other selected row–column pairs, conditioned on the selection of the rows and columns. For large m and n, $\mathbb{P}(\xi(a \times b, \boldsymbol{C}) = F)$ is very close to $t(\boldsymbol{F}, \boldsymbol{C})$ that resembles a likelihood function. (The more precise formulation with induced and injective homomorphisms is to be found in [3]).

Definition 6 *We say that the sequence $(\boldsymbol{C}_{m\times n})$ of contingency tables is convergent if the sequence $t(\boldsymbol{F}, \boldsymbol{C}_{m\times n})$ converges for any simple binary table $\boldsymbol{F}$ as $m, n \to \infty$.*

The convergence means that the tables $\boldsymbol{C}_{m\times n}$ become more and more similar in small details as they are probed by smaller 0-1 tables $(m, n \to \infty)$.

The limit object is a measurable function $U : [0,1]^2 \to [0,1]$ and we call it *contingon*, which is the non-symmetric generalization of a graphon (see Section 1) and was introduced in [3]. The step-function contingon U_C is assigned to $\boldsymbol{C}$ in the following way: the sides of the unit square are divided into intervals $I_1, \dots, I_m$ and $J_1, \dots, J_n$ of lengths $\alpha_1/\alpha_C, \dots, \alpha_m/\alpha_C$ and $\beta_1/\beta_C, \dots, \beta_n/\beta_C$, respectively; then over the rectangle $I_i \times J_j$ the step-function takes on the value c_{ij}.

In fact, the above convergence of contingency tables can be formulated in terms of the cut distance. First we define it for contingons.

Definition 7 *The cut distance between the contingons U and V is*

$$\delta_\square(U, V) = \inf_{\mu,\nu} \|U - V^{\mu,\nu}\|_\square \tag{6}$$

where the cut-norm of the contingon U is defined by

$$\|U\|_\square = \sup_{S,T \subset [0,1]} \left| \iint_{S\times T} U(x,y)\, dx\, dy \right|,$$

and the infimum in (6) is taken over all measure preserving bijections $\mu, \nu : [0,1] \to [0,1]$, while $V^{\mu,\nu}$ denotes the transformed V after performing the measure preserving bijections μ and ν on the sides of the unit square, respectively.

An equivalence relation is defined over the set of contingons: two contingons belong to the same class if they can be transformed into each other by measure preserving map, i.e. their cut distance is zero. In the sequel, we consider contingons modulo measure preserving maps, and under contingon we understand the whole equivalence class.

Definition 8 *The cut distance between the contingency tables $\boldsymbol{C}, \boldsymbol{C}' \in \mathcal{C}$ is*

$$\delta_\square(\boldsymbol{C}, \boldsymbol{C}') = \delta_\square(U_C, U_{C'}).$$

By the above remarks, this distance of $\boldsymbol{C}$ and $\boldsymbol{C}'$ is indifferent to permutations of the rows or columns of $\boldsymbol{C}$ and $\boldsymbol{C}'$. In the special case when $\boldsymbol{C}$ and $\boldsymbol{C}'$ are of the same size, $\delta_\square(\boldsymbol{C}, \boldsymbol{C}')$ is $\frac{1}{mn}$ times the usual cut distance of matrices, based on the cut-norm (see Definition **??**).

The following reversible relation between convergent contingency table sequences and contingons also holds, as a rectangular analogue of Fact 1.

Fact 5 *For any convergent sequence $(\boldsymbol{C}_{m\times n}) \subset \mathcal{C}$ there exists a contingon such that $\delta_\square(U_{C_{m\times n}}, U) \to 0$ as $m, n \to \infty$. Conversely, any contingon can be obtained as the limit of a sequence of contingency tables in $\mathcal{C}$. The limit of a convergent contingency table sequence is essentially unique: if $\boldsymbol{C}_{m\times n} \to U$, then also $\boldsymbol{C}_{m\times n} \to U'$ for precisely those contingons U' for which $\delta_\square(U, U') = 0$.*

It also follows that a sequence of contingency tables in $\mathcal{C}$ is convergent if and only if it is a Cauchy sequence in the metric $\delta_\square$.

A simple binary random $a \times b$ table $\xi(a \times b, U)$ can also be randomized based on the contingon U in the following way. Let $X_1, \dots, X_a$ and $Y_1, \dots, Y_b$ be i.i.d., uniformly distributed random numbers on [0,1]. The entries of $\xi(a \times b, U)$ are independent Bernoulli random variables, namely the entry in the ith row and jth column is 1 with probability $U(X_i, Y_j)$ and 0, otherwise. It is easy to see that the distribution of the previously defined $\xi(a \times b, \boldsymbol{C})$ and that of $\xi(a \times b, U_C)$ is the same. Further, $\delta_\square(\boldsymbol{C}_{m\times n}, \xi(a \times b, \boldsymbol{C}_{m\times n}))$ tends to 0 in probability, for fixed a and b as $m, n \to \infty$.

Note, that in the above way, we can theoretically randomize an infinite simple binary table $\xi(\infty \times \infty, U)$ out of the contingon U by generating countably infinitely many i.i.d. uniform random numbers on [0,1]. The distribution of the infinite binary array $\xi(\infty \times \infty, U)$ is denoted by $\mathbb{P}_U$. Because of the symmetry of the construction, this is an *exchangeable* array in the sense that the joint distribution of its entries is invariant under permutations of the rows and columns. Furthermore, any exchangeable binary array is a mixture of such $\mathbb{P}_U$'s. More precisely, the Aldous–Hoover Representation Theorem (see [15]) states that for every infinite exchangeable binary array ξ there exists a probability distribution μ (over the contingons) such that $\mathbb{P}(\xi \in A) = \int \mathbb{P}_U(A)\, \mu(dU)$.

A function $f : \boldsymbol{C} \to \mathbb{R}$ is called a *contingency table parameter* if it is invariant under isomorphism and scaling of the rows/columns. In fact, it is a statistic evaluated on the table, and hence, we are interested in contingency table parameters that are not sensitive to minor changes in the entries of the table.

Definition 9 *A contingency table parameter f is testable if for every $\varepsilon > 0$ there are positive integers a and b such that if the row- and column-weights of $\boldsymbol{C}$ satisfy*

$$\max_i \frac{\alpha_i}{\alpha_C} \le \frac{1}{a}, \qquad \max_j \frac{\beta_j}{\beta_C} \le \frac{1}{b}, \tag{7}$$

then

$$\mathbb{P}(|f(\boldsymbol{C}) - f(\xi(a \times b, \boldsymbol{C}))| > \varepsilon) \le \varepsilon.$$

Consequently, such a contingency table parameter can be consistently estimated based on a fairly large sample. Now, we introduce some equivalent statements of the testability, indicating that a testable parameter depends continuously on the whole table. This is the generalization of Theorem 1.

Theorem 4 *For a testable contingency table parameter f the following are equivalent:*

- *For every $\varepsilon > 0$ there are positive integers a and b such that for every contingency table $\boldsymbol{C} \in \mathcal{C}$ satisfying the condition (7),*

$$|f(\boldsymbol{C}) - \mathbb{E}(f(\xi(a \times b, \boldsymbol{C})))| \le \varepsilon.$$

- *For every convergent sequence $(\boldsymbol{C}_{m\times n})$ of contingency tables with no dominant row- or column-weights, $f(\boldsymbol{C}_{m\times n})$ is also convergent ($m, n \to \infty$).*
- *f is continuous in the cut distance.*

For example, in the case of a simple binary table, the singular spectrum is testable, since $\boldsymbol{C}_{m\times n}$ can be regarded as part of the adjacency matrix of a bipartite graph on $m+n$ vertices, where Row_C and Col_C are the two independent vertex sets; further, the ith vertex of Row_C and the jth vertex of Col_C are connected by an edge if and only if $c_{ij} = 1$. The non-zero real eigenvalues of the symmetric $(m+n) \times (m+n)$ adjacency matrix of this bipartite graph are the numbers $\pm s_1, \dots, \pm s_r$, where $s_1, \dots, s_r$ are the non-zero singular values of $\boldsymbol{C}$, and $r \leq \min\{m, m\}$ is the rank of $\boldsymbol{C}$ (see Proposition **??**). Consequently, the convergence of the adjacency spectra implies the convergence of the singular spectra. Therefore, by Theorem 4, any property of a large contingency table based on its SVD (e.g., correspondence decomposition) can be concluded from a smaller part of it.

Using the notation of Section **??**, analogously to the symmetric case, it can be proved that special blown-up tables (see Definition **??**) burdened with a general kind of noise (see Definition **??**) are convergent.

Proposition 7 *For any sequence $\boldsymbol{W}_{m\times n}$ of rectangular Wigner-noises*

$$\lim_{m,n\to\infty} \|U_{\boldsymbol{W}_{m\times n}}\|_{\square} = 0$$

almost surely, where $(U_{\boldsymbol{W}_{m\times n}})$ is the step-function contingon assigned to $\boldsymbol{W}_{m\times n}$.

Now, let us fix the pattern-matrix matrix $\boldsymbol{P}_{a\times b}$ and blow it up to obtain matrix $\boldsymbol{B}_{m\times n}$.

Proposition 8 *Let the block sizes of the blown-up matrix $\boldsymbol{B}_{m\times n}$ be $m_1, \dots, m_a$ horizontally, and $n_1, \dots, n_b$ vertically ($\sum_{i=1}^{a} m_i = m$ and $\sum_{j=1}^{b} n_j = n$). Let $\boldsymbol{A}_{m\times n} = \boldsymbol{B}_{m\times n} + \boldsymbol{W}_{m\times n}$ and $m, n \to \infty$ is such a way that $m_i/m \to r_i$ ($i = 1, \dots, a$), $n_j/n \to q_j$ ($j = 1, \dots, b$), where r_i's and q_j's are fixed ratios. Under these conditions, the 'noisy' contingency table sequence $(\boldsymbol{A}_{m\times n})$ converges almost surely.*

In many applications we are looking for clusters of the rows and columns of a rectangular array such that the the densities within the cross-products of the clusters be as homogeneous as possible. For example, in microarray analysis we are looking for clusters of genes and conditions such that genes of the same cluster equally influence conditions of the same cluster. The following theorem ensures the existence of such a structure with possibly many clusters. However, the number of clusters does not depend on the size of the array, it merely depends on the accuracy of the approximation. The following statement is a straightforward generalization of the Weak Regularity Lemma 1.

Proposition 9 *For every $\varepsilon > 0$ and $\boldsymbol{C}_{m\times n} \in \mathcal{C}$ there exists a blown-up matrix $\boldsymbol{B}_{m\times n}$ of an $a \times b$ pattern matrix with $a + b \leq 4^{1/\varepsilon^2}$ (independently of m and n) such that $\delta_{\square}(\boldsymbol{C}, \boldsymbol{B}) \leq \varepsilon$.*

The statement can be proved by embedding $\boldsymbol{C}$ into the adjacency matrix of an edge-weighted bipartite graph. The statement itself is closely related to the testability of the following contingency table parameter. For fixed integers $1 \le a \ll m$ and $1 \le b \ll n$,

$$S_{a,b}^2(\boldsymbol{C}) = \min_{\substack{R_1,\dots,R_a \\ C_1,\dots,C_b}} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k \in R_i} \sum_{\ell \in C_j} (c_{k\ell} - \bar{c}_{i,j})^2 \tag{8}$$

where the minimum is taken over balanced a- and b-partitions $R_1, \dots, R_a$ and $C_1, \dots, C_b$ of Row_C and Col_C, respectively; further,

$$\bar{c}_{i,j} = \frac{1}{|R_i| \cdot |C_j|} \sum_{k \in R_i} \sum_{\ell \in C_j} c_{kl}$$

is the center of the bicluster $R_i \times C_j$ $(i = 1, \dots, a;\ j = 1, \dots, b)$. Note that, instead of $c_{k\ell}$, we may take $\alpha_k \beta_\ell c_{k\ell}$ in the row- and column-weighted case, provided there are no dominant rows/columns.

An objective function reminiscent of (8) is minimized in [21, 29] in a more general block clustering problem, where not only Cartesian product type biclusters ($R_i \times C_j$, also called chess-board patterns in [23]) are allowed, but the contingency table is divided into disjoint rectangles forming as homogeneous blocks of the heterogeneous data as possible. (8) can as well be regarded as within-custer variance in a Two-way ANOVA setup. In [29], applications to microarrays is presented, where the rows correspond to genes, the columns to different conditions, whereas the entries are expression levels of genes under the distinct conditions. In this framework, biclusters identify subsets of genes sharing similar expression patterns across subsets of conditions. Recall that in Subsection **??** we looked for regular cluster pairs by means of spectral methods, but there we used Cartesian product type biclusters, moreover, the number of row and column clusters was the same.

The *Gardner–Ashby's connectance* c_n of a not necessarily symmetric, quadratic array $\boldsymbol{A}_{n \times n}$ is the percentage of nonzero entries in the matrix, that is the ratio of actual row-column interactions to all possible ones in the network. In social and ecological models, a random array $\boldsymbol{A}_{n \times n}$ of independent entries is considered. Assume that the entries have symmetric distribution (consequently, zero expectation) and common variance σ_n^2, where σ_n is called *average interaction strength.* The *stability* of the system is characterized by the stability of the equilibrium solution $\mathbf{0}$ of the differential equation $d\mathbf{x}/dt = \boldsymbol{A}_{n \times n}\mathbf{x}$ (sometimes this is achieved by linearization techniques in the neighborhood of the equilibrium solution). Based on Wigner's famous semicircle law (see Theorem **??**), [30] proved that the equilibrium solution is stable in the $\sigma_n^2 n c_n < 1$, and unstable in the $\sigma_n^2 n c_n > 1$ case; further, the transition region between stability and instability becomes narrow as $n \to \infty$. Hence, it seems that high connectance and high interaction strength destroy stability, but only in this simple model. If $\boldsymbol{A}_{n \times n}$ is a block matrix, like a noisy matrix before, it has some structural, possibly complex eigenvalues (see [22]). If all their real parts are negative, the system is stable, see [18]. In fact, in many natural ecosystems and other networks the interactions are arranged in blocks.

References

[1] Arora S, Karger D and Karpinski M 1995 Polynomial time approximation schemes for dense instances of NP-hard problems. In *Proc. 27th Annual ACM Symposium on the Theory of Computing (STOC 1995)*, pp. 284-293.

[2] Ballester C, Calvó-Armengol A and Zenou Y 2006 Who's who in networks. Wanted: The key player. *Econometrica* **74** (5), 1403–1417.

[3] Bolla M 2010 Statistical inference on large contingency tables: convergence, testability, stability. In *Proc. 19th International Conference on Computational Statistics (COMPSTAT 2010), Paris* (Lechevallier Y and Saporta G eds), pp. 817-824. Physica-Verlag, Springer.

[4] Bolla M, Kói T and Krámli A 2012 Testability of minimum balanced multiway cut densities. *Discret. Appl. Math.* **160**, 1019–1027.

[5] Borgs C, Chayes JT, Lovász L, T.-Sós V and Szegedy B 2006 Graph limits and parameter testing. In *Proc. 38th Annual ACM Symposium on the Theory of Computing (STOC 2006)*, pp. 261–270.

[6] Borgs C, Chayes JT, Lovász L, T.-Sós V and Vesztergombi K 2008 Convergent graph sequences I: Subgraph Frequencies, metric properties, and testing. *Advances in Math.* **219**, 1801–1851.

[7] Borgs C, Chayes JT, Lovász L, T.-Sós V and Vesztergombi K 2011 Limits of randomly grown graph sequences. *European J. Comb.* **32**, 985–999.

[8] Borgs C, Chayes JT, Lovász L, T.-Sós V and Vesztergombi K 2012 Convergent sequences of dense graphs II: Multiway cuts and statistical physics. *Ann. Math.* **176**, 151–219.

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

[10] Chung F and Lu L. 2005 Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics* **3**, 79–127.

[11] Clauset A, Newman MEJ and Moore C 2004 Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111.

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

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

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

[15] Diaconis P and Janson S 2008 Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (VII. Ser.) **28**, 33–61.

[16] 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.

[17] Elek G 2008 L^2-spectral invariants and convergent sequences of finite graphs. *J. Funct. Anal.* **254**, 2667–2689.

[18] Érdi P and Tóth J 1990 What is and what is not stated by the May-Wigner theorem. *J. Theor. Biol.* **145**, 137–140.

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

[20] Goldreich O, Goldwasser S and Ron D 1998 Property testing and its connection to learning and approximation. *J. ACM* **45**, 653–750.

[21] Hartigan JA 1972 Direct clustering of a data matrix. *J. Am. Stat. Assoc.* **67**, 123–129.

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

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

[24] 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.

[25] Lovász L and Szegedy B 2006 Limits of dense graph sequences. *J. Comb. Theory B* **96**, 933–957.

[26] Lovász L and Szegedy B 2007 Szemerédi's Lemma for the analyst. *Geom. Func. Anal.* **17**, 252–270.

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

[28] Lovász L and Szegedy B 2011 Finitely forcible graphons. *J. Comb. Theory B.* **101**, 269–301.

[29] Madeira SC and Oliveira AL 2004 Biclustering algorithms for biological data analysis: A survey. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **1** (1), 24–45.

[30] May RM 1972 Will a large complex system be stable? *Nature* **238**, 413–414.

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