Testability of minimum balanced multiway cut densities

Marianna Bolla [a,*], Tamás Kói [a], András Krámli [b]

[a] *Institute of Mathematics, Budapest University of Technology and Economics*

[b] *Bolyai Institute, University of Szeged*

Abstract

Testability of certain balanced minimum multiway cut densities is investigated for vertex- and edge-weighted graphs with no dominant vertex-weights. We apply the results for fuzzy clustering and noisy graph sequences.

Key words: Testable weighted graph parameters, Minimum balanced multiway cuts, Fuzzy clustering, Wigner-noise
1991 MSC: 05C90, 62H30, 68R10

---

1 Introduction

The testability of the maximum cut density is stated in [4] based on earlier algorithmic results of [1,8,7]. Roughly speaking, testable parameters are nonparametric statistics on a large graph that can be approximated by appropriate sampling. We are 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

* Corresponding author. Research supported in part by the Hungarian National Research Grants OTKA 76481 and OTKA-NKTH 77778.

Email addresses: `marib@math.bme.hu` (Marianna Bolla), `koitomi@math.bme.hu` (Tamás Kói), `kramli@informatika.ilab.sztaki.hu` (András Krámli).

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 large networks' vertices with sparse inter-cluster connections, where the clusters do not differ significantly in sizes.

The organization of the paper is as follows. In Section 2, we adapt the notion and equivalent statements of testability introduced by Borgs, Chayes, Lovász, Sós, and Vesztergombi [5] to vertex- and edge-weighted graphs with no dominant vertex-weights. In Section 3, testability of different kinds of minimum multiway cut densities is investigated using the equivalent statements of Section 2 and statistical physics facts of Borgs, Chayes, Lovász, Sós, and Vesztergombi [6]. In Section 4, continuous extensions of testable minimum multiway cut densities to graphons are constructed and applied for fuzzy clustering. In Section 5, convergence of special noisy weighted graph sequences is established.

2 Testable weighted graph parameters

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 $\beta_{ij} = \beta_{ji} \in [0,1]$ $(i, j \in [n])$, while the vertex-weights $\alpha_i > 0$ $(i \in [n])$ indicate relative significance of the vertices. Let $\mathcal{G}$ denote the set of all such weighted graphs.

We will intensively use the following notions of [5]. The *volume* of $G \in \mathcal{G}$ is $\alpha_G = \sum_{i=1}^n \alpha_i$, while that of the vertex-subset T is $\alpha_T = \sum_{i \in T} \alpha_i$. Further, $e_G(S,T) = \sum_{s \in S} \sum_{t \in T} \alpha_s \alpha_t \beta_{st}$ denotes the *weighted cut* between the vertex-subsets S and T. The *homomorphism density* between the simple graph F (on vertex set $V(F) = [k]$) and the above weighted graph G is

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

In their theory, the authors of [5] indirectly 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* $\xi(k, G)$.

The weighted graph sequence (G_n) is said to-be (left-)convergent, if the sequence $t(F, G_n)$ converges for any simple graph F $(n \to \infty)$. As other kinds of convergence are not discussed here, in the sequel we omit the word left and simply use *convergence*. The authors in [9] 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. Graphons are considered modulo measure preserving maps, and under graphon the whole equivalence class is understood. We also cite the definition of the cut-distance between weighted graphs and between a graphon and a graph:

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

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. The testability results of [5] for simple graphs remain valid if we consider weighted graph sequences (G_n) with *no dominant vertex-weights*, that is $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0$ as $n \to \infty$.

To use this condition imposed on the vertex-weights, we slightly modify the definition of a testable graph parameter for weighted graphs.

Definition 1 *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*

$$\max_i \frac{\alpha_i(G)}{\alpha_G} \le \frac{1}{k}, \tag{2}$$

then

$$\mathbb{P}(|f(G) - f(\xi(k, G))| > \varepsilon) \leq \varepsilon, \tag{3}$$

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

Note that for simple G, the condition (2) implies that $|V(G)| \geq k$, giving back the definition in [5]. 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. Now, we introduce some equivalent statements of the testability, indicating that a testable parameter depends continuously on the whole graph. This is the straightforward generalization of Theorem 6.1 of [5] stated for simple graphs. We omit the proof as it uses the ideas of the proof in [5], where some details for such a generalization are also elaborated.

Theorem 2 *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 $\max_i \alpha_i(G)/\alpha_G \leq 1/k$,

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

(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 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 $\max_i \alpha_i(G_1)/\alpha_{G_1} \leq 1/n_0$, $\max_i \alpha_i(G_2)/\alpha_{G_2} \leq 1/n_0$, and $\delta_\square(G_1, G_2) < \varepsilon_0$, then $|f(G_1) - f(G_2)| < \varepsilon$.

3 Balanced multiway cuts

In this section we investigate the testability of some balanced multiway cut densities. For the proofs we use Theorem 2 and some notions of statistical physics in the same way as in [6].

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 $q \leq n$ 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. 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

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

where

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

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 $\mathbf{J}$ 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* of such a spin configuration is

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

where $\mathbf{J}$ 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 shall use only special $\mathbf{J}$ 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 ($c \le 1/q$). Let $\mathcal{P}_q^c$ denote the set of q-partitions of V such that $\frac{\alpha_{V_i}}{\alpha_G} \ge c \quad (i \in [q])$, or equivalently, $c \le \frac{\alpha_{V_i}}{\alpha_{V_j}} \le \frac{1}{c}$ $(i \ne j)$. 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 $\left|\frac{\alpha_{V_i}}{\alpha_G} - a_i\right| \le \frac{\alpha_{\max}(G)}{\alpha_G} \quad (i = 1, \dots, q)$. 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 $\mathbf{J}$ ($\mathbf{h} = \mathbf{0}$) is

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

Fact 3 *In Theorem 2.14 of [6] it is proved that 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* $\mathbf{J}$. *Also, it is equivalent to the convergence of its* q*-quotients in Hausdorff distance for any* q.

Fact 4 *Under the same conditions, Theorem 2.15 of [6] states that the convergence of the above* (G_n) *implies the convergence of its ground state energies for any* q, $\mathbf{J}$, *and* $\mathbf{h}$.

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

Definition 5 *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 6 *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 7 *$f_q(G)$ is testable for any $q \le |V(G)|$.*

PROOF. Observe that $f_q(G)$ is a special ground state energy: $f_q(G) = \hat{\mathcal{E}}_q(G, \mathbf{J}, \mathbf{0})$, where the magnetic field is $\mathbf{0}$ and the $q \times q$ symmetric matrix $\mathbf{J}$ is the following: $J_{ii} = 0$ $(i \in [q])$, further $J_{ij} = -1/2$ $(i \neq j)$. By Fact 4 and the equivalent statement (c) of Theorem 2, the minimum q-way cut density is testable for any q. □

However, this statement is of not 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 8 *$f_q^{\mathbf{a}}(G)$ is testable for any $q \le |V(G)|$ and probability vector* **a** *over $[q]$.*

PROOF. Choose $\mathbf{J}$ as in the proof of Proposition 7. In this way, $f_q^{\mathbf{a}}(G)$ is a special microcanonical ground state energy:

$$f_q^{\mathbf{a}}(G) = \hat{\mathcal{E}}_q^{\mathbf{a}}(G, \mathbf{J}). \tag{4}$$

Hence, by Fact 3, the convergence of (G_n) is equivalent to the convergence of $f_q^{\mathbf{a}}(G_n)$ for any q and any distribution $\mathbf{a}$ over $[q]$. Therefore, by the equivalent statement (c) of Theorem 2, the testability of the minimum $\mathbf{a}$-balanced q-way cut density also follows. □

Proposition 8 and Fact 3 together imply the following less obvious statement.

Theorem 9 *$f_q^c(G)$ is testable for any $q \le |V(G)|$ and $c \le 1/q$.*

PROOF. Theorem 4.7 and Theorem 5.5 of [6] imply that for any two weighted graphs G, G'

$$|\hat{\mathcal{E}}_q^{\mathbf{a}}(G, \mathbf{J}) - \hat{\mathcal{E}}_q^{\mathbf{a}}(G', \mathbf{J})| \le (3/2 + \kappa) \cdot d^{\mathrm{Hf}}(\hat{S}_q(G), \hat{S}_q(G')), \tag{5}$$

where $\kappa = o(\min\{|V(G)|, |V(G')|\})$ is a negligible small constant, provided the number of vertices of G and G' is sufficiently large. By Fact 3, we know that if (G_n) converges, its q-quotients also converge in Hausdorff distance, consequently, form a Cauchy-sequence. This means that for any $\varepsilon > 0$ there is an N_0 such that for $n, m > N_0$: $d^{\mathrm{Hf}}(\hat{S}_q(G_n), \hat{S}_q(G_m)) < \varepsilon$. We want to prove that for $n, m > N_0$: $|f_q^c(G_n) - f_q^c(G_m)| < 2\varepsilon$. On the contrary, suppose that there are $n, m > N_0$ such that $|f_q^c(G_n) - f_q^c(G_m)| \geq 2\varepsilon$. Say, $f_q^c(G_n) \geq f_q^c(G_m)$. Let $A := \{\mathbf{a} : a_i \geq c\,,\; i = 1, \dots, q\}$ is the subset of special c-balanced distributions over $[q]$. On the one hand,

$$f_q^c(G_m) = \min_{\mathbf{a} \in A} f_q^{\mathbf{a}}(G_m) = f_q^{\mathbf{a}^*}(G_m)$$

for some $\mathbf{a}^* \in A$. On the other hand, by (4) and (5), $f_q^{\mathbf{a}^*}(G_n) - f_q^{\mathbf{a}^*}(G_m) \leq (\frac{3}{2} + \kappa)\varepsilon$, that together with the indirect assumption implies that $f_q^c(G_n) - f_q^{\mathbf{a}^*}(G_n) \geq (\frac{1}{2} - \kappa)\varepsilon > 0$ for this $\mathbf{a}^* \in A$. But this contradicts to the fact that $f_q^c(G_n)$ is the minimum of $f_q^{\mathbf{a}}(G_n)$'s over A. Thus, $f_q^c(G_n)$ is also a Cauchy sequence, and being a real sequence, it is also convergent. □

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

Theorem 10 *$\mu_q^{\mathbf{a}}(G)$ is testable for any $q \leq |V(G)|$ and probability vector $\mathbf{a}$ over $[q]$.*

PROOF. By the definition of Hausdorff distance, the convergence of q-quotients guarantees the convergence of

$$\mu_q^{\mathbf{a}}(G) = \min_{P \in \mathcal{P}_q^{\mathbf{a}}} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij}(G/P) \tag{6}$$

for any $\mathbf{a}$ and q in the following way. Let $\hat{\mathcal{S}}_q^{\mathbf{a}}(G)$ denote the set of factor graphs of G with respect to partitions in $\mathcal{P}_q^{\mathbf{a}}$. As a consequence of Lemma 4.5 and Theorem 5.4 of [6], for any two weighted graphs G, G'

$$\max_{\mathbf{a}} d^{Hf}(\hat{\mathcal{S}}_q^{\mathbf{a}}(G), \hat{\mathcal{S}}_q^{\mathbf{a}}(G')) \leq (3 + \kappa) \cdot d^{Hf}(\hat{\mathcal{S}}_q(G), \hat{\mathcal{S}}_q(G')), \tag{7}$$

where $\kappa = o(\min\{|V(G)|, |V(G')|\})$.

By Fact 3, for a convergent graph-sequence (G_n), the sequence $\hat{\mathcal{S}}_q(G_n)$ converges, and by the inequality (7), $\hat{\mathcal{S}}_q^{\mathbf{a}}(G_n)$ also converges in Hausdorff distance for any distribution $\mathbf{a}$ over $[q]$. As they form a Cauchy sequence, $\forall\varepsilon\ \exists N_0$ such that for $n, m > N_0$

$$d^{Hf}(\hat{\mathcal{S}}_q^{\mathbf{a}}(G_n), \hat{\mathcal{S}}_q^{\mathbf{a}}(G_m)) < \varepsilon$$

uniformly for any $\mathbf{a}$. In view of the Hausdorff distance's definition, this means that for any q-quotient $H \in \hat{\mathcal{S}}_q^{\mathbf{a}}(G_n)$ there exists (at least one) q-quotient $H' \in \hat{\mathcal{S}}_q^{\mathbf{a}}(G_m)$, and vice versa, for any $H' \in \hat{\mathcal{S}}_q^{\mathbf{a}}(G_m)$ there exists (at least one) $H \in \hat{\mathcal{S}}_q^{\mathbf{a}}(G_n)$ such that $d_1(H, H') < \varepsilon$. Therefore, the maximum distance between the elements of the above pairs is less than ε. (Note that the symmetry in the definition of the Hausdorff distance is important: the pairing exhausts the sets even if they have different cardinalities.)

Using the fact that the vertex-weights of such a pair H and H' are almost the same (the coordinates of the vector $\mathbf{a}$), by the notation $a = \min_{i\in[q]} a_i$, the following argument is valid for n, m large enough:

$$2a^2 \sum_{i\neq j} |\beta_{ij}(H) - \beta_{ij}(H')| \le \sum_{i,j=1}^{q} a^2 |\beta_{ij}(H) - \beta_{ij}(H')| \le$$
$$\le \sum_{i,j=1}^{q} |\alpha_i(H)\alpha_j(H)\beta_{ij}(H) - \alpha_i(H')\alpha_j(H')\beta_{ij}(H')| = d_1(H, H') < \varepsilon.$$

Therefore

$$|\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \beta_{ij}(H) - \sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \beta_{ij}(H')| < \frac{\varepsilon}{2a^2} := \varepsilon',$$

and because $\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \beta_{ij}(H)$ and $\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \beta_{ij}(H')$ are individual terms behind the minimum in (6), the above inequality holds for their minima over $\mathcal{P}_q^{\mathbf{a}}$ as well:

$$|\mu_q^{\mathbf{a}}(G_n) - \mu_q^{\mathbf{a}}(G_m)| < \varepsilon'. \tag{8}$$

Consequently, the sequence $\mu_q^{\mathbf{a}}(G_n)$ is a Cauchy sequence, and being a real sequence, it is also convergent. Thus $\mu_q^{\mathbf{a}}$ is testable. □

Proposition 11 *$\mu_q^c(G)$ is testable for any $q \le |V(G)|$ and $c \le 1/q$.*

PROOF. The proof is analogous to that of Theorem 9 using equation (8) instead of equation (5). By the pairing argument of the proof of Theorem 10, the real sequence $\mu_q^c(G_n)$ is a Cauchy sequence, and therefore, convergent. This immediately implies the testability of μ_q^c. □

4 Balanced minimum cuts and fuzzy clustering

In cluster analysis of large data sets, the testable parameters $f_q^c(G)$ and $\mu_q^c(G)$ have the greatest relevance as they require only a slight balancing between the clusters. Now, they are continuously extended to graphons by an explicit construction.

Proposition 12 *Let us define the graphon functional $\tilde{f}_q^c$ in the following way:*

$$\tilde{f}_q^c(W) := \inf_{Q \in \mathcal{Q}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} W(x,y)\, dx\, dy = \inf_{Q \in \mathcal{Q}_q^c} \tilde{f}_q(W; S_1, \dots, S_q) \quad (9)$$

where the infimum is taken over all the c-balanced Lebesgue-measurable partitions $Q = (S_1, \dots, S_q)$ of [0,1]. For these, $\sum_{i=1}^{q} \lambda(S_i) = 1$ and $\lambda(S_i) \ge c$ ($i \in [q]$), where λ denotes the Lebesgue-measure, and $\mathcal{Q}_q^c$ denotes the set of c-balanced q-partitions of [0,1]. We state that $\tilde{f}_q^c$ is the extension of f_q^c in the following sense: If (G_n) is a convergent weighted graph sequence with no dominant vertex-weights and edge-weights in the [0,1] interval, then denoting by W the essentially unique limit graphon of the sequence, $f_q^c(G_n) \to \tilde{f}_q^c(W)$ as $n \to \infty$.

PROOF. First we show that $\tilde{f}_q^c$ is continuous in the cut-distance. As $\tilde{f}_q^c(W) = \tilde{f}_q^c(W^\nu)$, where $\nu : [0,1] \to [0,1]$ is a measure preserving bijection, it suffices to prove that to any ε we can find ε' such that for any two graphons W, U with $\|W - U\|_\square < \varepsilon'$, the relation $|\tilde{f}_q^c(W) - \tilde{f}_q^c(U)| < \varepsilon$ also holds. Indeed, by the definition of the cut-norm, for any Lebesgue-measurable q-partition $(S_1, \dots, S_q)$ of [0,1], the relation

$$\left| \iint_{S_i \times S_j} (W(x,y) - U(x,y))\, dx\, dy \right| \le \varepsilon' \qquad (i \ne j)$$

holds. Summing up for the $i \ne j$ pairs

$$\left| \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} W(x,y)\, dx\, dy - \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} U(x,y)\, dx\, dy \right| \le \binom{q}{2} \varepsilon'.$$

Therefore

$$\inf_{(S_1,\dots,S_q) \in \mathcal{Q}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} W(x,y)\, dx\, dy \ge \inf_{(S_1,\dots,S_q) \in \mathcal{Q}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} U(x,y)\, dx\, dy - \binom{q}{2} \varepsilon'$$

and vice versa,

$$\inf_{(S_1,\dots,S_q) \in \mathcal{Q}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} U(x,y)\, dx\, dy \ge \inf_{(S_1,\dots,S_q) \in \mathcal{Q}_q^c} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \iint_{S_i \times S_j} W(x,y)\, dx\, dy - \binom{q}{2} \varepsilon'.$$

Consequently the absolute difference of the two infima is bounded from above by $\binom{q}{2}\varepsilon'$. Thus, $\varepsilon' = \varepsilon / \binom{q}{2}$ will do.

Let (G_n) be a convergent weighted graph sequence with no dominant vertex-weights and edge-weights in the [0,1] interval. By Theorem 3.9 of [5], there is an essentially unique graphon W such that $G_n \to W$, i.e., $\delta_\Box(W_{G_n}, W) \to 0$ as $n \to \infty$. By the continuity of $\tilde{f}_q^c$,

$$\tilde{f}_q^c(W_{G_n}) \to \tilde{f}_q^c(W), \qquad n \to \infty. \tag{10}$$

Suppose that

$$\tilde{f}_q^c(W_{G_n}) = \tilde{f}_q(W_{G_n}; S_1^*, \dots, S_q^*),$$

that is the infimum in (9) is attained at the c-balanced Lebesgue-measurable q-partition $(S_1^*, \dots, S_q^*)$ of [0,1].

Let G_{nq}^* be the q-fold blown-up of G_n with respect to $(S_1^*, \dots, S_q^*)$. It is a weighted graph on at most nq vertices defined in the following way. Let $I_1, \dots, I_n$ be consecutive intervals of [0,1] such that $\lambda(I_j) = \alpha_j(G_n)$, $j = 1, \dots, n$. The weight of the vertex labeled by ju of G_{nq}^* is $\lambda(I_j \cap S_u^*)$, $u \in [q]$, $j \in [n]$, while the edge-weights are $\beta_{ju,iv}(G_{nq}^*) = \beta_{ji}(G_n)$. Trivially, the graphons W_{G_n} and $W_{G_{nq}^*}$ essentially define the same stepfunction, hence $\tilde{f}_q^c(W_{G_n}) = \tilde{f}_q^c(W_{G_{nq}^*})$. Therefore, by (10),

$$\tilde{f}_q^c(W_{G_{nq}^*}) \to \tilde{f}_q^c(W), \qquad n \to \infty. \tag{11}$$

As $\delta_\Box(G_n, G_{nq}^*) = \delta_\Box(W_{G_n}, W_{G_{nq}^*}) = 0$, by part (e) of Theorem 2 it follows that

$$|f_q^c(G_{nq}^*) - f_q^c(G_n)| \to 0, \qquad n \to \infty. \tag{12}$$

Finally, by the construction of G_{nq}^*, $\tilde{f}_q^c(W_{G_{nq}^*}) = f_q^c(G_{nq}^*)$, and hence,

$$|f_q^c(G_n) - \tilde{f}_q^c(W)| \le |f_q^c(G_n) - f_q^c(G_{nq}^*)| + |\tilde{f}_q^c(W_{G_{nq}^*}) - \tilde{f}_q^c(W)|$$

which, in view of (11) and (12), implies the required statement. □

The above continuous extension of $f_q^c(G)$ to graphons is essentially unique (the proof of Theorem 6.1 of [5] can be adapted to this situation), and part (d) of Theorem 2 implies that for a weighted graph sequence (G_n) with $\max_i \frac{\alpha_i(G_n)}{\alpha_{G_n}} \to 0$, the limit relation $\tilde{f}_q^c(W_{G_n}) - f_q^c(G_n) \to 0$ also holds as $n \to \infty$. This gives rise to approximate the minimum c-balanced q-way cut density of a weighted graph on a large number of vertices with no dominant vertex-weights by the extended c-balanced q-way cut density of the stepfunction graphon assigned to the graph. In this way, the discrete optimization problem can be formulated as a quadratic programming task with linear equality and inequality constraints, and fuzzy clusters are obtained.

To this end, let us investigate a fixed weighted graph G on n vertices (n is large). To simplify notation we will drop the subscript n, and G in the arguments of the vertex- and edge-weights. As $f_q^c(G)$ is invariant under scaling the vertex-weights, we can suppose that $\alpha_G = \sum_{i=1}^n \alpha_i = 1$. As $\beta_{ij} \in [0,1]$, W_G is uniformly bounded by 1. Recall that $W_G(x,y) = \beta_{ij}$, if $x \in I_i$, $y \in I_j$, where $\lambda(I_j) = \alpha_j$ ($j = 1,\dots,n$) and $I_1,\dots,I_n$ are consecutive intervals of [0,1].

For fixed q and $c \le 1/q$, $f_q(G;V_1,\dots,V_q) = \frac{1}{\alpha_G^2}\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} e_G(V_i,V_j)$ is a function taking on discrete values over c-balanced q-partitions $P = (V_1,\dots,V_q) \in \mathcal{P}_q^c$ of the vertices of G. As $n \to \infty$, this function approaches $\tilde{f}_q(W_G;S_1,\dots,S_q)$ that is already a continuous function over c-balanced q-partitions $(S_1,\dots,S_q) \in \mathcal{Q}_q^c$ of [0,1]. In fact, this continuous function is a multilinear function of the variable

$$\mathbf{x} = (x_{11},\dots,x_{1n},x_{21},\dots,x_{2n},\dots,x_{q1},\dots,x_{qn})^T \in \mathbb{R}^{nq}$$

where the coordinate indexed by ij is

$$x_{ij} = \lambda(S_i \cap I_j), \quad j = 1,\dots,n; \quad i = 1,\dots q.$$

Hence,

$$\tilde{f}_q(W_G;S_1,\dots,S_q) = \tilde{f}_q(\mathbf{x}) = \sum_{i=1}^{q-1}\sum_{i'=i+1}^{q}\sum_{j=1}^{n}\sum_{j'=1}^{n} x_{ij}x_{i'j'}\beta_{jj'} = \frac{1}{2}\mathbf{x}^T(\mathbf{A}\otimes\mathbf{B})\mathbf{x},$$

where – denoting by $\mathbf{1}_{q\times q}$ and $\mathbf{I}_{q\times q}$ the $q\times q$ all 1's and the identity matrix, respectively – the eigenvalues of the $q\times q$ symmetric matrix $\mathbf{A} = \mathbf{1}_{q\times q} - \mathbf{I}_{q\times q}$ are the number $q-1$ and -1 with multiplicity $q-1$, while those of the $n\times n$ symmetric matrix $\mathbf{B} = (\beta_{ij})$ are $\lambda_1 \ge \dots \ge \lambda_n$. The eigenvalues of the Kronecker-product $\mathbf{A}\otimes\mathbf{B}$ are the numbers $(q-1)\lambda_i$ ($i = 1,\dots,n$) and $-\lambda_i$ with multiplicity $q-1$ ($i = 1,\dots,n$). Therefore the above quadratic form is indefinite.

Hence, we have the following *quadratic programming* task:

$$\begin{aligned} &\text{minimize} \quad \tilde{f}_q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T(\mathbf{A}\otimes\mathbf{B})\mathbf{x} \\ &\text{subject to} \quad \mathbf{x} \ge 0; \quad \sum_{i=1}^{q} x_{ij} = \alpha_j \quad (j \in [n]); \quad \sum_{j=1}^{n} x_{ij} \ge c \quad (i \in [q]). \end{aligned} \tag{13}$$

The feasible region is the closed convex polytope of (13), and it is, in fact, in an $n(q-1)$-dimensional hyperplane of $\mathbb{R}^{nq}$. The gradient of the objective function $\nabla\tilde{f}_q(\mathbf{x}) = (\mathbf{A}\otimes\mathbf{B})\mathbf{x}$ cannot be $\mathbf{0}$ in the feasible region, provided the weight matrix $\mathbf{B}$, or equivalently, $\mathbf{A}\otimes\mathbf{B}$ is not singular.

The arg-min $\mathbf{x}^*$ of the quadratic programming task (13) is one of the Kuhn–Tucker points (giving relative minima of the indefinite quadratic form over the feasible region), that can be found by numerical algorithms, see [2]. In this way, for large n, we also get fuzzy clustering of the vertices, whereas $x^*_{ij}/\lambda(S_i)$ is the probability that vertex j belongs to cluster i. The index i giving the largest proportion can be regarded as the cluster membership of vertex j. We conjecture that most of the vertices j can be uniquely assigned to a cluster i for which $x^*_{ij} = \alpha_j$, and there will be $\Theta(q)$ vertices with $0 < x^*_{ij} < \alpha_j$ $(i = 1, \dots, q)$. Indeed, the minimum is more likely to be attained at a low-dimensional face, as here a lot of inequality constraints of (13) are satisfied by equalities that causes many coordinates of $\mathbf{x}^*$ to be zeros. On higher dimensional faces the small number of equalities may come from the last q ones.

More generally, the above problem can be solved with other equality or inequality constraints making it possible to solve individual fuzzy clustering problems by quadratic programming. Similarly, the extension of the testable weighted graph parameter $\mu^c_q(G)$ to graphons can be given and used for fuzzy clustering.

5 Noisy graph sequences

Now, we use the above theory for perturbations, showing that special noisy weighted graph sequences converge in the sense of Section 2. If not stated otherwise, the vertex-weights are equal (say 1), and a weighted graph G on n vertices is determined by its $n \times n$ symmetric weight matrix $\mathbf{A}$. Let $G_{\mathbf{A}}$ denote the weighted graph with unit vertex-weights and edge-weights that are entries of $\mathbf{A}$. We will use the following definitions of [3].

Definition 13 *Let w_{ij} $(1 \le i \le j \le n)$ be independent random variables defined on the same probability space, and $w_{ji} = w_{ij}$. $\mathbb{E}(w_{ij}) = 0$ $(\forall i, j)$ and the w_{ij}'s are uniformly bounded, i.e., there is a constant $K > 0$ – that does not depend of n – such that $|w_{ij}| \le K$, $\forall i, j$. The $n \times n$ symmetric real random matrix $\mathbf{W}_n = (w_{ij})_{1 \le i \le n,\, 1 \le j \le n}$ is called a Wigner-noise.*

Definition 14 *The $n \times n$ symmetric real matrix $\mathbf{B}$ is a blown-up matrix, if there is a $q \times q$ symmetric so-called pattern matrix $\mathbf{P}$ with entries $0 < p_{ij} < 1$, and there are positive integers $n_1, \dots, n_q$ with $\sum_{i=1}^q n_i = n$, such that – after rearranging its rows and columns – the matrix $\mathbf{B}$ can be divided into $q \times q$ blocks, where block (i, j) is an $n_i \times n_j$ matrix with entries all equal to p_{ij} $(1 \le i, j \le n)$.*

Let us fix $\mathbf{P}$, blow it up to an $n \times n$ matrix $\mathbf{B}_n$, and consider the noisy matrix $\mathbf{A}_n = \mathbf{B}_n + \mathbf{W}_n$ as $n_1, \dots, n_q \to \infty$ at the same rate. While perturbing $\mathbf{B}_n$ by

$\mathbf{W}_n$, suppose that for the uniform bound of the entries of $\mathbf{W}_n$ the condition

$$K \le \min\{\min_{i,j\in[q]} p_{ij}\,,\, 1-\max_{i,j\in[q]} p_{ij}\} \tag{14}$$

is satisfied. In this way, the entries of $\mathbf{A}_n$ are in the [0,1] interval, and hence, $G_{\mathbf{A}_n} \in \mathcal{G}$. We remark that $G_{\mathbf{W}_n} \notin \mathcal{G}$, but $W_{G_{\mathbf{W}_n}} \in \mathcal{W}$ and the theory of bounded graphons applies to it. In [3] we show that by adding an appropriate Wigner-noise to $\mathbf{B}_n$, we can achieve that $\mathbf{A}_n$ 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 $G_{\mathbf{A}_n}$ is a random simple graph, called generalized random graph, see e.g., [11].

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

Theorem 15 *For any sequence $\mathbf{W}_n$ of Wigner-noises*

$$\lim_{n\to\infty} \|W_{G_{\mathbf{W}_n}}\|_{\square} = 0 \qquad (n\to\infty)$$

almost surely.

The main idea of the proof is that by the definition of the cut-norm of a stepfunction graphon and by using formulas (7.2), (7.3) of [5]:

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

where the entries behind the latter double sum are independent random variables. Hence, the Azuma's inequality is applicable, and the statement follows by the Borel-Cantelli lemma.

Let $\mathbf{A}_n := \mathbf{B}_n + \mathbf{W}_n$ 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 $\mathbf{W}_n$ the condition (14) is satisfied. Under these conditions, Theorem 15 implies that the noisy graph sequence $(G_{\mathbf{A}_n}) \subset \mathcal{G}$ 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 $(G_{\mathbf{A}_n})$ becomes a generalized quasirandom graph sequence with the model graph H, see [10].

Acknowledgements

We are indebted to László Lovász and Katalin Friedl for inspiring discussions.

References

[1] Arora, S., Karger, D., Karpinski, M., Polynomial time approximation schemes for dense instances of NP-hard problems, Proc. 27-th ACM Symp. Theory of Comp. (1995), 284-293.

[2] Bazaraa, M. S., Shetty, C. M., Nonlinear Programming, Theory and Algorithms, Wiley, New York (1979).

[3] Bolla, M., Recognizing Linear Structure in Noisy Matrices, Lin. Alg. Appl 402 (2005), 228-244.

[4] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., Szegedy, B., Graph limits and parameter testing, Proc. 38-th ACM Symp. Theory of Comp. (2006), 261-270.

[5] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K., Convergent Sequences of Dense Graphs I: Subgraph Frequences, Metric Properties and Testing, Advances in Math. 219 (2008), 1801-1851.

[6] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K., Convergent Sequences of Dense Graphs II: Multiway Cuts and Statistical Physics (2007), http://www.cs.elte.hu/∼lovasz/ConvRight.pdf

[7] Frieze, A., Kannan, R., Quick approximation to matrices and applications, Combinatorica 19 (1999), 175-220.

[8] Goldreich, O., Goldwasser, S., Ron, D, Property testing and its connection to learning and approximation, J. ACM 45 (1998), 653-750.

[9] Lovász, L., Szegedy, B., Limits of Dense Graph Sequences, J. Comb. Theory B 96 (2006), 933-957.

[10] Lovász, L., Sós, V. T., Generalized quasirandom graphs, J. Comb. Theory B 98 (2008), 146-163.

[11] McSherry, F., Spectral partitioning of random graphs, Proc. 42-nd FOCS (2001), 529-537.