Limits of dense graph sequences

László Lovász and Balázs Szegedy

Microsoft Research
One Microsoft Way
Redmond, WA 98052

Microsoft Technical Report TR-2004-79
August 2004
Revised April 2006

Contents

Abstract

We show that if a sequence of dense graphs G_n has the property that for every fixed graph F, the density of copies of F in G_n tends to a limit, then there is a natural "limit object", namely a symmetric measurable function $W: [0,1]^2 \to [0,1]$. This limit object determines all the limits of subgraph densities. Conversely, every such function arises as a limit object. We also characterize graph parameters that are obtained as limits of subgraph densities by the "reflection positivity" property.

Along the way we introduce a rather general model of random graphs, which seems to be interesting on its own right.

1 Introduction

Let G_n be a sequence of simple graphs whose number of nodes tends to infinity. For every fixed simple graph F, let $\hom(F,G)$ denote the number homomorphisms of F into G (edge-preserving maps $V(F) \to V(G)$). We normalize this number to get the *homomorphism density*

$$t(F,G) = \frac{\hom(F,G)}{|V(G)|^{|V(F)|}}.$$

This quantity is the probability that a random mapping $V(F) \to V(G)$ is a homomorphism.

Suppose that the graphs G_n become more and more similar in the sense that $t(F,G_n)$ tends to a limit $t(F)$ for every F. Let $\mathcal{T}$ denote the set of graph parameters $t(F)$ arising this way. The goal of this paper is to give characterizations of graph parameters in $\mathcal{T}$. (This question is only interesting if the graphs G_n are dense (i.e., they have $\Omega(|V(G_n)|^2)$ edges); else, the limit is 0 for every simple graph F with at least one edge.)

One way to characterize members of $\mathcal{T}$ is to define an appropriate limit object from which the values $t(F)$ can be read off.

For example, let G_n be a random graph with density 1/2 on n nodes. It can be shown that this converges with probability 1. A natural guess for the limit object would be the random countable graph. This is a very nice object, uniquely determined up to automorphism. However, this graph is too "robust": the limit of random graphs with edge-density 1/3 would be the same, while the homomorphism densities have different limits than in the case of edge-density 1/2.

The main result of this paper is to show that indeed there is a natural "limit object" in the form of a symmetric measurable function $W: [0,1]^2 \to [0,1]$ (we call W *symmetric* if $W(x,y) = W(y,x)$). Conversely, every such function arises as the limit of an appropriate graph sequence. This limit object determines all the limits of subgraph densities: if F is a simple graph with $V(F) = [k] = \{1, \dots, k\}$, then

$$t(f, W) = \int_{[0,1]^n} \prod_{ij \in E(F)} W(x_1, x_j)\, dx_1 \dots dx_n. \qquad (1)$$

The limit object for random graphs of density p is the constant function p.

Another characterization of graph parameters $t(F)$ that are limits of homomorphism densities can be given by describing a complete system of inequalities between the values $t(F)$ for different finite graphs F. One can give such a complete system in terms of the positive semidefiniteness of a certain sequence of matrices which we call *connection matrices* (see section 2.3 for details). This property is related to *reflection positivity* in statistical mechanics. Our results in this direction can be thought of as analogues of the characterization of homomorphism density functions given in [5] in the limiting case.

We can also look at this result as an analogue of the well known characterization of moment sequences in terms of the positive semidefiniteness of the moment matrix. A "2-variable" version of a sequence is a graph parameter, and representation in form of moments of a function (or random variable) can be replaced by the integral representation (1). The positive semidefiniteness of connection matrices is analogous to the positive semidefiniteness of moment matrices.

Every symmetric measurable function $W : [0,1]^2 \to [0,1]$ gives rise to a rather general model of random graphs, which we call W-random. Their main role in this paper is that they provide a graph sequence that converges to W; but they seem to be interesting on their own right.

We show that every random graph model satisfying some rather natural criteria can be obtained as a W-random graph for an appropriate W.

The set $\mathcal{T}$ was introduced by Erdős, Lovász and Spencer [4], where the dimension of its projection to any finite number of coordinates (graphs F) was determined.

Limit objects of graph sequences were constructed by Benjamini and Schramm [1] for sequences of graphs with bounded degree; this was extended by Lyons [10] to sequences of graphs with bounded average degree. (The normalization in that case is different.)

2 Definitions and main results

2.1 Weighted graphs and homomorphisms

A *weighted graph* G is a graph with a weight $\alpha_G(i)$ associated with each node and a weight $\beta_G(i, j)$ associated with each edge ij. (Here we allow that G has loops, but no multiple edges.) In this paper we restrict our attention to positive real weights between 0 and 1. An edge with weight 0 will play the same role as no edge between those nodes, so we could assume that we only consider weighted complete graphs with loops at all nodes (but this is not always convenient). The adjacency matrix of a weighted graph is obtained by replacing the 1's in the adjacency matrix by the weights of the edges. An *unweighted graph* is a weighted graph where all the node-and edgeweights are 1. We set

$$\alpha_G = \sum_{i \in V(G)} \alpha_G(i).$$

Recall that for two unweighted graphs F and G, $\hom(F,G)$ denotes the number of homomorphisms (adjacency preserving maps) from F to G. We extend this notion to the case when G is a weighted graph. To every $\phi:\ V(F) \to V(G)$, we assign the weights

$$\alpha_\phi = \prod_{u \in V(F)} \alpha_G(\phi(u)),$$

and

$$\hom_\phi(F,G) = \prod_{uv \in E(F)} \beta_G(\phi(u), \phi(v)).$$

We then define the *homomorphism function*

$$\hom(F,G) = \sum_{\phi:\ V(F) \to V(G)} \alpha_\phi \hom_\phi(F,G). \tag{2}$$

and the *homomorphism density*

$$t(F,G) = \frac{\hom(F,G)}{\alpha_G^{|V(F)|}}.$$

We can also think of $t(F,G)$ as a homomorphism function after the nodeweights of G are scaled so that their sum is 1.

It will be convenient to extend the notation $\hom_\phi$ as follows. Let $\phi:\ V' \to V(G)$ be a map defined on a subset $V' \subseteq V(F)$. Then define

$$\alpha_\phi = \prod_{u \in V'} \alpha_G(\phi(u)),$$

and

$$\hom_\phi(F,G) = \sum_{\substack{\psi: V(F) \to V(G) \\ \psi \text{ extends } \phi}} \frac{\alpha_\psi}{\alpha_\phi} \hom_\psi(F,G).$$

If $V' = \emptyset$, then $\alpha_\phi = 1$ and $\hom_\phi(F,G) = \hom(F,G)$.

2.2 Convergence of graph sequences

Let (G_n) be a sequence of weighted graphs. We say that this sequence is *convergent*, if the sequence $(t(F,G_n))$ has a limit as $n \to \infty$ for every simple unweighted graph F. (Note that it would be enough to assume this for connected graphs F.) We say that the sequence *converges to a finite weighted graph G* if

$$t(F, G_n) \longrightarrow \hom(F,G)$$

for every simple graph F. A convergent graph sequence may not converge to any finite weighted graph; it will be our goal to construct appropriate limit objects for convergent graph sequences which do not have a finite graph as a limit.

A *graph parameter* is a function defined on simple graphs that is invariant under isomorphism. Every weighted graph G defines graph parameters $\hom(.,G)$, $\mathrm{inj}(.,G)$, $t(.,G)$ and $t_0(.,G)$.

Often we can restrict our attention to graph parameters f satisfying $f(K_1) = 1$, which we call *normalized*. Of the four parameters above, $t(.,G)$ and $t_0(.,G)$ are normalized. We say that a graph parameter f is *multiplicative*, if $f(G_1G_2) = f(G_1)f(G_2)$, where G_1G_2 denotes the disjoint union of two graphs G_1 and G_2. The parameters $\hom(.,G)$ and $t(.,G)$ are multiplicative.

The same graph parameter $\hom(.,G)$, defined by a weighted graph G, arises from infinitely many graphs. Replace a node i of G by two nodes i' and i'', whose weights are chosen so that $\alpha(i') + \alpha(i'') = \alpha(i)$; define the edge weights $\beta(i'j) = \beta(i''j) = \beta(ij)$ for every node j; and keep all the other nodeweights and edgeweights. The resulting weighted graph G' will define the same graph parameter $\hom(.,G') = \hom(.,G)$. Repeating this operation we can create arbitrarily large weighted graphs defining the same graph parameter.

If we want to stay among unweighted graphs, then the above operation cannot be carried out, and the function $\hom(.,G)$ in fact determines G [8]. But for the $t(.,G)$ parameter, the situation is different: if we replace each node of an unweighted graph G by N copies (where copies of two nodes are connected if and only if the originals were), then $t(.,G)$ does not change.

In particular, if we consider a convergent graph sequence, we need not assume that the number of nodes tends to infinity: we could always achieve this without changing the limit.

2.3 Reflection positivity

A *k-labeled graph* ($k \geq 0$) is a finite graph in which k nodes are labeled by $1, 2, \dots k$ (the graph can have any number of unlabeled nodes). For two k-labeled graphs F_1 and F_2, we define the graph F_1F_2 by taking their disjoint union, and then identifying nodes with the same label, and then cancelling the resulting multiplicities of edges. Hence for two 0-labeled graphs, F_1F_2 is just their disjoint union.

Let f be any graph parameter defined on simple graphs. For every integer $k \geq 0$, we define the *connection matrix* $M(k,f)$ as follows. This is infinite matrix, whose rows and columns are indexed by (isomorphism types of) k-labeled graphs. The entry in the intersection of the row corresponding to F_1 and the column corresponding to F_2 is $f(F_1F_2)$. We say that the parameter f is *reflection positive*, if $M(k,f)$ is positive semidefinite for every $k \geq 0$.

In [5], a related matrix was defined. In that paper, the test graphs F may have multiple edges and the target graphs G have arbitrary edgeweights. Let us call a graph parameter defined on graphs which may have multiple edges a *multigraph parameter*. The only difference in the definition of the connection matrix is that edge multiplicities are not cancelled when F_1F_2 is defined. It was shown that $\hom(.,G)$ as a multigraph parameter is reflection positive for every weighted graph G, and the matrix $M(f,k)$ has rank at most $|V(G)|^k$. It was also shown that these two properties characterize which multigraph parameters arise in this form. (In this paper we restrict our attention to simple test graphs F and edge-weights between 0 and 1. See also section 6.2.)

For graph parameters defined on simple graphs, there is a simpler matrix whose positive semidefiniteness could be used to define reflection positivity. Let $M_0(k,f)$ denote the submatrix of $M(k,f)$ formed by those rows and columns that are indexed by k-labeled graphs on k nodes (so that every node is labeled). The equivalence of these definitions (under some further conditions) will follow from our main theorem.

We could combine all these matrices into single matrix $M_0(f)$: the rows and columns of $M_0(f)$ are indexed by all finite graphs whose nodes form a finite subset of $\mathbb{N}$. To get the entry in the intersection of row F_1 and column F_2, we take the union $F_1 \cup F_2$, and evaluate f on this union. Clearly every $M_0(k,f)$ is a minor of $M_0(f)$, and every finite minor of $M_0(f)$ is a minor of $M_0(k,f)$ for every large enough k. So $M_0(f)$ is positive semidefinite if and only if every $M_0(k,f)$ is.

2.4 Homomorphisms, subgraphs, induced subgraphs

Sometimes it is more convenient to work with injective maps. For two unweighted graphs F and G, let $\mathrm{inj}(F,G)$ denote the number of injective homomorphisms from F to G (informally, the number of copies of F in G). We also introduce the *injective homomorphism density*

$$t_0(F,G) = \frac{\mathrm{inj}(F,G)}{(|V(G)|)_{|V(F)|}}$$

(where $(n)_k = n(n-1)\cdot(n-k+1)$).

From a graph-theoretic point of view, it is also important to count induced subgraphs. More precisely, if F and G are two unweighted graphs, then let $\mathrm{ind}(F,G)$ denote the number of embeddings of F into G as an induced subgraph. We define the *induced homomorphism density* by

$$t_1(F,G) = \frac{\mathrm{ind}(F,G)}{(|V(G)|)_{|V(F)|}}.$$

If G is weighted, then we define $\mathrm{inj}(F,G)$ by the same type of sum as for homomorphisms:

$$\mathrm{inj}(F,G) = \sum_{\phi} \alpha_\phi \mathrm{hom}_\phi(F,G),$$

except that the summation is restricted to injective maps. Let

$$t_0(F,G) = \frac{\mathrm{inj}(F,G)}{(\alpha_G)_{|V(F)|}}$$

(where for a sequence $\alpha = (\alpha_1,\dots,\alpha_n)$, $(\alpha)_k$ denotes the k-th elementary symmetric polynomial of the α_i). Note that the normalization was chosen so that $t(F,G) = t_0(F,G) = 1$ if F has no edges.

We also extend the ind function to the case when G is weighted: we define

$$\mathrm{ind}_\phi(F,G) = \prod_{uv\in E(F)} \beta_G(\phi(u),\phi(v)) \prod_{uv\in E(\overline{F})} (1-\beta_G(\phi(u),\phi(v)))$$

(here $\overline{F}$ denotes the complement of the graph F),

$$\mathrm{ind}(F,G) = \sum_{\phi} \alpha_\phi \mathrm{ind}_\phi(F,G),$$

and

$$t_1(F,G) = \frac{\mathrm{ind}(F,G)}{(\alpha_G)_{|V(F)|}}.$$

In the definition of convergence, we could replace $t(F,G)$ by the number $t_0(F,G)$ of embeddings (injective homomorphisms); this is more natural from the graph theoretic point if view. This would not change the notion of convergence or the value of the limit, as the following simple lemma shows:

Lemma 2.1 *For every weighted graph G and unweighted simple graph F, we have*

$$|t(F,G) - t_0(F,G)| < \frac{1}{|V(G)|}\binom{|V(F)|}{2}.$$

We could also replace the hom function by ind function. Indeed,

$$\mathrm{inj}(F,G) = \sum_{F' \supset F} \mathrm{ind}(F',G)$$

(where F' ranges over all supergraphs of F on the same set of nodes), and by inclusion-exclusion,

$$\mathrm{ind}(F,G) = \sum_{F' \supset F} (-1)^{|E(F')\setminus E(F)|} \mathrm{ind}(F',G).$$

Hence it follows that

$$t_1(F,G) = \sum_{F' \supset F} (-1)^{|E(F')\setminus E(F)|} t_0(F',G).$$

It will be convenient to introduce the following operator: if f is any graph parameter, then $f^\dagger$ is the graph parameter defined by

$$f^\dagger(F) = \sum_{F' \supset F} (-1)^{|E(F')\setminus E(F)|} f(F').$$

Thus $t_1 = t_0^\dagger$.

(There is a similar precise relation between the numbers of homomorphisms and injective homomorphisms as well, but we will not have to appeal to it.)

2.5 The limit object

We'll show that every convergent graph sequence has a limit object, which can be viewed as an infinite weighted graph on the points of the unit interval. To be more precise, for every symmetric measurable function $W: [0,1]^2 \to [0,1]$, we can define a graph parameter $t(.,W)$ by

$$t(F,W) = \int_{[0,1]^k} \prod_{ij\in E(F)} W(x_i,x_j)\, dx_1 \ldots dx_k$$

(where F is a simple graph with $V(F) = [k]$).

Let $\phi : V' \to [0,1]$ is a map defined on a subset $V' \subseteq [k]$. Similarly as in the case of homomorphisms into finite graphs, we define t_ϕ as follows. Let, say, $V' = [k']$ $(1 \le k' \le k)$, and $x_i = \phi(i)$ $(i = 1, \dots, k')$. Define

$$t_\phi(F, W) = \int_{[0,1]^{k-k'}} \prod_{ij \in E(F)} W(x_i, x_j)\, dx_{k'+1} \dots dx_k.$$

It is easy to see that for every weighted finite graph H, the simple graph parameter $t(., H)$ is a special case. Indeed, define a function $W_H : [0,1]^2 \to [0,1]$ as follows. Let α_i be the nodeweights of H and β_{ij}, the edgeweights of H. We may assume that $\sum_i \alpha_i = 1$. For $(x, y) \in [0,1]^2$, let a and b determined by

$$\alpha_1 + \dots + \alpha_{a-1} \le x < \alpha_1 + \dots + \alpha_a,$$
$$\alpha_1 + \dots + \alpha_{b-1} \le y < \alpha_1 + \dots + \alpha_b,$$

and let

$$W_H(x, y) = \beta_{ab}.$$

The main result in this paper is the following. Recall that $\mathcal{T}$ denotes the set of graph parameters f that are limits of graph parameters $t(., G)$; i.e., there is a convergent sequence of simple graphs G_n such that

$$f(.) = \lim_{n\to\infty} t(., G_n)$$

for every simple graph F.

Theorem 2.2 *For a simple graph parameter f the following are equivalent:*

(a) $f \in \mathcal{T}$;

(b) *There is a symmetric measurable function $W : [0,1]^2 \to [0,1]$ for which $f = t(., W)$.*

(c) *The parameter f is normalized, multiplicative and reflection positive.*

(d) *The parameter f is normalized, multiplicative and $M_0(f)$ is positive semidefinite.*

(e) *The parameter f is normalized, multiplicative and $f^\dagger \ge 0$.*

Remarks 1. The theorem gives four characterizations of the set $\mathcal{T}$: one analytic, two algebraic and one combinatorial. Characterizations (c), (d) and (e) are closely related (even though a direct proof of the equivalence of (c) and (d) is not easy). Any of these three on the one hand, and (b) on the other, form a "dual" pair in the spirit of NP-coNP: (b) tells us why a graph parameter is in $\mathcal{T}$, while (c) (or (d) or (e)) tells us why it is not.

2. Corollary 2.6 below shows that finite weighted graphs are limits of simple unweighted graphs. This implies that in the definition of $\mathcal{T}$, we could take convergent sequences of weighted graphs instead of unweighted graphs.

3. We could define a more general limit object as a probability space $(\Omega, \mathcal{A}, \pi)$ and a symmetric measurable function on $W: \Omega \times \Omega \to [0,1]$. This would not give rise to any new invariants. However, some limit objects may have a simpler or more natural representation on other σ-algebras (cf. Corollary 2.3 below).

4. One might think that (c) and (d) are equivalent for the more direct reason that $M_0(k, f)$ is positive semidefinite if and only if $M(k, f)$ is. This implication, however, does not hold for a fixed k (even if we assume that f is normalized and multiplicative). For example, $M_0(1, f)$ is positive semidefinite for every normalized multiplicative graph parameter, but $M(1, f)$ is not if $f(F)$ is the number of matchings in F.

5. In the case when $f = \hom(F, G)$ for some finite graph G, $f^\dagger \geq 0$ in condition (e) in Theorem 2.2 expresses that counting induced subgraphs in G we get non-negative values.

As an immediate application of Theorem 2.2, we prove the following fact:

Proposition 2.3 *If $t_1, t_2 \in \mathcal{T}$, then $t_1 t_2 \in \mathcal{T}$.*

This follows from condition (c) in Theorem 2.2, using that positive semidefiniteness is preserved under Schur product. It may be instructive to see how a representation of the product of type (b) can be constructed. Let $t_i = t(., W_i)$, and define W as the 4-variable function $W_1(x_1 y_1) W_2(x_2, y_2)$. We can consider W as a function in two variables x, y, where $x = (x_1, x_2) \in [0,1]^2$ and $y = (y_1, y_2) \in [0,1]^2$. Then W gives rise to graph parameter $t(., W)$, and it is straightforward to check that $t = t_1 t_2$.

2.6 W-random graphs

Given any symmetric measurable function $W: [0,1]^2 \to [0,1]$ and an integer $n > 0$, we can generate a random graph $\mathbf{G}(n, W)$ on node set $[n]$ as follows. We generate n independent numbers $X_1, \ldots, X_n$ from the uniform distribution on $[0,1]$, and then connect nodes i and j with probability $W(X_i, X_j)$.

As a special case, if W is the identically p function, we get "ordinary" random graphs $\mathbf{G}(n, p)$. This sequence is convergent with probability 1, and in fact it converges to the graph $K_1(p)$, the weighted graph with one node and one loop with weight p. The limiting simple graph parameter is given by $t(F) = p^{|E(F)|}$.

More generally, let $W_H: [0,1]^2 \to [0,1]$ be defined by a (finite) weighted graph H with $V(H) = [q]$, whose node weights α_i satisfy $\alpha_1 + \cdots + \alpha_q = 1$. Then $\mathbf{G}(n, W_H)$ can be described as follows. We open q bins $V_1, \ldots, V_q$. Create n nodes, and put each of them independently in bin i with probability α_i. For every pair u, v of nodes, connect them by an edge with probability β_{ij} if $u \in V_i$ and $v \in V_j$. We call $\mathbf{G}(n, W_H)$ a *random graph with model H*.

We show that the homomorphisms densities into $\mathbf{G}(n, F)$ are close to the homomorphism densities into W. Let us fix a simple graph F, let $V(F) = [k]$ and $\mathbf{G} = \mathbf{G}(n, W)$.

The following lemma summarizes some simple properties of W-random graphs.

Lemma 2.4 *For every simple graph F,*

(a) $\mathsf{E}\big(t_0(F,\mathbf{G}(n,W))\big) = t(F,W)$;

(b) $\big|\mathsf{E}\big(t(F,\mathbf{G}(n,W))\big) - t(F,W)\big| < \frac{1}{n}\binom{|V(F)|}{2}$;

(c) $\mathsf{Var}(t(F,\mathbf{G}(n,W))) \le \frac{3}{n}|V(F)|^2$.

This lemma implies, by Chebyshev's inequality, that

$$\Pr\big(|t(F,\mathbf{G}(n,W)) - t(F,W)| > \varepsilon\big) \le 3|V(F)|^2\frac{1}{n\varepsilon^2}.$$

Much stronger concentration results can be proved for $t(F,\mathbf{G})$, using deeper techniques (Azuma's inequality):

Theorem 2.5 *Let F be a graph with k nodes. Then for every $0<\varepsilon<1$,*

$$\Pr\Big(|t_0(F,\mathbf{G}(n,W)) - t(F,W)| > \varepsilon\Big) \le 2\exp\left(-\frac{\varepsilon^2}{2k^2}n\right). \tag{3}$$

and

$$\Pr\Big(|t(F,\mathbf{G}(n,W)) - t(F,W)| > \varepsilon\Big) \le 2\exp\left(-\frac{\varepsilon^2}{18k^2}n\right). \tag{4}$$

From this Theorem it is easy to show:

Corollary 2.6 *The graph sequence $\mathbf{G}(n,W)$ is convergent with probability 1, and its limit is the function W.*

Indeed, the sum of the right hand sides is convergent for every fixed $\varepsilon > 0$, so it follows by the Borell–Cantelli Lemma that $t(F,\mathbf{G}(n,W)) \to t(F,W)$ with probability 1. There is only a countable number of graphs F, so this holds with probability 1 for every F.

This way of generating random graphs is quite general in the following sense. Suppose that for every $n \ge 1$, we are given a distribution on simple graphs on n given nodes, say $[n]$; in other words, we have a random variable $\mathbf{G}_n$ whose values are simple graphs on $[n]$. We call this random variable a *random graph model.* Clearly, every symmetric function $W:\ [0,1]^2 \to [0,1]$ gives rise to a random graph model $\mathbf{G}(n,W)$. The following theorem shows that every model satisfying rather general conditions is of this form:

Theorem 2.7 *A random graph model is of the form $\mathbf{G}(n,W)$ for some symmetric function $W:\ [0,1]^2 \to [0,1]$ if and only if it has the following three properties:*

(i) *the distribution of $\mathbf{G}_n$ is invariant under relabeling nodes;*

(ii) *if we delete node n from $\mathbf{G}_n$, the distribution of the resulting graph is the same as the distribution of $\mathbf{G}_{n-1}$;*

(iii) *for every $1<k<n$, the subgraphs of $\mathbf{G}$ induced by $[k]$ and $\{k+1,\dots,n\}$ are independent as random variables.*

3 Examples

3.1 Quasirandom graphs

Graph sequences converging to $K_1(p)$ are well studied under the name of *quasirandom graphs* (see [3]). More generally, graph sequences converging to a finite weighted graph H are called *quasirandom graphs with model H*. The name is again justified since random graphs $\mathbf{G}(n, W_H)$ with model H converge to H with probability 1. These generalized quasirandom graphs are characterized in [9].

3.2 Half-graphs

Let $H_{n,n}$ denote the bipartite graph on $2n$ nodes $\{1, \dots, n, 1', \dots, n'\}$, where i is connected to j' if and only if $i \le j$. It is easy to see that this sequence is convergent. Indeed, let F be a simple graph with k nodes; we show that the limit of $t(F, H_{n,n})$ exists. We may assume that F is connected. If F is non-bipartite, then $t(F, H_{n,n}) = 0$ for all n, so suppose that F is bipartite; let $V(F) = V_1 \cup V_2$ be its (unique) bipartition. Then every homomorphism of F into H preserves the 2-coloring, and so the homomorphisms split into two classes: those that map V_1 into $\{1, \dots, n\}$ and those that map it into $\{1', \dots, n'\}$. By the symmetry of the half-graphs, these two classes have the same cardinality.

Now F defines a partial order P on $V(F)$, where $u \le v$ if and only if $u = v$ or $u \in V_1$, $v \in V_2$, and $uv \in E$. $(1/2)\hom(F, H_{n,n})$ is just the number of order-preserving maps of P to the chain $\{1, \dots, n\}$, and so

$$\frac{(1/2)\hom(F, H_{n,n})}{n^k} = 2^{k-1} \cdot \frac{\hom(F, H_{n,n})}{(2n)^k} = 2^{k-1} t(F, H_{n,n})$$

is the probability that a random map of $V(F)$ into $\{1, \dots, n\}$ is order-preserving. As $n \to \infty$, the fraction of non-injective maps tends to 0, and hence it is easy to see that $2^{k-1}t(F, H_{n,n})$ tends to a number $2^{k-1}t(F)$, which is the probability that a random ordering of $V(F)$ is compatible with P. In other words, $k!2^{k-1}t(F)$ is the number of linear extensions of P.

However, the half-graphs do not converge to any finite weighted graph. To see this, let S_k denote the star on k nodes, and consider the (infinite) matrix M defined by $M_{k,l} = t(S_{k+l-1})$. If $t(F) = t(F, G_0)$ for some finite weighted graph G_0, then it follows from the characterization of homomorphism functions in [5] that this matrix has rank at most $|V(G_0)|$; on the other hand, it is easy to compute that

$$M_{k,l} = \frac{2^{k+l-1}}{k+l-1},$$

and this matrix (up to row and column scaling, the *Hilbert matrix*) has infinite rank (see e.g [2]).

It is easy to see that in the limit, we are considering order-preserving maps of the poset P into the interval $[0, 1]$; equivalently, the limit object is the characteristic function $W : [0, 1]^2 \to [0, 1]$ of the set $\{(x, y) \in [0, 1]^2 : |x - y| \ge 1/2\}$.

4 Tools

4.1 Distances of functions, graphs and matrices

For any integrable function $U : [0,1]^2 \to \mathbb{R}$, we define its *rectangle norm* by

$$\|U\|_\square = \sup_{\substack{A\subseteq[0,1]\\ B\subseteq[0,1]}} \left| \int_A \int_B U(x,y)\,dx\,dy \right|. \tag{5}$$

It is easy to see that this norm could be defined by the formula

$$\|U\|_\square = \sup_{0\le f,g\le 1} \left| \int_0^1 \int_0^1 U(x,y)f(x)g(y) \right|. \tag{6}$$

The rectangle norm is related to other norms known from analysis. It is not hard to see that

$$\frac{1}{4}\|U\|_{\infty\to 1} \le \|U\|_\square \le \|U\|_{\infty\to 1},$$

where

$$\|U\|_{\infty\to 1} = \sup_{-1\le f,g\le 1} \int_0^1 \int_0^1 U(x,y)f(x)g(y)$$

is the $L_\infty \to L_1$ norm of the operator defined by

$$f \mapsto \int_0^1 U(.,y)f(y)\,dy.$$

It is also easy to see that

$$\|U\|_\square \le \|U\|_1,$$

where

$$\|U\|_1 = \int_0^1 \int_0^1 |U(x,y)|\,dx\,dy$$

is the L_1-norm of U as a function.

The following lemma relates the rectangle norm and homomorphism densities.

Lemma 4.1 *Let $U, W : [0,1]^2 \to [0,1]$ be two symmetric integrable functions. Then for every simple finite graph F,*

$$|t(F,U) - t(F,W)| \le |E(F)| \cdot \|U - W\|_\square.$$

Proof. Let $V(F) = [n]$ and $E(F) = \{e_1, \dots, e_m\}$. Let $e_t = i_t j_t$. Define $E_t = \{e_1, \dots, e_t\}$. Then

$$t(F,U) - t(F,W) = \int_{[0,1]^n} \Big(\prod_{ij\in E(F)} W(x_i,x_j) - \prod_{ij\in E(F)} U(x_i,x_j) \Big)\,dx$$

We can write

$$\prod_{ij\in E(F)} W(x_i,x_j) - \prod_{ij\in E(F)} U(x_i,x_j) = \sum_{t=0}^{m-1} X_t(x_1,\dots,x_n),$$

where

$$X_t(x_1,\dots,x_n) = \Big(\prod_{ij\in E_{t-1}} W(x_i,x_j)\Big)\Big(\prod_{ij\in E(F)\setminus E_t} U(x_i,x_j)\Big)(W(x_{i_t},x_{j_t}) - U(x_{i_t},x_{j_t})).$$

To estimate the integral of a given term, let us integrate first the variables x_{i_t} and x_{j_t}; then by (6),

$$\Big|\int_0^1\int_0^1 X_t(x_1,\dots,x_n)\,dx_{i_t}\,dx_{j_t}\Big| \le \|U-W\|_\square,$$

and so

$$|t(F,U)-t(F,W)| \le \sum_{t=0}^{m-1}\Big|\int_{[0,1]^n} X_t(x_1,\dots,x_n)\,dx\Big| \le m\|U-W\|_\square$$

as claimed. □

Let G and G' be two edge-weighted graphs on the same set V of nodes. We define their *rectangular distance* by

$$d_\square(G,G') = \frac{1}{n^2}\max_{S,T\subseteq V}\Big|\sum_{i\in S,\, j\in T}(\beta_G(i,j)-\beta_{G'}(i,j))\Big|.$$

Clearly, this is a metric, and the distance of any two graphs is a real number between 0 and 1.

Finally, the *rectangular norm* (also called *cut norm* of a matrix $A=(a_{ij})_{i,j=1}^n$ is defined by

$$\|A\|_\square = \max_{S,T}\left|\sum_{i\in S}\sum_{j\in T} a_{ij}\right|,$$

where S and T range over all subsets of $[n]$.

These norms and distances are closely related. If G and G' are two graphs on the same set of nodes, then their distance can be expressed in terms of the associated symmetric functions W_G and $W_{G'}$, and in terms of their (weighted) adjacency matrices A_G and $A_{G'}$ as

$$d_\square(G,G') = \|W_G - W_{G'}\|_\square = \|A_G - A_{G'}\|_\square.$$

Hence by Lemma 4.1,

$$|t(F,G)-t(F,G')| \le |E(F)|\cdot d_\square(G,G') \tag{7}$$

for any simple graph F.

4.2 Szemerédi partitions

A weak form of Szemerédi's lemma (see e.g. [6]; this weak form is all we need) asserts that every graph G can be approximated by a weighted graph with a special structure. Let $\mathcal{P}=(V_1,\dots,V_k)$ be a partition of a finite set V and let Q be a symmetric $k\times k$ matrix with all entries between 0 and 1. We define the graph $K(\mathcal{P},Q)$ as the complete graph on V with a loop at each node, in which the weight of an edge uv ($u\in V_i, v\in V_j$) is Q_{ij}.

Lemma 4.2 (Weak form of Szemerédi's Lemma) *For every $\varepsilon > 0$ there is an integer $k(\varepsilon) > 0$ such that for every simple graph G there exists a partition $\mathcal{P}$ of $V(G)$ into $k \le k(\varepsilon)$ classes $V_1, \dots, V_k$, and a symmetric $k \times k$ matrix Q with all entries between 0 and 1, such that*

$$\big||V_i| - |V_j|\big| \le 1 \qquad (1 \le i, j \le k),$$

and for every set $S \subseteq V(G)$,

$$d_\square(G, K(\mathcal{P}, Q)) \le \varepsilon.$$

We call the partition $\mathcal{P}$ a *weak ε-regular partition of G with density matrix Q.*

The best known bound $k(\varepsilon)$ is of the order $2^{O(1/\varepsilon^2)}$. If the number of nodes of G is less than this, then $\mathcal{P}$ can be chosen to be the partition into singletons, and $K(\mathcal{P}, Q) = G$.

It is not hard to see that (at the cost of increasing $k(\varepsilon)$) we can impose additional conditions on the partition $\mathcal{P}$. We'll need the following condition: the partition $\mathcal{P}$ refines a given partition $\mathcal{P}_0$ of $V(G)$ (the value $k(\varepsilon)$ will also depend on the number of classes in $\mathcal{P}_0$).

It follows from the results of this paper (but it would not be hard to prove it directly), that the above weak form of Szemerédi's Lemma extends to the limit objects in the following form. A symmetric function $U \to [0,1]^2 \to [0,1]$ is called a *symmetric stepfunction with k steps* if there exists a partition $[0,1] = S_1 \cup \dots \cup S_k$ such that U is constant on every set $S_i \times S_j$.

Corollary 4.3 *For every $\varepsilon > 0$ there is an integer $k(\varepsilon) > 0$ such that for every symmetric function $W : [0,1]^2 \to [0,1]$ there exists a symmetric stepfunction $U : [0,1]^2 \to [0,1]$ with k steps such*

$$\|W - U\|_\square \le \varepsilon.$$

Let $\mathcal{P} = \{V_1, \dots, V_k\}$ be a partition of $[0,1]$ into measurable sets. Let $W_\mathcal{P}$ be the stepfunction defined by

$$W_\mathcal{P}(x, y) = \frac{1}{\lambda(V_i)\lambda(V_j)} \int_{V_i \times V_j} W(u, v)\, du\, dv$$

if $x \in V_i$ and $y \in V_j$ (this is not defined if $\lambda(V_i)\lambda(V_j) = 0$, in which case we define $W_\mathcal{P}(x, y) = 0$). It is not hard to see [6] that if U is any stepfunction and $\mathcal{P}$ is the partition of $[0,1]$ into the steps of U, then

$$\|W - W_\mathcal{P}\|_\square \le 2\|W - U\|_\square.$$

So (at the cost of increasing the bound $k(\varepsilon)$), we could state the "weak" Szemerédi Lemma as follows:

Corollary 4.4 *For every $\varepsilon > 0$ there is an integer $k(\varepsilon) > 0$ such that for every symmetric function $W : [0,1]^2 \to [0,1]$ there exists a partition $\mathcal{P}$ of $[0,1]$ into at most $k(\varepsilon)$ measurable sets such that*

$$\|W - W_\mathcal{P}\|_\square \le \varepsilon.$$

5 Proofs

5.1 Proof of Lemma 2.1

We have trivially

$$\hom(F,G) \ge \mathrm{inj}(F,G),$$

and so

$$t(F,G) = \frac{\hom(F,G)}{n^k} \ge \frac{\mathrm{inj}(F,G)}{n^k} = t_0(F,G)\frac{(n)_k}{n^k}.$$

Here

$$\frac{(n)_k}{n^k} = \prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right) \ge 1 - \binom{k}{2}\frac{1}{n},$$

and so

$$t(F,G) \ge t_0(F,G)\left(1-\binom{k}{2}\frac{1}{n}\right) \ge t_0(F,G) - \binom{k}{2}\frac{1}{n}.$$

On the other hand, we have by the beginning of inclusion-exclusion,

$$\mathrm{inj}(F,G) \ge \hom(F,G) - \sum_{F'} \hom(F',G),$$

where the summation ranges over all graphs F' arising from F by identifying two of the nodes. The number of such graphs is $\binom{k}{2}$. Hence

$$\begin{aligned} t_0(F,G) = \frac{\mathrm{inj}(F,G)}{(n)_k} &\ge \frac{\mathrm{inj}(F,G)}{n^k} \ge \frac{\hom(F,G)}{n^k} - \sum_{F'}\frac{\hom(F',G)}{n^k} \\ &= t(F,G) - \frac{1}{n}\sum_{F'} t(F',G) \ge t(F,G) - \binom{k}{2}\frac{1}{n}. \end{aligned}$$

This completes the proof.

5.2 Proof of Lemma 2.4

Consider any injective map $\phi : V(F) \to V(\mathbf{G})$. For a fixed choice of $X_1,\dots,X_n$, the events $\phi(i)\phi(j) \in E(\mathbf{G})$ are independent for different edges ij of F, and so the probability that ϕ is a homomorphism is

$$\prod_{ij\in E(F)} W(X_{\phi(i)}, X_{\phi(j)})$$

Now choosing $X_1,\dots,X_n$ at random, we get that the probability that ϕ is a homomorphism is

$$\mathsf{E}\Big(\prod_{ij\in E(F)} W(X_{\phi(i)}, X_{\phi(j)})\Big) = t(F,W).$$

Summing over all injective maps ϕ, we get (a). By (2.1), we get (b).

Finally, we estimate the variance of $t(F,\mathbf{G})$. Let F_2 denote the disjoint union of 2 copies of F. Then

$$t(F_2,\mathbf{G}) = t(F,\mathbf{G})^2, \qquad \text{and} \qquad t(F_2,W) = t(F,W)^2.$$

Let $R = |V(F)|^2/n$. By Lemma 2.1

$$\mathsf{E}(t(F,G)^2) = \mathsf{E}(t(F_2,G)) \le \mathsf{E}(t_0(F_2,\mathbf{G}) + 2R) = t(F_2,W) + 2R = t(F,W)^2 + 2R,$$

and

$$\mathsf{E}(t(F,G))^2 \ge \mathsf{E}(t_0(F,G) - R/2)^2 \ge \mathsf{E}(t_0(F,\mathbf{G}))^2 - R = t_0(F,W)^2 - R.$$

Hence

$$\mathsf{Var}(t(F,\mathbf{G})) = \mathsf{E}(t(F,G)^2) - \mathsf{E}(t(F,\mathbf{G}))^2 \le 3R = \frac{3}{n}|V(F)|^2.$$

5.3 Proof of Theorem 2.5

The idea of the proof is to form a martingale as follows. In the m-th step ($m = 1,\dots,n$), we generate $X_m \in [0,1]$, and the edges of $\mathbf{G}$ connecting the new node to previously generated nodes. The probability that a random injection of $V(F)$ into $V(\mathbf{G})$ is a homomorphism (conditioning on the part of $\mathbf{G}$ we already generated) is a martingale. We are going to apply Azuma's inequality to this martingale.

To be precise: For every injective map $\phi :\ [k] \to [m]$, let A_ϕ denote the event that ϕ is a homomorphism from F to the random graph $\mathbf{G}$. Let $\mathbf{G}_m$ denote the subgraph of $\mathbf{G}$ induced by nodes $1,\dots,m$. Define

$$B_m = \frac{1}{(n)_k}\sum_\phi \Pr(A_\phi \mid \mathbf{G}_m).$$

Clearly the sequence $(B_0, B_1, \dots)$ is a martingale. Furthermore,

$$\Pr(A_\phi) = t(F,W),$$

and

$$\Pr(A_\phi \mid \mathbf{G}_n) = \begin{cases} 1 & \text{if } \phi \text{ is a homomorphism from } F \text{ to } \mathbf{G}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$B_0 = \sum_\phi \Pr(A_\phi) = t(F,W),$$

and

$$B_n = \frac{1}{(n)_k}\mathrm{inj}(F,\mathbf{G}) = t_0(F,\mathbf{G}).$$

Next we estimate $|B_m - B_{m-1}|$:

$$|B_m - B_{m-1}| = \frac{1}{(n)_k}\left|\sum_\phi (\Pr(A_\phi \mid \mathbf{G}_m) - \Pr(A_\phi \mid \mathbf{G}_{m-1}))\right|$$
$$\le \frac{1}{(n)_k}\sum_\phi \Big|\Pr(A_\phi \mid \mathbf{G}_m) - \Pr(A_\phi \mid \mathbf{G}_{m-1})\Big|.$$

In this sum, every term for which m is not in the range of ϕ is 0, and the other terms are at most 1. The number of terms of the latter kind is $k(n-1)_{k-1}$, and so

$$|B_m - B_{m-1}| \le \frac{k(n-1)_{k-1}}{(n)_k} = \frac{k}{n}.$$

Thus we can invoke Azuma's Inequality:

$$\Pr\Big(B_n - B_0 > \varepsilon\Big) \le \exp\left(-\frac{\varepsilon^2}{2n(k/n)^2}\right) = \exp\left(-\frac{\varepsilon^2}{2k^2}n\right),$$

and similarly

$$\Pr\Big(B_n - B_0 > \varepsilon(n)_k\Big) \le \exp\left(-\frac{\varepsilon^2}{2k^2}n\right).$$

Hence

$$\Pr\Big(|B_n - (n)_k t(F,W)| > \varepsilon(n)_k\Big) \le 2\exp\left(-\frac{\varepsilon^2}{2k^2}n\right).$$

This proves (3).

To get (4), we use Lemma 2.1. We may assume that $n > k^2/\varepsilon$ (else the inequality is trivial). Then

$$|t_0(F,\mathbf{G}) - t(F,\mathbf{G})| \le \frac{\varepsilon}{3},$$

and similarly

$$|t_0(F,\mathbf{G}) - t(F,H)| \le \frac{\varepsilon}{3},$$

so (4) follows by applying (3) with $\varepsilon/3$ in place of ε.

5.4 Proof of Theorem 2.2: (a)⇒(b)

Let (G_n) be a convergent graph sequence and

$$f(F) = \lim_{n\to\infty} t(F,G_n)$$

for every n. We want to construct a function $W:\ [0,1]^2 \to [0,1]$ such that $f = t(.,W)$.

We start with constructing a subsequence of (G_n) whose members have well-behaved Szemerédi partitions.

Lemma 5.1 *Every graph sequence* $(G_n:\ n = 1, 2, \dots)$ *has a subsequence* $(G'_m:\ m = 1, 2, \dots)$ *for which there exists a sequence* $(k_m:\ m = 1, 2, \dots)$ *of integers and a sequence* $(Q_m:\ m = 1, 2, \dots)$ *of matrices with the following properties.*

(i) Q_m *is a* $k_m \times k_m$ *symmetric matrix, all entries of which are between* 0 *and* 1*.*

(ii) *If* $i < j$*, then* $k_i \mid k_j$*, and the matrix* Q_i *is obtained from the matrix* Q_j *by partitioning its rows and columns into* k_i *consecutive blocks of size* k_j/k_i*, and replacing each block by a single entry, which is the average of the entries in the block.*

(iii) *For all* $j < m$*,* G'_m *has a weakly* $(1/m)$*-regular partition* $\mathcal{P}_{m,j}$ *with density matrix* $Q_{m,j}$ *such that*

$$\|Q_{m,j} - Q_j\|_\square < 1/j, \tag{8}$$

and for $1 \le i < j \le m$*,* $\mathcal{P}_{m,j}$ *is a refinement of* $\mathcal{P}_{m,i}$*.*

Proof. For every integer $m \ge 1$, we construct a subsequence (G_n^m) so that all graphs in the subsequence have a weakly $(1/m)$-regular partition $\mathcal{P}_{n,m}$ into the same number k_m of classes and with almost the same density matrix.

The first sequence (G_n^1) is selected from (G_n) so that the edge density of G_n^1 converges to a fix constant c between 0 and 1 if n tends to infinity. Furthermore, for every graph G_n^1 let $\mathcal{P}_{n,1} = \{V(G_n^1)\}$ be the 1-block partition and let $Q_{n,1}$ be the 1 by 1 matrix containing the edge density of G_n^1. We set $k_1 = 1$.

Suppose that for some integer $m > 0$, we have constructed the sequence (G_n^m). For every graph G_n^m, consider a weakly $1/(m+1)$-regular partition $\mathcal{P}_{n,m+1} = \{V_1, \dots, V_{K_n}\}$ of G_n^m with density matrix $Q_{n,m+1}$. We may choose this partition so that it refines the previous partition $\mathcal{P}_{n,m}$, and each class of $\mathcal{P}_{n,m}$ is split into the same number of classes $r_{n,m}$; the number K_n of classes remains bounded, and so the numbers $r_{n,m}$ also remain bounded (m is fixed, $n \to \infty$). So we can thin the sequence so that all remaining graphs have the same $r_m = r_{n,m}$. We set $k_{m+1} = k_m r_m$. Furthermore, we can select a subsequence so that the density matrices $Q_{n,m+1}$ converge to a fixed matrix Q_{m+1} if n tends to infinity. Finally we drop all the elements G_i^{m+1} from the remaining sequence for which $\|Q_{i,m+1} - Q_{m+1}\|_\square > 1/(2m+2)$. By renumbering the indices, we obtain the subsequence G_n^{m+1}.

Let $G'_m = G_1^m$. For every $1 \le j < m$, the graph G'_m has a tower of partitions $\mathcal{P}_{m,j}$ $(j = 1, \dots, m)$ so that $\mathcal{P}_{m,j}$ has k_j almost equal classes, and has a density matrix $Q_{m,j}$ such that $\lim_{m\to\infty} Q_{m,j} = Q_j$ and $\|Q_{m,j} - Q_{m',j}\|_\square < 1/j$ for all $m, m' > j$.

For a fixed graph G'_m, the partitions $\mathcal{P}_{m,1}, \mathcal{P}_{m,2}, \dots, \mathcal{P}_{m,m}$ are successively refinements of each other. We may assume that the classes are labeled so that the i-th class of $\mathcal{P}_{m,j+1}$ is the union of consecutive classes $(i-1)r_m + 1, \dots, ir_m$. Let $\widehat{Q}_{m,j}$ $(1 \le j \le m)$ be the $k_j \times k_j$ matrix obtained from Q_m by partitioning its rows and columns into k_j consecutive blocks of size k_m/k_j, and replacing each block by a single entry, which is the average of the corresponding entries of Q_m. Using that $\lim_{n\to\infty} |V(G_n)| = \infty$ and that all the sets in $\mathcal{P}_{m,j}$ have almost the same size one gets that $\widehat{Q}_{m,j} = Q_j$.

Thus we constructed a sequence $(G'_1, G'_2, \dots)$ of graphs, an increasing sequence $(k_1, k_2, \dots)$ of positive integers, and a sequence $(Q_1, Q_2 \dots)$ of matrices. We claim that these sequences satisfy the properties required in the Lemma. (i), (ii) and the second assertion of (iii) are trivial by construction. The first assertion in (iii) follows on noticing that Q_j is the limit of matrices $Q_{m,j}$, and $\|Q_{m,j}, Q_{m',j}\|_\square < 1/j$ for all $j \le m \le m'$. $\square$

Lemma 5.2 *Let* (k_m) *be a sequence of positive integers and* (Q_m), *a sequence of matrices satisfying* (i) *and* (ii) *in Lemma 5.1. Then there exists a symmetric measurable function* $W:\ [0,1]^2 \to [0,1]$ *such that*

(a) $W_{Q_m} \to W$ $(m \to \infty)$ *almost everywhere;*

(b) *for all* m *and* $1 \le i, j \le k_m$

$$(Q_M)_{ij} = k_m^2 \int_{(i-1)/k_m}^{i/k_m} \int_{(j-1)/k_m}^{j/k_m} W(x,y)\,dx\,dy.$$

Proof. Define a map $\phi_m :\ [0,1) \to [k_m]$ by mapping the interval $[(i-1)/k_m, i/k_m)$ to i.

Claim 5.3 *Let* X *and* Y *be two uniformly distributed random elements of* $[0,1]$. *Then the sequence* $Z_m = (Q_m)_{\phi_m(X),\phi_m(Y)}$, $i = 1, 2, \ldots$ *is a martingale.*

We want to show that

$$\mathsf{E}(Z_{m+1} \mid Z_1, \ldots, Z_m) = Z_m. \tag{9}$$

In fact, we show that

$$\mathsf{E}(Z_{m+1} \mid \phi_1(X), \phi_1(Y), \ldots, \phi_m(X), \phi_m(Y)) = Z_m. \tag{10}$$

Since $\phi_m(X)$ and $\phi_m(Y)$ determine $\phi_i(X)$ and $\phi_i(Y)$ for $i < m$, it suffices to show that

$$\mathsf{E}(Z_{m+1} \mid \phi_m(X) = a, \phi_m(Y) = b) = (Q_m)_{a,b}. \tag{11}$$

The condition $\phi_m(X) = a$ and $\phi_m(Y) = b$ force X to be uniform in the interval $[(a-1)/k_m, a_{k_m}]$, and so $\phi_{m+1}(X)$ is a uniform integer in the interval $[(k_{m+1}/k_m)(a-1), (k_{m+1}/k_m)a]$. Similarly, $\phi_{m+1}(Y)$ is a uniform integer in the interval $[(k_{m+1}/k_m)(b-1), (k_{m+1}/k_m)b]$. So $(Q_{m+1})_{\phi_{m+1}(X),\phi_{m+1}(Y)}$ is a uniformly distributed entry of the submatrix formed by these rows and columns. By condition (ii) in Lemma 5.1, the average of these matrix entries is exactly $(Q_m)_{a,b}$. This proves the claim.

Since Z_m is also bounded, we can invoke the Martingale Convergence Theorem, and conclude that $\lim_{m\to\infty} Z_m$ exists with probability 1. This means that

$$W(x,y) = \lim_{m\to\infty} (Q_m)_{\phi_m(x),\phi_m(y)}$$

exists for almost all pairs (x,y), $0 \le x, y \le 1$. Let us define $W(x,y) = 0$ whenever the limit does not exist.

It is trivial that W is symmetric, $0 \le W \le 1$, and W satisfies condition (a) in the Lemma. Furthermore,

$$\int_{(i-1)/k_m}^{i/k_m} \int_{(j-1)/k_m}^{j/k_m} W(x,y)\,dx\,dy = \int_{(i-1)/k_m}^{i/k_m} \int_{(j-1)/k_m}^{j/k_m} \lim_{n\to\infty} (Q_n)_{\phi_n(x),\phi_n(y)}\,dx\,dy$$
$$= \lim_{n\to\infty} \int_{(i-1)/k_m}^{i/k_m} \int_{(j-1)/k_m}^{j/k_m} (Q_n)_{\phi_n(x),\phi_n(y)}\,dx\,dy$$
$$= \frac{1}{k_m^2}\Big(\lim_{n\to\infty} \frac{k_m^2}{k_n^2} \sum_{a=(i-1)(k_n/k_m)+1}^{i(k_n/k_m)} \sum_{b=(j-1)(k_n/k_m)+1}^{j(k_n/k_m)} (Q_n)_{a,b}\Big)$$
$$= \frac{1}{k_m^2}(Q_m)_{i,j}$$

(the last equation follows from assumption (ii)). This proves (b). □

Now it easy to conclude the proof of the necessity of the condition in Theorem 2.2. Let us apply Lemma 5.1 to the given convergent graph sequence. The sequence $(G_1', G_2', \dots)$ of graphs it gives is a subsequence of the original sequence, so it is convergent and defines the same limit parameter f. The lemma also gives a sequence of integers and a sequence of matrices satisfying (i) and (ii). We can use Lemma 5.3 to construct an integrable function $W: [0,1]^2 \to [0,1]$ with properties (a), (b) and (c).

It remains to show that $f = t(.,W)$. For $1 \le j \le m$, let $G_{m,j}^* = G(\mathcal{P}_{m,j}, Q_{m,j})$ and $G_{m,j}^{**} = G(\mathcal{P}_{m,j}, Q_j)$. Then

$$d(G_m', G_{m,j}^*) \le \frac{1}{j} \tag{12}$$

(since $\mathcal{P}_{m,j}$ is a weakly $(1/j)$-regular partition of G_m'), and

$$d(G_{m,j}^*, G_{m,j}^{**}) \le \frac{1}{j} \tag{13}$$

by (8).

Let $W_{m,j} = W_{G_{m,j}^{**}}$ and $W_j = W_{Q_j}$. Clearly

$$t(F, W_{m,j}) = t(F, G_{m,j}^{**}).$$

Furthermore,

$$W_{m,j} \to W_j \qquad (m \to \infty) \tag{14}$$

almost everywhere. Indeed, the functions $W_{m,j}$ and W_j differ only if the classes in $\mathcal{P}_{m,j}$ are not all equal; but even in this case, if $W_{m,j}(x,y) \ne W_j(x,y)$ then either x or y must be closer to one of the numbers a/k_j than $1/|V(G_m')|$. Finally, we have

$$W_j \to W \qquad (j \to \infty) \tag{15}$$

almost everywhere.

Now let $\varepsilon > 0$, and choose a positive integer m_0 so that for $m > m_0$, we have

$$|t(F, G'_m) - f(F)| < \frac{\varepsilon}{4}.$$

By (15), we can choose a positive integer j so that

$$|t(F, W) - t(F, W_j)| < \frac{\varepsilon}{4}.$$

We may also assume that $j > 8/\varepsilon$ and $j > m_0$. By (14), we can choose an $m > j$ so that

$$|t(F, W_{m,j}) - t(F, W_j)| \leq \frac{\varepsilon}{4}.$$

By Lemma 7 and inequalities (12) and (13), we have

$$|t(F, G'_m) - t(F, G^{**}_{m,j})| \leq \frac{\varepsilon}{4}.$$

Combining these inequalities, we get that

$$|f(F) - t(F, W)| \leq \varepsilon,$$

which completes the proof of (a)$\Rightarrow$ (b).

5.5 Proof of Theorem 2.2: (b)⇒(c)

Let $f = t(., W)$. It is obvious that f is normalized and multiplicative.

To prove that f is reflection positive, consider any finite set $F_1, \dots, F_m$ of k-labeled graphs, and real numbers $y_1, \dots, y_m$. We want to prove that

$$\sum_{p,q=1}^{m} f(F_p F_q) y_p y_q \geq 0.$$

For every k-labeled graph F with node set $[n]$, let F' denote the subgraph of F induced by the labeled nodes, and F'' denote the graph obtained from F by deleting the edges spanned by the labeled nodes. Define

$$\tau(F, x_1, \dots, x_k) = \int_{[0,1]^{n-k}} \prod_{ij \in E(F'')} W(x_i, x_j)\, dx_{k+1} \dots dx_n,$$

and for every graph F with $V(F) = [k]$,

$$W(F, x_1, \dots, x_k) = \prod_{ij \in E(F)} W(x_i, x_j).$$

Then

$$\sum_{p,q=1}^{m} y_p y_q f(F_p F_q) = \int_{[0,1]^k} \sum_{p,q=1}^{m} y_p y_q \tau(F_p, x_1, \dots, x_k) \tau(F_q, x_1, \dots, x_k)$$
$$W(F'_p \cup F'_q, x_1, \dots, x_k)\, dx_1 \dots dx_k.$$

We prove that the integrand is nonnegative for every $x_1, \dots, x_k$:

$$\sum_{p,q=1}^{m} \overline{y}_p \overline{y}_q W(F'_p \cup F'_q) \geq 0,$$

where $\overline{y}_p = y_p \tau(F_p, x_1, \dots, x_k)$, and the x_i are suppressed for clarity). Let

$$\hat{W}(F) = \prod_{ij \in E(F)} W(x_i, x_j) \prod_{ij \in E(\overline{F})} (1 - W(x_i, x_j)),$$

then clearly $W(\overline{F}) \geq 0$, and for every $F \in \mathcal{F}_k$,

$$W(F) = \sum_{H \supset F} \hat{W}(H)$$

(where the summation extends over all $H \in \mathcal{F}_k$ containing F as a subgraph). Thus

$$\sum_{p,q=1}^{m} \overline{y}_p \overline{y}_q W(F'_p \cup F'_q) = \sum_{p,q=1}^{m} \overline{y}_p \overline{y}_q \sum_{H \supset F'_p \cup F'_q} \hat{W}(H)$$
$$= \sum_{H \in \mathcal{F}_k} \hat{W}(H) \sum_{p,q: F_p, F_q \subseteq H} \overline{y}_p \overline{y}_q = \sum_{H \in \mathcal{F}_k} \hat{W}(H) \left(\sum_{p: F_p \subseteq H} \overline{y}_p \right)^2 \geq 0.$$

5.6 Proof of Theorem 2.2: (c)⇒(d)

This is trivial, since $M_0(k, f)$ is a symmetric submatrix of $M(k, f)$.

5.7 Proof of Theorem 2.2: (d)⇒(e)

The proof of this implication uses the fact that the entry in the (F_1, F_2)-position in $M_0(k, f)$ depends on the union of F_1 and F_2 only. The Lindström–Wilf Formula gives a nice diagonalization of such matrices as follows. Let $\mathcal{F}_k$ denote the set of all graphs with nodes $[k]$. Let Z denote the $\mathcal{F}_k \times \mathcal{F}_k$ matrix defined by

$$Z_{F_1,F_2} = \begin{cases} 1 & \text{if } F_1 \subseteq F_2, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Let D be the diagonal matrix

$$D_{F_1,F_2} = \begin{cases} f^\dagger(F_1) & \text{if } F_1 = F_2, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Then

$$B = Z^\mathsf{T} D Z. \tag{18}$$

This implies that $M_0(k, f)$ is positive semidefinite if and only if $f^\dagger \geq 0$ for all graphs with k nodes.

5.8 Proof of Theorem 2.2: (e)⇒(a)

Let f be a normalized and multiplicative graph parameter such that $f^\dagger \geq 0$. Fix any $k \geq 1$. As a first step we construct a random variable $\mathbf{G}_k$, whose values are graphs with k labeled nodes: Let $\mathbf{G}_k = F$ with probability $f^\dagger(F)$. Since $f^\dagger \geq 0$ by hypothesis and

$$\sum_F f^\dagger(F) = f(\overline{K_k}) = 1$$

(where the summation extends over all graphs F with $V(F) = [k]$), this is well defined. It is also clear that this distribution does not depend on the labeling of the nodes.

Next we show that for every graph F with k nodes,

$$f(F) = \mathsf{E}(t_0(F, \mathbf{G}_k)). \tag{19}$$

Indeed, we have

$$\begin{aligned}\mathsf{E}(t_0(F, \mathbf{G}_k)) &= \sum_{F' \supseteq F} \mathsf{E}(t_1(F', \mathbf{G}_k)) \\ &= \sum_{F' \supseteq F} \Pr(F' = \mathbf{G}_k) = \sum_{F' \supseteq F} f^\dagger(F') = f(F).\end{aligned}$$

We claim that for every graph with $k \leq n$ nodes,

$$f(F) = \mathsf{E}(t_0(F, \mathbf{G}_n)). \tag{20}$$

Indeed, add $n - k$ isolated nodes to F to get a graph F' with n nodes. Then $f(F') = f(F)$ by multiplicativity and $f(K_1) = 1$, while $t_0(F', G) = t_0(F, G)$ for every graph G. Thus

$$f(F) = f(F') = \mathsf{E}(t_0(F', \mathbf{G}_n)) = \mathsf{E}(t_0(F, \mathbf{G}_n)).$$

We need a bound on the variance of $t_0(F', \mathbf{G}_n)$: By (20),

$$\mathsf{Var}(t_0(F, \mathbf{G}_n)) = \mathsf{E}(t_0(F, \mathbf{G}_n)^2) - (\mathsf{E}(t_0(F, \mathbf{G}_n))^2.$$

Here

$$(\mathsf{E}(t_0(F, \mathbf{G}_n))^2 = f(F)^2 = f(FF) = \mathsf{E}(t_0(FF, \mathbf{G}_n)$$

(by multiplicativity), so

$$\mathsf{Var}(t_0(F, \mathbf{G}_n)) = \mathsf{E}\Big(t_0(F, \mathbf{G}_n)^2 - t_0(FF, \mathbf{G}_n)\Big).$$

Now for any graph G,

$$t(F, G)^2 = t(FF, G),$$

and so

$$\begin{aligned}|t_0(F, \mathbf{G}_n)^2 - t_0(FF, \mathbf{G}_n)| &\leq |t(F, \mathbf{G}_n)^2 - t_0(F, \mathbf{G}_n)^2| + |t(FF, \mathbf{G}_n) - t_0(FF, \mathbf{G}_n)| \\ &\leq 2|t(F, \mathbf{G}_n) - t_0(F, \mathbf{G}_n)| + |t(FF, \mathbf{G}_n) - t_0(FF, \mathbf{G}_n)| \\ &\leq 2\binom{k}{2}\frac{1}{n} + \binom{2k}{2}\frac{1}{n} < \frac{3k^2}{n}.\end{aligned}$$

Thus for every graph F with $k \le n$ nodes,

$$\mathsf{Var}(t_0(F, \mathbf{G}_n)) \le \frac{3k^2}{n}. \tag{21}$$

By Chebyshev's Inequality,

$$\Pr(|t_0(F, \mathbf{G}_n) - f(F)| > \varepsilon) < \frac{3k^2}{\varepsilon^2 n}.$$

It follows by the Borell-Cantelli Lemma that if we take (say) the graph sequence $(\mathbf{G}_{n^2} : n = 1, 2, \dots)$, then with probability 1,

$$t_0(F, \mathbf{G}_{n^2}) \to f(F) \quad (n \to \infty).$$

Since there are only a countable number of graphs F, this convergence holds with probability 1 for every F. So we see that

$$f(.) = \lim_{n\to\infty} t(., \mathbf{G}_{n^2})$$

for almost all choices of the sequence $(\mathbf{G}_{n^2})$. This completes the proof of Theorem 2.2.

Remarks. 1. There are alternatives for certain parts of the proof. Instead of verifying (b)⇒(c) directly, we could argue that (b)⇒(a) (which follows e.g. from Corollary 2.6), and then that (a)⇒(c) (which follows from the characterization of homomorphism functions in [5]).

2. Equation (20), satisfied by the random graph $\mathbf{G}_n$, is the same as equation (a) in Lemma 2.4, satisfied by the random graph $\mathbf{G}(n, W)$. It is not hard to see that this equation uniquely determines the distribution on n-node graphs, and hence $\mathbf{G}_n$ and $\mathbf{G}(n, W)$ have the same distribution.

3. The construction of the limit object W in the proof shows that every convergent graph sequence (G_n) has a subsequence (G'_n) such that (with an appropriate labeling of the nodes)

$$\|W_{G'_n} - W\|_\square \to 0.$$

5.9 Proof of Theorem 2.7

It is trivial that $\mathbf{G}(n, W)$ satisfies (a), (b) and (c). Conversely, suppose that $\mathbf{G}_n$ is a graph model with properties (i), (ii) and (iii). Define a graph parameter f by

$$f(F) = \Pr(F \subseteq \mathbf{G}_k),$$

where $V(F) = [k]$. By condition (i), f is independent of the labeling of the nodes, so it is indeed a graph parameter.

We claim that $f \in \mathcal{T}$, by verifying (e) in Theorem 2.2. It is trivial that f is normalized. Multiplicativity is an easy consequence of (iii) and (ii): Let $V(F_1) = [k]$ and $V(F_2) = \{k + 1, \dots, k + l\}$, then

$$\begin{aligned} f(F_1 \cup F_2) &= \Pr(F_1 \cup F_2 \subseteq \mathbf{G}_{k+l}) = \Pr(F_1 \subseteq \mathbf{G}_{k+l}) \Pr(F_2 \subseteq \mathbf{G}_{k+l}) \\ &= \Pr(F_1 \subseteq \mathbf{G}_k) \Pr(F_2 \subseteq \mathbf{G}_l) = f(F_1) f(F_2). \end{aligned}$$

Next we show that

$$f(F) = \mathsf{E}(t_0(F, \mathbf{G}_n)) \tag{22}$$

for every $n \geq k$. For $n = k$, this is obvious. Let $n > k$, and let us add $n - k$ isolated nodes to F to get graph F'. Then, using condition (ii),

$$f(F) = \Pr(F \subseteq \mathbf{G}_k) = \Pr(F' \subseteq \mathbf{G}_n) = \mathsf{E}(t_0(F', \mathbf{G}_n)) = \mathsf{E}(t_0(F, \mathbf{G}_n))$$

as claimed.

Applying (22) we obtain

$$f^\dagger(F) = \mathsf{E}(t_0^\dagger(F, \mathbf{G}_n)) = \mathsf{E}(t_1(F, \mathbf{G})) \geq 0.$$

This proves that $f \in \mathcal{T}$.

By Theorem 2.2, there exists a symmetric measurable function $W: [0,1]^2 \to [0,1]$ such that

$$f(F) = t(F, W)$$

for every graph F. By Lemma 2.4(a) and equation (22), we have

$$\mathsf{E}(t_0(F, \mathbf{G}_k)) = \mathsf{E}(t_0(F, \mathbf{G}(k, W)).$$

Applying the "†" operator, this implies that

$$\mathsf{E}(t_1(F, \mathbf{G}_k)) = \mathsf{E}(t_1(F, \mathbf{G}(k, W)),$$

and hence

$$\Pr(F = \mathbf{G}_k) = \Pr(F = \mathbf{G}(k, W)),$$

which proves that $\mathbf{G}_k$ and $\mathbf{G}(k, W)$ have the same distribution.

6 Concluding remarks

We mention some results and problems related to our work. Details (exact formulations, results and conjectures) will be discussed elsewhere.

6.1 Uniqueness

The limit function of a graph sequence is "essentially unique". In other words, if two functions $U, W: [0,1]^2 \to [0,1]$ are such that

$$t(F, U) = t(F, W)$$

for every simple graph F, then U and W are "essentially the same". Unfortunately, it is nontrivial to characterize what "essentially the same" means; for example, U could be obtained from W by applying the same measure-reserving permutation in both coordinates.

6.2 Weighted graphs and multiple edges

It seems to be quite straightforward to extend our results to the case when the graphs G_n can have multiple edges, or more generally, edge-weights (not restricted to $[0,1]$): We simply have to drop the bounds on the limit function W. However, several technical issues arise concerning integrability conditions and the applicability of the Martingale Theorem.

Allowing multiple edges in the "sample graphs" F leads to a more complicated question. Assume that we consider a sequence of simple graphs G_n. If we define F' as the underlying simple graph of a multigraph F, then

$$\hom(F', G_n) = \hom(F, G_n) \qquad \text{and} \qquad t(F', G_n) = t(F, G_n),$$

but

$$t(F', W) < t(F, W)$$

if W is a function that is strictly between 0 and 1. Since, as we have remarked, the limit function W is essentially unique, the formula for $t(F, W)$ does not define the limit of $t(F, G_n)$ correctly.

6.3 Extremal graph theory

There are many results in graph theory, especially in extremal graph theory, that can be formulated as inequalities between the numbers $t(F, G)$ for a fixed G and various graphs F. For example, Goodman's theorem relating the number of edges to the number of triangles can be stated as

$$t(K_3, G) \geq t(K_2, G)(2t(K_2, G) - 1).$$

This inequality is equivalent to saying that for every graph parameter $t \in \mathcal{T}$,

$$t(K_3) \geq t(K_2)(2t(K_2) - 1). \tag{23}$$

By Theorem 2.2, such an inequality must be a consequence of reflection positivity, multiplicativity, and the trivial condition that t is normalized. In fact, (23) can be easily derived from these conditions (this is left to the reader as an exercise). Many other results in extremal graph theory follow in a similar way.

Acknowledgement

We are grateful to Jeong-Han Kim for his kind advice.

References

[1] I. Benjamini and O. Schramm: Recurrence of Distributional Limits of Finite Planar Graphs, *Electronic J. Probab.* **6** (2001), paper no. 23, 1–13.

[2] M.-D. Choi: Tricks or Treats with the Hilbert Matrix, *Amer. Math. Monthly* **90** (1983), 301–312.

[3] F. Chung, R.L. Graham and R.M. Wilson: Quasi-random graphs, *Combinatorica* **9** (1989), 345–362.

[4] P. Erdős, L. Lovász, J. Spencer: Strong independence of graphcopy functions, in: *Graph Theory and Related Topics*, Academic Press (1979), 165–172.

[5] M. Freedman, L. Lovász, A. Schrijver: Reflection positivity, rank connectivity, and homomorphism of graphs (MSR Tech Report # MSR-TR-2004-41)

`ftp://ftp.research.microsoft.com/pub/tr/TR-2004-41.pdf`

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

[7] W.T. Gowers: Lower bounds of tower type for Szemerédi's Uniformity Lemma, *Geom. Func. Anal.* **7** (1997), 322–337.

[8] L. Lovász: Operations with structures, *Acta Math. Hung.* **18** (1967), 321–328.

[9] L. Lovász and V.T. Sós: Generalized quasirandom graphs, manuscript.

[10] R. Lyons: Asymptotic enumeration of spanning trees *Combin. Probab. Comput.* **14** (2005), 491–522.

[11] A. Thomason: Pseudorandom graphs, in: *Random graphs '85* North-Holland Math. Stud. **144**, North-Holland, Amsterdam, 1987, 307–331.