Edit distance and its computation

József Balogh* Ryan Martin†

Submitted: Sep 26, 2007; Accepted: Jan 17, 2008; Published: Jan 28, 2008
Mathematics Subject Classification: 05C35, 05C80

Abstract

In this paper, we provide a method for determining the asymptotic value of the maximum edit distance from a given hereditary property. This method permits the edit distance to be computed without using Szemerédi's Regularity Lemma directly.

Using this new method, we are able to compute the edit distance from hereditary properties for which it was previously unknown. For some graphs H, the edit distance from $\mathrm{Forb}(H)$ is computed, where $\mathrm{Forb}(H)$ is the class of graphs which contain no induced copy of graph H.

Those graphs for which we determine the edit distance asymptotically are $H = K_a + E_b$, an a-clique with b isolated vertices, and $H = K_{3,3}$, a complete bipartite graph. We also provide a graph, the first such construction, for which the edit distance cannot be determined just by considering partitions of the vertex set into cliques and cocliques.

In the process, we develop weighted generalizations of Turán's theorem, which may be of independent interest.

1 Introduction

Throughout this paper, we use standard terminology in the theory of graphs. See, for example, [6]. A subgraph devoid of edges, usually called an independent set, is referred to in this paper as a **coclique**, so that it parallels the notion of a **clique**.

1.1 Background

The edit distance of graphs was defined in [4] as follows:

*Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, jobal@math.uiuc.edu. This author's research supported in part by NSF grant DMS-0600303, UIUC Campus Research Board #07048 and by OTKA grants T034475 and T049398.

†Department of Mathematics, Iowa State University, Ames, IA 50011, rymartin@iastate.edu. This author's research supported in part by NSA grant H98230-05-1-0257.

Definition 1 *Let $\mathcal{P}$ denote a class of graphs. If G is a fixed graph, then **the edit distance from G to $\mathcal{P}$** is*

$$\mathrm{Dist}(G,\mathcal{P}) = \min\left\{|E(F)\triangle E(G)| : F\in\mathcal{P}, V(F)=V(G)\right\}$$

*and **the edit distance from n-vertex graphs to $\mathcal{P}$** is*

$$\mathrm{Dist}(n,\mathcal{P}) = \max\left\{\mathrm{Dist}(G,\mathcal{P}) : |V(G)|=n\right\}.$$

It is natural to consider hereditary properties of graphs. A **hereditary property** is one that is closed under the deletion of vertices. In fact, edge-modification for such properties is an important question in computer science, as described in Alon and Stav [1] and biology, as shown in [4].

Clearly, $\mathrm{Forb}(H)$ is a hereditary property for any graph H. In fact, every hereditary property, $\mathcal{H}$, can be expressed as $\bigcap_{H\in\mathcal{F}(\mathcal{H})}\mathrm{Forb}(H)$, where the intersection is over the family $\mathcal{F}(\mathcal{H})$, which consists of all graphs H which are the minimal elements of $\overline{\mathcal{H}}$.

In [1], Alon and Stav prove that, for every hereditary property $\mathcal{H}$, there exists a $p^*=p^*(\mathcal{H})$ such that, with high probability, $\mathrm{Dist}(n,\mathcal{H}) = \mathrm{Dist}\left(G(n,p^*),\mathcal{H}\right)+o(n^2)$, where $G(n,p)$ denotes the usual Erdős-Rényi random graph. This fact can be used to prove the existence of

$$d^*(\mathcal{H}) \stackrel{\mathrm{def}}{=} \lim_{n\to\infty}\mathrm{Dist}(n,\mathcal{H})/\binom{n}{2}.$$

1.2 Previous results

The previously-known general bounds for $\mathrm{Dist}(n,\mathrm{Forb}(H))$ are expressed in terms of the so-called binary chromatic number:

Definition 2 *The **binary chromatic number** of a graph G, $\chi_B(G)$ is the least integer $k+1$ such that, for all $c\in\{0,\ldots,k+1\}$, there exists a partition of $V(G)$ into c cliques and $k+1-c$ cocliques.*

The binary chromatic number [4] is called the "colouring number" of a hereditary property by Bollobás and Thomason [9] and again by Bollobás [5] and is called the parameter $\tau(H)$ in Prömel and Steger [21]. The term indicates its generalizibility to multicolorings of the edges of K_n, or $K_{n,n}$ as in [3].

The binary chromatic number gives the value of $\mathrm{Dist}(n,\mathrm{Forb}(H))$ to within a multiplicative factor of 2, asymptotically:

Theorem 3 ([4]) *If H is a graph with binary chromatic number $\chi_B(H)=k+1$, then $\left(\frac{1}{2k}-o(1)\right)\binom{n}{2}\leq\mathrm{Dist}\left(n,\mathrm{Forb}(H)\right)\leq\frac{1}{k}\binom{n}{2}$.*

1.2.1 Known values of d^* and p^*

In [4], a large class of graphs H for which $d^*(\mathrm{Forb}(H))$ is known to be the lower bound in Theorem 3 is described. Namely, if a graph H has the property that $\chi_B(H) = k+1$ and there exist (A, c) and (a, C) such that each of the following occurs

- $V(H)$ cannot be partitioned into c cliques and A cocliques,
- $V(H)$ cannot be partitioned into C cliques and a cocliques,
- $A + c = a + C = k$ and $c \leq k/2 \leq C$,

then

$$d^*(\mathrm{Forb}(H)) = \frac{1}{2k}.$$

It is observed in [1] that if H and (A, c) and (a, C) satisfy the conditions above, then $p^*(\mathrm{Forb}(H)) = 1/2$. Furthermore, if H is a self-complementary graph, then $A = C$ and $a = c$. So, $C + c = k$, which implies $c \leq k/2 \leq C$ and $(p^*(\mathrm{Forb}(H)), d^*(\mathrm{Forb}(H))) = (1/2, 1/(2k))$.

The edit distance from monotone properties is also well-known. A **monotone property** is, without loss of generality, closed under the removal of either vertices or edges. Let $\mathcal{M}$ be a monotone property of graphs. The theorems of Erdős and Stone [15] and Erdős and Simonovits [14] give that

$$d^*(\mathcal{M}) = 1/r, \qquad \text{where } r = \min\{\chi(F) - 1 : F \notin \mathcal{M}\} \qquad \text{and} \qquad p^*(\mathcal{M}) = 1.$$

Alon and Stav, in [2], prove that $d^*(\mathrm{Forb}(K_{1,3})) = p^*(\mathrm{Forb}(K_{1,3})) = 1/3$. In this paper, we generalize this result to compute the pairs (p^*, d^*) for hereditary properties of the form $\mathrm{Forb}(K_a + E_b)$ and $\mathrm{Forb}\left(\overline{K_a + E_b}\right)$, where K_a is a complete graph on a vertices, E_b is an empty graph on b vertices and the "+" denotes a disjoint union of graphs. The claw $K_{1,3}$ is $\overline{K_3 + E_1}$.

In both [4] and in [2], more precise results for determining $\mathrm{Dist}(n, \mathcal{H})$ are given for several families of hereditary properties. For this paper, we concern ourselves exclusively with the first-order asymptotics.

Finally, in [2], a formula is given for the asymptotic value of the distance $\mathrm{Dist}(G(n, 1/2), \mathcal{H})$ for an arbitrary hereditary property $\mathcal{H}$. It generalizes the result, stated in [1] and implicit from arguments in [4], that almost surely, $\mathrm{Dist}(G(n, 1/2), \mathrm{Forb}(H)) = \frac{1}{2(\chi_B(\mathrm{Forb}(H))-1)}\binom{n}{2} - o(n^2)$. In this paper, we will further generalize this by determining an asymptotic expression for $\mathrm{Dist}(G(n, p), \mathcal{H})$ for all $p \in [0, 1]$.

1.3 Colored homomorphisms

Next we recall three definitions from [1] which are convenient for us.

Definition 4 *A* ***colored regularity graph (CRG)****, K, is a complete graph for which the vertices are partitioned $V(K) = \mathrm{VW}(K) \,\dot\cup\, \mathrm{VB}(K)$ and the edges are partitioned $E(K) = \mathrm{EW}(K) \,\dot\cup\, \mathrm{EG}(K) \,\dot\cup\, \mathrm{EB}(K)$. The sets* VW *and* VB *are the white and black vertices, respectively, and the sets* EW, EG *and* EB *are the white, gray and black edges, respectively.*

Bollobás and Thomason ([8],[10]) originate the use of this structure to define so-called basic hereditary properties. In particular, the paper [10] generalizes the enumeration of graphs with a given property $\mathcal{P}$ to the problem of computing the probability that $G(n,p) \in \mathcal{P}$. The problems are equivalent if $p = 1/2$. The papers use many of the techniques that are repeated or cited in the subsequent works on edit distance and use other nontrivial ideas.

Definition 5 *Let K be a CRG with $V(K) = \{v_1, \ldots, v_k\}$. The graph property $\mathcal{P}_{K,n}$ consists of all graphs J on n vertices for which there is an equipartition $\mathbf{A} = \{A_i : 1 \leq i \leq k\}$ of the vertices of J satisfying the following conditions for $1 \leq i < j \leq k$:*

- *if $v_i \in \mathrm{VW}(K)$, then A_i spans an empty graph in J,*
- *if $v_i \in \mathrm{VB}(K)$, then A_i spans a complete graph in J,*
- *if $\{v_i, v_j\} \in \mathrm{EW}(K)$, then (A_i, A_j) spans an empty bipartite graph in J,*
- *if $\{v_i, v_j\} \in \mathrm{EB}(K)$, then (A_i, A_j) spans a complete bipartite graph in J,*
- *if $\{v_i, v_j\} \in \mathrm{EG}(K)$, then (A_i, A_j) is unrestricted.*

*If all of the above holds, we say that the equipartition **witnesses the membership of J in $\mathcal{P}_{K,n}$**.*

Definition 6 *A **colored-homomorphism** from a (simple) graph F to a CRG, K, is a mapping $\varphi : V(F) \to V(K)$, which satisfies the following:*

1. *If $\{u, v\} \in E(F)$ then either $\varphi(u) = \varphi(v) = t \in \mathrm{VB}(K)$, or $\varphi(u) \neq \varphi(v)$ and $\{\varphi(u), \varphi(v)\} \in \mathrm{EB}(K) \cup \mathrm{EG}(K)$.*
2. *If $\{u, v\} \notin E(F)$ then either $\varphi(u) = \varphi(v) = t \in \mathrm{VW}(K)$, or $\varphi(u) \neq \varphi(v)$ and $\{\varphi(u), \varphi(v)\} \in \mathrm{EW}(K) \cup \mathrm{EG}(K)$.*

Moreover, a colored-homomorphism can be defined from a CRG, K', to another CRG, K'', that satisfies the following:

0. *If $v \in \mathrm{VB}(K')$, then $\varphi(v) \in \mathrm{VB}(K'')$. If $v \in \mathrm{VW}(K')$, then $\varphi(v) \in \mathrm{VW}(K'')$.*
1. *If $(u, v) \in \mathrm{EB}(K')$ then either $\varphi(u) = \varphi(v) = t \in \mathrm{VB}(K'')$, or $\varphi(u) \neq \varphi(v)$ and $(\varphi(u), \varphi(v)) \in \mathrm{EB}(K'') \cup \mathrm{EG}(K'')$.*
2. *If $(u, v) \in \mathrm{EW}(K')$ then either $\varphi(u) = \varphi(v) = t \in \mathrm{VW}(K'')$, or $\varphi(u) \neq \varphi(v)$ and $(\varphi(u), \varphi(v)) \in \mathrm{EW}(K'') \cup \mathrm{EG}(K'')$.*

Note that we can use the second definition to include the first, by defining $V(F) = \mathrm{VW}(F) \,\dot\cup\, \mathrm{VB}(F)$ in such a way as to make the colored-homomorphism legal with respect to the edge set.

Definition 7 *A CRG, K', is **induced** in another CRG, K, if there is a colored-homomorphism $\varphi : V(K') \to V(K)$ such that*

- *φ is an injection and*
- *for any $u, v \in V(K')$ for which $\{\varphi(u), \varphi(v)\} \in \mathrm{EG}(K)$, then $\{u, v\} \in \mathrm{EG}(K')$.*

Definition 8 *A CRG, K, is an **$\mathcal{H}$-colored regularity graph ($\mathcal{H}$-CRG)** for a hereditary property $\mathcal{H}$ if, for every graph $J \notin \mathcal{H}$, there is no colored-homomorphism from J to K.*

Denote $\mathcal{K}(\mathcal{H})$ to be the family of all CRGs K such that for every graph $J \notin \mathcal{H}$ there is no colored-homomorphism from J to K. If there is no colored-homomorphism from J to K, then this is denoted as $J \not\mapsto_c K$. If there is a colored-homomorphism from J to K, then this is denoted as $J \mapsto_c K$.

Observe that if $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$, then an $\mathcal{H}$-CRG, K, is one such that for all $H \in \mathcal{F}(\mathcal{H})$, there is no colored-homomorphism from H into K.

1.4 Functions of colored regularity graphs

1.4.1 Binary chromatic number

Previous edit distance results were expressed in terms of the so-called **binary chromatic number**, which can be viewed as an invariant on CRGs for which the edge set is gray.

Definition 9 *Let $K(a, c)$ denote the CRG with a white vertices, c black vertices and all edges gray.*

*The **binary chromatic number** of a hereditary property $\mathcal{H}$, denoted $\chi_B(\mathcal{H})$, is the least integer $k+1$ such that, $K(a, c) \notin \mathcal{K}(\mathcal{H})$ for all a, c such that $a + c = k + 1$. This definition means that $\chi_B(\mathrm{Forb}(H)) = \chi_B(H)$ for any graph H.*

This quantity is too specific for our purposes. We need to introduce a function that accounts for nongray edges in CRGs.

1.4.2 The function f

Given a CRG, K, we define two functions. If K has k vertices, with the usual notation for the edge sets and the vertex sets, then let

$$f_K(p) \stackrel{\text{def}}{=} \frac{1}{k^2}\left[p\left(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|\right) + (1-p)\left(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|\right)\right].$$

The function that defines $f_K(p)$ was introduced in [1] and corresponds to equipartitioning the vertex set of some G which is chosen according to the distribution $G(n, p)$ and mapping the parts of the partition to the vertices of K. So, f represents the expected proportion of edges that are changed under the rule that if an edge is mapped to a white edge or its endvertices are mapped to the same white vertex, then the edge is removed and if a nonedge is mapped to a black edge or its endvertices are mapped to the same black vertex, then the edge is added.

The function $f_K(p)$, as a function of p, is a line with a slope in $[-1, 1]$.

1.4.3 The function g

The function $g_K(p)$ is defined by a quadratic program. It corresponds not necessarily to an equipartition, but a partition with optimal sizes.

In order to define g, we first define some matrices: Let $\mathbf{W}_K$ denote the adjacency matrix of the graph defined by the white edges, along with the first $|\mathrm{VW}(K)|$ diagonal entries being 1 (corresponding to the white vertices) and the other diagonal entries being 0. Let $\mathbf{B}_K$ denote the adjacency matrix of the graph defined by the black edges along with the last $|\mathrm{VB}(K)|$ diagonal entries being 1 (corresponding to the black vertices) and the other diagonal entries being 0. We define the matrix $\mathbf{M}_K(p)$ as follows:

$$\mathbf{M}_K(p) = p\mathbf{W}_K + (1-p)\mathbf{B}_K.$$

With this, we define $g_K(p)$:

$$g_K(p) := \left\{ \begin{array}{lrcl} \min & \mathbf{u}^T\mathbf{M}_K(p)\mathbf{u} & & \\ \text{s.t.} & \mathbf{u}^T\mathbf{1} & = & 1 \\ & \mathbf{u} & \geq & \mathbf{0}. \end{array} \right. \tag{1}$$

If an optimal solution $\mathbf{u}'$ has zero entries, then $g_K(p) = g_{K^*}(p)$ for the CRG, K^*, induced in K, whose vertices correspond to the nonzero entries of $\mathbf{u}'$. (Note that K^* may depend on $\mathbf{u}'$.)

Lemma 10 *For any CRG K, and any $p \in [0,1]$, there exists a CRG K^*, where K^* is defined as a CRG induced in K by the vertices which correspond to nonzero entries of $\mathbf{u}'$, such that $g_K(p) = g_{K^*}(p) = \frac{1}{\mathbf{1}^T\mathbf{M}_{K^*}^{-1}(p)\mathbf{1}}$.*

We prove Lemma 10 in Section 3.2.

2 Results

2.1 General bounds

Theorem 11 is our main theorem, relating the functions f and g. For $p \in (0,1)$, the notation $G(n,p)$ is the random variable that represents a graph on n vertices chosen by a random process in which each edge is present independently with probability p. For $m \geq 1$, $G(n,m)$ is the random variable that represents a graph on n vertices chosen uniformly at random from all n vertex graphs with $\lfloor m \rfloor$ edges.

Theorem 11 *For a hereditary property $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$, let $\mathcal{K}(\mathcal{H})$ denote all CRGs K such that $H \not\mapsto_c K$ for each $H \in \mathcal{F}(\mathcal{H})$. Then, $d^*(\mathcal{H}) \stackrel{\text{def}}{=} \lim_{n\to\infty} \mathrm{Dist}(n,\mathcal{H})/\binom{n}{2}$ exists. Define*

$$f(p) \stackrel{\text{def}}{=} \inf_{K\in\mathcal{K}(\mathcal{H})} f_K(p) \qquad \textit{and} \qquad g(p) \stackrel{\text{def}}{=} \inf_{K\in\mathcal{K}(\mathcal{H})} g_K(p).$$

Then it is the case that $f(p) = g(p)$ for all $p \in [0,1]$,

$$d^*(\mathcal{H}) = \max_{p\in[0,1]} f(p) = \max_{p\in[0,1]} g(p),$$

and $p^*(\mathcal{H})$ is the value of p at which f achieves its maximum. In addition, the function $f(p) = g(p)$ is concave.

Furthermore, for all $p \in (0,1)$,

$$\max_{G:e(G)=p\binom{n}{2}} \{\mathrm{Dist}(G,\mathcal{H})\} = f(p)\binom{n}{2} + o(n^2),$$

and for all $\epsilon > 0$, $\mathrm{Dist}\left(G\left(n, p\binom{n}{2}\right), \mathcal{H}\right) \geq f(p)\binom{n}{2} - \epsilon n^2$, with probability approaching 1 as $n \to \infty$.

Of course, by definition, $\mathrm{Dist}(n, \mathrm{Forb}(\mathcal{H})) = d^*(\mathcal{H})\binom{n}{2} + o(n^2)$.

Remark: The main theorem of Alon and Stav [1] states, informally, that there exists a $p^* = p^*(\mathcal{H})$ such that $\mathrm{Dist}(n,\mathcal{H}) = \mathrm{Dist}(G(n,p^*),\mathcal{H})$. Here, we compute the first-order asymptotic of the edit distance and show that $f(p)\binom{n}{2}$ is asymptotically the maximum edit distance among all graphs of density p, and is achieved by the random graph $G(n, p\binom{n}{2})$. Informally, $\mathrm{Dist}\left(G(n, p\binom{n}{2}), \mathcal{H}\right) = f(p)\binom{n}{2} + o(n^2)$ and in the proof, we show, that $\mathrm{Dist}\left(G(n,p), \mathcal{H}\right) = f(p)\binom{n}{2} + o(n^2)$ as well.

In addition, Theorem 11 has the advantage that the edit distance can be computed, asymptotically, without direct use of Szemerédi's Regularity Lemma. As we see in Theorems 12, 13, 14 and 15, the function $f(p)$ is very useful in computing the values of $(p^*(\mathcal{H}), d^*(\mathcal{H}))$.

The method for computing (p^*, d^*) in this paper follows the same pattern for every hereditary property.

Method for computing edit distance:

Upper bound: Carefully choose CRGs, $K', K'' \in \mathcal{K}(\mathcal{H})$ (possibly $K' = K''$) and compute $\max_{p\in[0,1]} \min\{g_{K'}(p), g_{K''}(p)\}$. This maximum is an upper bound for $d^*(\mathcal{H})$.

Lower bound: Let p^* be the value of p at which the function $\min\{g_{K'}(p), g_{K''}(p)\}$ achieves its maximum. For any $K \in \mathcal{K}(\mathcal{H})$, we try to show that $f_K(p^*)$ is at least the upper bound value. If this is the case, then we have computed $d^*(\mathcal{H})$; moreover, $p^*(\mathcal{H})$ is the p^* provided above. In order to do this, we use a type of weighted Turán theorem.

2.2 The edit distance of $K_a + E_b$

We give a class of graphs in which neither the upper nor the lower bounds given by the binary chromatic number hold.

Theorem 12 *Let $a \geq 2$ and $b \geq 1$ be positive integers. Let $H = K_a + E_b$, the disjoint union of an a-clique and a b-coclique. Then,*

$$d^*\left(\mathrm{Forb}(K_a + E_b)\right) = \frac{1}{a+b-1} \qquad \textit{and} \qquad p^*\left(\mathrm{Forb}(K_a + E_b)\right) = \frac{a-1}{a+b-1},$$

i.e., $\mathrm{Dist}(n, \mathrm{Forb}(K_a \cup E_b)) = \frac{1}{a+b-1}\binom{n}{2} - o(n^2)$.

We note that $\chi_B(K_a + E_b) = \max\{a, b+1\}$ and so Theorem 12 is an improvement over [4] in the case when $a \neq b+1$. It is also an improvement over Proposition 17, which appears below, in the case when $b > 1$ and $a > 2$. Alon and Stav [2] prove the case when $a = 3$ and $b = 1$, the complement of the "claw," $K_{1,3}$.

2.3 A few specific graphs

In all known examples of hereditary properties $\mathcal{H}$, the point at which $(p^*(\mathcal{H}), d^*(\mathcal{H}))$ occurs is either the intersection of two curves $g_{K'}(p)$, $g_{K''}(p)$ or is the maximum of a single curve $g_{K'}(p)$. In either case, each CRG can be chosen to be one with only gray edges.

We compute the edit distance of two hereditary properties that demonstrate the complexity of both p^* and d^*.

2.3.1 The graph $K_{3,3}$

The graph $K_{3,3}$ has d^* and p^* defined by the local maximum of a single curve $g_{K'}(p)$.

Theorem 13 *The complete bipartite graph $K_{3,3}$ satisfies*

$$p^*\left(\mathrm{Forb}(K_{3,3})\right) = \sqrt{2} - 1 \quad \textit{and} \quad d^*\left(\mathrm{Forb}(K_{3,3})\right) = 3 - 2\sqrt{2}.$$

Moreover, p^* is the local maximum of $g_{K'}(p)$, where K' consists of one white vertex, two black vertices and all gray edges.

It should be noted that neither p^* nor d^* could be determined for this hereditary property by the intersection of a finite number of f curves, simply because such intersections would occur at rational points. So, a sequence of CRGs would be required. By using the g curves, however, we need only to use a single CRG.

2.3.2 The graph H_9

Here, the graph we construct is formed by taking C_9^2 and adding a triangle. That is, if the vertices are $\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, then $i \sim j$ iff $i - j \in \{\pm 1, \pm 2\} \pmod 9$ or both i and j are congruent to 0 modulo 3. For notational simplicity, we call this graph H_9. See Figure 1.

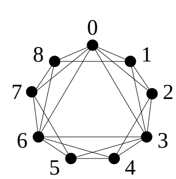

Figure 1: The graph H_9.

An upper bound on $d^*(\mathrm{Forb}(H_9))$ is defined by the intersection of two curves, $g_{K'}(p)$, $g_{K''}(p)$, one of which corresponds to a CRG that has one black edge. For this graph H_9, it is impossible to only consider CRGs which have all edges gray. It was a folklore belief that for every graph it is sufficient to consider CRGs which have all edges gray, H_9 is the first example showing that this belief is false.

Theorem 14 *The graph H_9 satisfies*

$$d^*\left(\mathrm{Forb}(H_9)\right) \leq \frac{3-\sqrt{5}}{4}.$$

Moreover, this value occurs at the intersection of $g_{K'}(p)$ and $g_{K''}(p)$, where K' consists of two black vertices and a gray edge and K'' consists of four white vertices, a black edge and 5 gray edges.

In the proof, we show that if only gray-edge CRGs are used, then the upper bound on d^* could be no less than $1/5 = 0.2$, but $\frac{3-\sqrt{5}}{4} \approx 0.191$. The lower bound, from Theorem 3, is $d^*(\mathrm{Forb}(H_9)) \geq 1/6 \approx 0.167$.

2.4 4-vertex graphs

In [2], Alon and Stav compute $(p^*(\mathrm{Forb}(H)), d^*(\mathrm{Forb}(H)))$ for all H on at most 4 vertices. Except for P_3+K_1 and its complement, all such graphs H are either covered by Theorem 16 (see also [4]) or are of the form $K_a + E_b$ or $\overline{K_a + E_b}$, which is covered by Theorem 12. Here we give a short and different proof, using Lemma 18, for $\overline{P_3 + K_1}$, which consists of a triangle and a pendant edge.

Theorem 15 *The graph $\overline{P_3 + K_1}$ satisfies*

$$p^*\left(\mathrm{Forb}(\overline{P_3+K_1})\right) = 2/3 \qquad \text{and} \qquad d^*\left(\mathrm{Forb}(\overline{P_3+K_1})\right) = 1/3.$$

3 Basic tools

3.1 Improved binary chromatic number bounds

Lemma 10, along with Theorem 11, yields a proof of a somewhat better upper bound for $\mathrm{Dist}(n, \mathcal{H})$, based on the binary chromatic number.

Recall that $K(a, c)$ denotes the CRG that consists of a white vertices, c black vertices and only gray edges. If $k = \chi_B(\mathcal{H}) - 1$, then let $c_{\min}$ be the least c so that $K(k - c, c) \notin \mathcal{K}(\mathcal{H})$. Let $c_{\max}$ be the greatest such number. For $\mathcal{H} = \mathrm{Forb}(H)$, there exists an upper bound that can be expressed in terms of the binary chromatic number of H and corresponding $c_{\min}$ and $c_{\max}$.

Theorem 16 ([4]) *Let H be a graph with binary chromatic number $k+1$ and $c_{\min}$ and $c_{\max}$ be defined as above. If $c_{\min} \leq k/2 \leq c_{\max}$, then*

$$d^*(\mathrm{Forb}(H)) = \frac{1}{2k}.$$

Otherwise, let c_0 be the one of $\{c_{\max}, c_{\min}\}$ that is closest to $k/2$. Then

$$d^*(\mathrm{Forb}(H)) \leq \left(\frac{1}{1+2\sqrt{\frac{c_0}{k}\left(1-\frac{c_0}{k}\right)}}\right)\frac{1}{k} \leq \frac{1}{k}.$$

Proposition 17 improves this general upper bound, not only trivially by extending it to general hereditary properties, but also by improving the case when $c_{\max} = 0$ or $c_{\min} = k$.

Proposition 17 *Let $\mathcal{H}$ be a hereditary property with $k+1 = \chi_B(\mathcal{H})$ and $c_0, c_{\max}, c_{\min}$ defined analogously to Theorem 16. The bounds in Theorem 16 hold for $\mathcal{H}$. Furthermore, if $\mathcal{H} \neq \mathrm{Forb}(K_{k+1})$, then*

$$d^*(\mathcal{H}) \leq \frac{1}{k+1}.$$

Note that $d^*(\mathrm{Forb}(K_{k+1})) = \frac{1}{k}$ by Turán's theorem.

Proof. If we restrict our attention to the CRGs in $\mathcal{K}(\mathcal{H})$ which are of the form $K(a,c)$, then Theorem 11 gives that

$$d^*(\mathcal{H}) \leq \max_{p\in[0,1]} \inf_{K(a,c)\in\mathcal{K}(\mathcal{H})} \left\{g_{K(a,c)}(p)\right\} = \max_{p\in[0,1]} \min_{K(a,c)\in\mathcal{K}(\mathcal{H})} \left\{\frac{p(1-p)}{a(1-p)+cp}\right\}.$$

A word on why the "inf" is made into a "min": Recall that if $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$, then $K \in \mathcal{K}(\mathcal{H})$ means that $H \not\mapsto_c K$ for all $H \in \mathcal{F}(\mathcal{H})$. Choose some $H_0 \in \mathcal{F}(\mathcal{H})$. In order for $K \in \mathcal{K}(\mathcal{H})$, it must be the case that $H_0 \not\mapsto_c K$. But, there are only a finite number of pairs (a,c) such that $H_0 \not\mapsto_c K(a,c)$. Indeed, $H_0 \mapsto_c K(a,c)$ if either $a \geq \chi(H_0)$ or $c \geq \chi(\overline{H_0})$. Therefore, regardless of $\mathcal{H}$, there are only a finite number of (a,c) for which $K(a,c) \in \mathcal{K}(\mathcal{H})$.

Suppose there exist different pairs (a,C) and (A,c) such that $a+C = A+c = k$ and $c \leq k/2 \leq C$. We bound $d^*(\mathcal{H})$ by $\max_{p\in[0,1]} \min\left\{g_{K(a,C)}(p), g_{K(A,c)}(p)\right\}$. If $p < 1/2$, then

$$\begin{aligned} C(1-2p) &> c(1-2p) \\ -C(1-p)+Cp &< -c(1-p)+cp \\ (k-C)(1-p)+Cp &< (k-c)(1-p)+cp \\ a(1-p)+Cp &< A(1-p)+cp \\ g_{K(a,C)}(p) &> g_{K(A,c)}(p). \end{aligned}$$

Similarly, if $p > 1/2$, then $g_{K(a,C)}(p) < g_{K(A,c)}(p)$. So, $\max_{p\in[0,1]} \min\left\{g_{K(a,C)}(p), g_{K(A,c)}(p)\right\}$ occurs at the intersection of the two curves, which is $(1/2, 1/(2k))$.

Otherwise, let c_0 be the value of c for which $K(k-c,c) \in \mathcal{K}(\mathcal{H})$ that is the closest to $k/2$. Without loss of generality, assume that $c_0 = c < k/2$. If $c_0 > 0$, then $k \geq 3$ and we may bound $d^*(\mathcal{H})$ by $g_{K(k-c_0,c_0)}(p)$, which achieves its maximum at $p = \frac{\sqrt{k-c_0}}{\sqrt{k-c_0}+\sqrt{c_0}}$ and this maximum is

$$\begin{aligned}
\frac{1}{\left(\sqrt{k-c_0}+\sqrt{c_0}\right)^2} &= \left(\frac{1}{1+2\sqrt{\frac{c_0}{k}\left(1-\frac{c_0}{k}\right)}}\right)\frac{1}{k} \\
&\leq \left(\frac{1}{1+2\sqrt{\frac{1}{k}\left(1-\frac{1}{k}\right)}}\right)\frac{1}{k} = \frac{1}{k+2\sqrt{k-1}} < \frac{1}{k+1}.
\end{aligned}$$

Finally, consider the case when $c_0 = 0$. If there exists some $K(\alpha,\gamma) \in \mathcal{K}(\mathcal{H})$ with $\gamma \geq 1$, then $g_{K(\alpha,\gamma)}(p)$ intersects $g_{K(k,0)}(p)$ at $p = \frac{k-\alpha}{k-\alpha+\gamma}$ and the value of each function at that point is $\frac{k-\alpha}{k(k-\alpha+\gamma)} \leq \frac{1}{k+1}$, which has equality only if $\gamma = 1$ and $\alpha = 0$. If there is no such $K(\alpha,\gamma)$, then, with $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$, each $H \mapsto_c K(0,1)$. Therefore, each H is a clique and so the smallest one defines $\mathcal{H}$. The fact that $\chi_B(\mathcal{H}) = k+1$ requires $\mathcal{H} = \mathrm{Forb}(K_{k+1})$.

So, $d^*(\mathcal{H}) \leq \frac{1}{k+1}$ unless $\mathcal{H} = \mathrm{Forb}(K_{k+1})$. □

3.2 Proof of Lemma 10

Let $\mathbf{u}'$ be an optimal solution of (1) and, among such solutions, it is one with the most number of zero entries. Let K^* be a CRG that corresponds to $\mathbf{u}'$. Let $\mathbf{M}_{K^*}(p)$ be the associated matrix and $\mathbf{u}^*$ be the vector formed by removing the zero entries from $\mathbf{u}'$. By assumption, $\mathbf{u}^*$ is an optimal solution of

$$g_{K^*(p)} := \left\{\begin{array}{lrcl} \min & \mathbf{u}^T\mathbf{M}_{K^*}(p)\mathbf{u} & & \\ \text{s.t.} & \mathbf{u}^T\mathbf{1} & = & 1 \\ & \mathbf{u} & \geq & \mathbf{0}, \end{array}\right. \tag{2}$$

where from K^* the vertices that correspond to the deleted 0 coordinates of $\mathbf{u}'$ and the corresponding rows and columns of $\mathbf{M}_{K^*}(p)$ are removed. Furthermore, all entries of $\mathbf{u}^*$ are strictly positive.

Suppose $\mathbf{M}_{K^*}(p)$ is not invertible, with $\mathbf{M}_{K^*}(p)\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}^T\mathbf{1} \neq 0$. Then rescale $\mathbf{x}$ so that $\mathbf{x}^T\mathbf{1} = 1$. Choose an $\epsilon > 0$ such that $(1-\epsilon)\mathbf{u}^* + \epsilon\mathbf{x}$ has all nonnegative entries. This is possible and is a feasible solution to the quadratic program (2), producing the value

$$((1-\epsilon)\mathbf{u}^* + \epsilon\mathbf{x})^T\,\mathbf{M}_{K^*}(p)\,((1-\epsilon)\mathbf{u}^* + \epsilon\mathbf{x}) = (1-\epsilon)^2(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^*,$$

which contradicts the presumed optimal value.

Suppose $\mathbf{M}_{K^*}(p)$ is not invertible, with $\mathbf{M}_{K^*}(p)\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}^T\mathbf{1} = 0$. Then rescale $\mathbf{x}$ so that $\mathbf{u}^* + \mathbf{x}$ has nonnegative entries and at least one zero entry. This is a feasible solution to (2), producing the value

$$(\mathbf{u}^* + \mathbf{x})^T \mathbf{M}_{K^*}(p) (\mathbf{u}^* + \mathbf{x}) = (\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^*,$$

but this contradicts the value of $\mathbf{u}^*$ because this vector has zero entries. More zero entries can be appended to create a solution of (1) which has more zeros than $\mathbf{u}'$.

Therefore, we may assume that $\mathbf{M}_{K^*}(p)$ is invertible. Note that both it and its inverse are symmetric matrices. Define the following vector: $\mathbf{z} := \mathbf{M}_{K^*}(p)^{-1}\mathbf{1}/\left(\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}\right)$. Choose $\epsilon > 0$ small enough so that both $\frac{1}{1+\epsilon}(\mathbf{u}^* + \epsilon\mathbf{z})$ and $\frac{1}{1-\epsilon}(\mathbf{u}^* - \epsilon\mathbf{z})$ have all entries nonnegative. Such an ϵ exists because all entries of $\mathbf{u}^*$ are positive.

These are each feasible solutions and when $\frac{1}{1\pm\epsilon}(\mathbf{u}^* \pm \epsilon\mathbf{z})$ is placed in (2), it gives the value

$$\begin{aligned}
&\frac{1}{(1\pm\epsilon)^2}(\mathbf{u}^* \pm \epsilon\mathbf{z})^T \mathbf{M}_{K^*}(p) (\mathbf{u}^* \pm \epsilon\mathbf{z}) \\
&= \frac{1}{(1\pm\epsilon)^2}\left[(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^* \pm 2\epsilon(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{z} + \epsilon^2\mathbf{z}^T\mathbf{M}_{K^*}(p)\mathbf{z}\right] \\
&= \frac{1}{(1\pm\epsilon)^2}\left[(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^* \pm 2\epsilon\frac{(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}}\right. \\
&\qquad\qquad \left. +\epsilon^2\frac{\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{M}_{K^*}(p)\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}}{\left(\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}\right)^2}\right] \\
&= \frac{1}{(1\pm\epsilon)^2}\left[(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^* + \frac{\epsilon^2 \pm 2\epsilon}{\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}}\right] \\
&= (\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^* + \epsilon(\pm 2 + \epsilon)\left[\frac{1}{\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}} - (\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^*\right].
\end{aligned}$$

If $(\mathbf{u}^*)^T\mathbf{M}_{K^*}(p)\mathbf{u}^* \neq \left(\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}\right)^{-1}$, then either $\mathbf{u}^* + \epsilon\mathbf{z}$ or $\mathbf{u}^* - \epsilon\mathbf{z}$ is a better solution to (2) than $\mathbf{u}^*$, a contradiction.

So, (2), hence (1), has value $\left(\mathbf{1}^T\mathbf{M}_{K^*}(p)^{-1}\mathbf{1}\right)^{-1}$. □

4 Proof of Theorem 11

Our proof has the following outline:

A. Show that every graph G on n vertices and $p\binom{n}{2}$ edges has $\mathrm{Dist}(G, \mathcal{H}) \leq f(p)\binom{n}{2}$.

B. Show that f is continuous and so it achieves its maximum.

C. Show that, for any fixed p and for ϵ small enough, $\mathrm{Dist}\left(G(n,p), \mathcal{H}\right) \geq f(p)\binom{n}{2} - 2\epsilon n^2$ for n sufficiently large.

D. Show that $g(p) = f(p)$ for all p.

E. Show that $g(p)$ is concave.

A: Upper bound.

Recall that $f(p) = \inf_{K \in \mathcal{K}(\mathcal{H})} f_K(p)$ and $g(p) = \inf_{K \in \mathcal{K}(\mathcal{H})} g_K(p)$.

Let G be an arbitrary graph on n vertices with $p\binom{n}{2}$ edges. Let $K \in \mathcal{K}(\mathcal{H})$ with $k = |\mathrm{VB}(K)| + |\mathrm{VW}(K)|$. We will randomly partition $V(G)$ into k pieces and delete and add edges in a manner determined by K. For each $v \in V(G)$, randomly, and independently from other vertices, place v into V_i with probability $1/k$. Moreover, label the vertices of K with $\{v_1, \ldots, v_k\}$. Create G' from G by performing the following action for each distinct i and j in $[k]$:

- If $v_i \in \mathrm{VW}(K)$, then delete the edges in G having both endpoints in V_i.
- If $v_i \in \mathrm{VB}(K)$, then add the non-edges in G having both endpoints in V_i.
- If $\{v_i, v_j\} \in \mathrm{EW}(K)$, then delete the edges in G having one endpoint in V_i and the other in V_j.
- If $\{v_i, v_j\} \in \mathrm{EB}(K)$, then add the edges in G having one endpoint in V_i and the other in V_j.

If there is an induced copy of H in G', then there is a colored-homomorphism from H to K. Since $K \in \mathcal{K}(\mathcal{H})$, there is no $H \in \mathcal{F}(\mathcal{H})$ for which $H \mapsto_c K$. Thus, $G' \in \mathcal{H}$.

The probability that an edge is deleted is $(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|)/k^2$ and the probability that a nonedge is added is $(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|)/k^2$. Therefore, expected number of changes is

$$p\binom{n}{2}\frac{|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|}{k^2} + (1-p)\binom{n}{2}\frac{|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|}{k^2} = f_K(p)\binom{n}{2}.$$

This implies that there is a partition which results in at most $f_K(p)\binom{n}{2}$ changes in order to transform G into some $G' \in \mathcal{H}$, i.e., $\mathrm{Dist}(G, \mathcal{H})/\binom{n}{2} \le f_K(p)$. Since this is true for any $K \in \mathcal{K}(\mathcal{H})$, $\mathrm{Dist}(G, \mathcal{H})/\binom{n}{2} \le \inf_{K \in \mathcal{K}(\mathcal{H})} f_K(p) = f(p)$.

B: Continuity of f.

We differ slightly from Alon and Stav in their approach in [1] to ensure the continuity of f. For terminology and citations of the theorems below, see chapter 7 of Rudin [22].

The set $\mathcal{K}(\mathcal{H})$ is countable since the set of finite CRGs is countable. Therefore, we can linearly order the members of $\mathcal{K}(\mathcal{H})$ as $K_1, K_2, \ldots$. Let $m_n(p) = \min_{i \le n} f_{K_i}(p)$ and $f(p) = \inf_i f_{K_i}(p)$. Since each $f_{K_i}(p)$ is a line with slope in $[-1, 1]$, each $m_n(p)$ is Lipschitz with coefficient 1. So, $\{m_n\}$ forms an equicontinuous, pointwise bounded family. As such, $\{m_n\}$ has a uniformly convergent subsequence. The limit must, therefore, be continuous. Since $m_n \to f$ pointwise, this limit is $f(p)$.

Since f is continuous, it achieves its maximum in the closed interval $[0, 1]$. Therefore, $\mathrm{Dist}(n, \mathcal{H}) \le \max_{p \in [0,1]} f(p)$. Define p^* so that $f(p^*)$ is this maximum. Note that if some lines are horizontal then p^* is not necessarily unique.

C: Lower bound for the random graph.

Fix $p \in (0,1)$ and $\epsilon > 0$. Let $S = S(\epsilon, \mathcal{H})$ the function provided by the generalization of the Regularity Lemma, cited as Lemma 2.7 in [1]. The proof below follows ideas similar to those in [1]. Let $G \sim G(n,p)$. A routine application of the Chernoff bound (see [16]) gives that the probability that every equipartition of $V(G)$ into $k \leq S$ pieces $V_1 \dot\cup \cdots \dot\cup V_k$ has the subgraphs $G[V_i]$ and the bipartite subgraphs $G[(V_i,V_j)]$ with density in $(p - n^{-0.4}, p + n^{-0.4})$ for all distinct $i,j \in [k]$ is at most $\exp\{-\Omega(n^{1.2})\}$, with p and S fixed. Choose n to be large enough for such a graph to exist and choose G to be one such graph.

Let $G' \in \mathcal{H}$ have the property that $\mathrm{Dist}(G,G') = \mathrm{Dist}(G,\mathcal{H})$. Apply the generalization of the Regularity Lemma to G', with parameters ϵ and $m = 2\epsilon^{-1}$. There is an $S = S(\epsilon, \mathcal{H})$ such that there is an equipartition of the vertex set: $V(G') = V_1 \dot\cup \cdots \dot\cup V_k$, with $m \leq k \leq S$. Each piece is of size either $L \stackrel{\mathrm{def}}{=} \lfloor n/k \rfloor$ or $\lceil n/k \rceil$.

The graph G'' is constructed from this partition in such a way as to ensure that $G''[V_i]$ is either an empty or complete graph and either $d_{G''}(V_i,V_j) = 0$ or $d_{G''}(V_i,V_j) = 1$ or $\epsilon/2 \leq d_{G''}(V_i,V_j) \leq 1 - \epsilon/2$. This is done by deleting edges from sparse clusters and pairs and adding edges to dense clusters and pairs. Consequently, $\mathrm{Dist}(G',G'') < (\epsilon/2)n^2$.

This naturally yields a CRG, K, on the vertex set $\{v_1, \ldots, v_k\}$ where v_i is $\{\text{white}, \text{black}\}$ iff $G''[V_i]$ is $\{\text{empty}, \text{complete}\}$ and $\{v_i, v_j\}$ is $\{\text{white}, \text{black}\}$ iff $\{d_{G''}(V_i,V_j) = 0,$ $d_{G''}(V_i,V_j) = 1\}$; otherwise $\{v_i, v_j\}$ is gray. If there is a colored-homomorphism from $H \in \mathcal{F}(\mathcal{H})$ to K, then the construction of G'' ensures that H is induced in both G'' and G'. Therefore, K must be in $\mathcal{K}(\mathcal{H})$.

Since the distance between graphs is simply a symmetric difference of edges, we see immediately that the triangle inequality applies:

$$\begin{aligned}
\mathrm{Dist}(G,G') &\geq \mathrm{Dist}(G,G'') - \mathrm{Dist}(G',G'') \\
&\geq \mathrm{Dist}(G,G'') - (\epsilon/2)n^2 \\
&\geq \left(p - n^{-0.4}\right)\binom{L}{2}|\mathrm{VW}(K)| + \left(1 - p - n^{-0.4}\right)\binom{L}{2}|\mathrm{VB}(K)| \\
&\quad + \left(p - n^{-0.4}\right)L^2|\mathrm{EW}(K)| + \left(1 - p - n^{-0.4}\right)L^2|\mathrm{EB}(K)| - (\epsilon/2)n^2 \\
&\geq \frac{1}{k^2}\left(p|\mathrm{VW}(K)| + (1-p)|\mathrm{VB}(K)|\right)\binom{n}{2}\frac{(n-k)(n-2k)}{n(n-1)} \\
&\quad + \frac{1}{k^2}\left(p|\mathrm{EW}(K)| + (1-p)|\mathrm{EB}(K)|\right)\binom{n}{2}\frac{(n-k)^2}{n(n-1)} \\
&\quad - \frac{n^{1.6}}{2k} - \frac{n^{1.6}}{2} - (\epsilon/2)n^2.
\end{aligned}$$

For n large enough,

$$\begin{aligned}
\mathrm{Dist}(G,G') &\geq \frac{1}{k^2}\left(p|\mathrm{VW}(K)|+(1-p)|\mathrm{VB}(K)|\right)\binom{n}{2}\left(1-\frac{3k}{n}\right) \\
&\quad +\frac{1}{k^2}\left(p|\mathrm{EW}(K)|+(1-p)|\mathrm{EB}(K)|\right)\binom{n}{2}\left(1-\frac{2k}{n}\right)-\frac{3\epsilon}{4}n^2 \\
&\geq f_K(p)\binom{n}{2}-\epsilon n^2.
\end{aligned}$$

So, for each sufficiently small $\epsilon > 0$, the probability that $G \sim G(n,p)$ satisfies $\mathrm{Dist}(G,\mathcal{H}) \geq f(p)\binom{n}{2} - \epsilon n^2$ approaches 1 as $n \to \infty$.

The only place where randomness is used above is to show that, with respect to any equipartition with $k \leq S$ parts, the density of the pairs is close to p. This is true for $G\left(n, p\binom{n}{2}\right)$ as well, therefore we conclude that for all ϵ sufficiently small, the probability that $G \sim G\left(n, p\binom{n}{2}\right)$ satisfies $\mathrm{Dist}(G,\mathcal{H}) \geq f(p)\binom{n}{2} - \epsilon n^2$ approaches 1 as $n \to \infty$.

Thus, $f(p)$ is the supremum of $\mathrm{Dist}(G,\mathcal{H})$ for graphs G of density p and $\mathrm{Dist}(n,\mathcal{H})/\binom{n}{2} = f(p^*) - o(1)$.

D: Equality of f and g.

We address the g functions. Recalling (1),

$$g_K(p) = \left\{\begin{array}{lll} \min & \mathbf{w}^T\mathbf{M}_K(p)\mathbf{w} & \\ \text{s.t.} & \mathbf{w}^T\mathbf{1} & = \;1 \\ & \mathbf{w} & \geq \;\mathbf{0}. \end{array}\right.$$

If K has k vertices, then $\mathbf{w} = \frac{1}{k}\mathbf{1}$ is a feasible solution, and $g_K(p) \leq f_K(p)$ for all $p \in [0,1]$. Thus, $g(p) \leq f(p)$.

Fix $p \in [0,1]$ and $\epsilon \in (0,1)$ and choose a $K^* \in \mathcal{K}(\mathcal{H})$ such that $g_{K^*}(p) \leq g(p) + \epsilon/2$ and an optimal solution in the corresponding quadratic program, $\mathbf{u}^* = (u_1,\ldots,u_k)$, has strictly positive entries. We will find a CRG, L (for which the ℓ clusters are equally weighted), that will approximate the weighted version of K^*. Set $\ell > 5k\epsilon^{-1}$. Construct a CRG, L, on ℓ vertices such that there are $\lfloor u_i\ell\rfloor$ or $\lceil u_i\ell\rceil$ copies of vertex x_i of K^* in the natural way: Let y' be a copy of x_i and y'' be a copy of x_j. The vertex y' has the same color as x_i and y'' has the same color as x_j. If $i \neq j$, then $\{y',y''\}$ has the same color as $\{x_i,x_j\}$. If $i = j$, then $\{y',y''\}$ has the same color as vertex x_i.

Let $\tilde{\mathbf{u}} = (\lceil u_1\ell\rceil,\ldots,\lceil u_k\ell\rceil)$ and $\mathbf{d} = \tilde{\mathbf{u}} - \ell\mathbf{u}^*$. Hence, coordinatewise, $\mathbf{0} \leq \mathbf{d} \leq k\mathbf{1}$. We can upper bound the f function of L:

$$
\begin{aligned}
f_L(p) &= \frac{1}{\ell^2}(\tilde{\mathbf{u}})^T \mathbf{M}_{K^*}(p)\tilde{\mathbf{u}} \\
&= \frac{1}{\ell^2}\left(\ell\mathbf{u}^* + \mathbf{d}\right)^T \mathbf{M}_{K^*}(p)\left(\ell\mathbf{u}^* + \mathbf{d}\right) \\
&= (\mathbf{u}^*)^T \mathbf{M}_{K^*}(p)\mathbf{u}^* + \frac{2}{\ell}\mathbf{u}^*\mathbf{M}_{K^*}(p)\mathbf{d} + \frac{1}{\ell^2}\mathbf{d}^T\mathbf{M}_{K^*}(p)\mathbf{d} \\
&\leq g_{K^*}(p) + \frac{2}{\ell}(\mathbf{u}^*)^T\mathbf{J}\mathbf{1} + \frac{1}{\ell^2}\mathbf{1}^T\mathbf{J}\mathbf{1} \\
&= g_{K^*}(p) + \frac{2k}{\ell}(\mathbf{u}^*)^T\mathbf{1} + \frac{k}{\ell^2}\mathbf{1}^T\mathbf{1} \\
&= g_{K^*}(p) + \frac{2k}{\ell} + \frac{k^2}{\ell^2},
\end{aligned}
$$

where $\mathbf{J}$ is the all ones $k \times k$ matrix. Since $k/\ell < \epsilon/5 < 1/5$, it is true that $2k/\ell + k^2/\ell^2 < \epsilon/2$. Therefore,

$$
f(p) \leq f_L(p) < g_{K^*}(p) + \frac{\epsilon}{2} < g(p) + \epsilon,
$$

for all $\epsilon \in (0, 1)$, yielding $f(p) = g(p)$.

E: Concavity of $f(p)$.

A function h is concave on an interval domain if, whenever a and b are in the domain of h, then $h(ta + (1-t)b) \geq th(a) + (1-t)h(b)$ for all $t \in [0, 1]$.

For the function f, the infimum of linear functions,

$$
\begin{aligned}
f(ta + (1-t)b) &= \inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(ta + (1-t)b)\} = \inf_{K\in\mathcal{K}(\mathcal{H})}\{tf_K(a) + (1-t)f_K(b)\} \\
&\geq t\left(\inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(a)\}\right) + (1-t)\left(\inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(b)\}\right) \\
&= tf(a) + (1-t)f(b).
\end{aligned}
$$

This concludes the proof of Theorem 11. □

5 The computation of p^* and d^* for specific families

Let $t(n, k)$ denote the number of edges in the Turán graph on n vertices with no clique of order $k + 1$. The following is a result of elementary computation:

$$
\frac{k-1}{k}\frac{n^2}{2} - \frac{k}{8} \leq t(n, k) = \frac{k-1}{k}\frac{n^2}{2} - \frac{k}{2}\left(\left\lceil\frac{n}{k}\right\rceil - \frac{n}{k}\right)\left(\frac{n}{k} - \left\lfloor\frac{n}{k}\right\rfloor\right) \leq \frac{k-1}{k}\frac{n^2}{2}.
$$

5.1 General approach

To prove upper bounds on d^*, we use (1) and choose CRGs whose curves intersect at (p^*, d^*) or a curve that achieves its maximum at (p^*, d^*).

To prove lower bounds on d^*, we need to use a weighted Turán approach which seems to be quite difficult in general. To see a simple application of the weighted Turán method, we provide a very short proof of the lower bound in Theorem 3 below:

Let H be a graph with binary chromatic number χ_B and let K be any CRG for which $H \not\mapsto_c K$. This immediately implies that K contains no clique of order $\chi_B + 1$ whose edges are all gray.

In particular, this implies that $\mathrm{EG}(K) \leq t(k, \chi_B)$. Setting $p = 1/2$, we see that

$$\begin{aligned} f_K(1/2) &= \frac{1}{k^2}\left[\frac{1}{2}\left(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|\right) + \frac{1}{2}\left(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|\right)\right] \\ &= \frac{1}{k^2}\left[\frac{k}{2} + \left(|\mathrm{EW}(K)| + |\mathrm{EB}(K)|\right)\right] \\ &\geq \frac{1}{k^2}\left[\frac{k}{2} + \left(\binom{k}{2} - t(k, \chi_B)\right)\right] \\ &\geq \frac{1}{2} - \frac{1}{k^2} t(k, \chi_B) \\ &\geq \frac{1}{2} - \frac{\chi_B - 1}{2\chi_B} = \frac{1}{2\chi_B}, \end{aligned}$$

and this proves the lower bound of Theorem 3.

5.2 Edit distance of $K_a + E_b$

5.2.1 Upper bound

Here, we choose K' to have $|\mathrm{VW}(K')| = a - 1$, $|\mathrm{VB}(K')| = 0$ and all edges gray. Furthermore, we choose K'' to have $|\mathrm{VW}(K'')| = 0$, $|\mathrm{VB}(K'')| = b$ and all edges gray. It is easy to see that both $K_a + E_b \not\mapsto_c K'$ and $K_a + E_b \not\mapsto_c K''$. An easy computation gives that $g_{K'}(p) = \frac{p}{a-1}$ and $g_{K''}(p) = \frac{1-p}{b}$. The intersection of the two functions is at the point $(p^*, d^*) = \left(\frac{a-1}{a+b-1}, \frac{1}{a+b-1}\right)$. Moreover, the fact that $\min\{g_{K'}(p), g_{K''}(p)\}$ is strictly unimodal, means that our proof below that $g(p^*) \geq d^*$ means that p^* is the unique value at which $g(p)$ achieves its maximum.

5.2.2 Weighted Turán lemma

The following lemma can be considered to be a generalization of Turán's theorem. That is, if from Lemma 18 we only apply condition (1) but not condition (2), then the answer is a basic consequence of Turán.

Lemma 18 *Let $a \geq 2$ and let K be a CRG with the property that any set A of a vertices has at least one of the following conditions:*

(1) A contains at least one white edge,

(2) A contains a spanning subgraph of black edges.

Then

$$(a-1)\mathrm{EW}(K) + \mathrm{EB}(K) \geq \left\lceil \frac{n}{2}(n-a+1) \right\rceil .$$

Proof. We fix an integer $a \geq 2$ and proceed via induction on n. The base case, $n \leq a$ is trivial.

Now, we assume that any CRG, K', on $s < n$ vertices that satisfies the conditions of the lemma has $(a-1)\mathrm{EW}(K') + \mathrm{EB}(K') \geq \lceil (s/2)(s-a) \rceil$.

Let K be a CRG on n vertices. If it consists of only white edges, then

$$(a-1)\mathrm{EW}(K) + \mathrm{EB}(K) = (a-1)\mathrm{EW}(K) = (a-1)\binom{n}{2} \geq \left\lceil \frac{n}{2}(n-a+1) \right\rceil .$$

Let $V(K) - S$ be a maximal set of vertices that does not span a white edge. We may assume that $S \neq \emptyset$ because otherwise the minimum black degree is at least $n-a+1$, proving the claim of the theorem. By the maximality of $V(K) - S$, for any $s \in S$ there exists a $t \in V(K) - S$ such that $st \in \mathrm{EW}(K)$. Moreover, since there is no white edge in $V(K) - S$, vertex s has at most $a-2$ gray neighbors in $V(K) - S$. Otherwise, s and $a-1$ gray neighbors in $V(K) - S$ will violate both conditions.

The total weight of K is as follows:

- In the CRG induced by the vertex subset $V(K) - S$, the weight is at least $\lceil (n-|S|)(n-|S|-a+1)/2 \rceil$, by induction.
- In the CRG induced by the pair $(S, V(K)-S)$, each $s \in S$ has at least one white neighbor and at most $a-2$ gray neighbors, so the weight from s into $V(K)-S$ is at least $(a-1)+(n-|S|-(a-2)-1) = n-|S|$. So the weight is at least $|S|(n-|S|)$.
- In the CRG induced by S, the total weight is at least $\lceil (|S|/2)(|S|-a+1) \rceil$, by induction.

Adding these together, the proof is complete. □

Remark: Note that equality holds when $S = \emptyset$ (i.e, there is no white edge) and the gray edges form a graph that is either $(a-2)$-regular or has $n-1$ vertices of degree $a-2$ and one vertex of degree $a-3$, depending on divisibility.

5.2.3 Lower bound

Fix $p^* = \frac{a-1}{a+b-1}$. Let K be any CRG for which $K_a + E_b \not\mapsto_c K$. To simplify notation, define K_{W} to be the CRG induced by $\mathrm{VW}(K)$. First we will give a lower bound on $p^*|\mathrm{EW}(K)| + (1-p^*)|\mathrm{EB}(K)|$.

- In the bipartite CRG induced by $(\mathrm{VW}(K), \mathrm{VB}(K))$, all edges must be black, otherwise $K_a + E_b \mapsto_c K$. These edges contribute a weight of $(1-p^*)|\mathrm{VW}(K)||\mathrm{VB}(K)|$.

- In the CRG induced by $\mathrm{VB}(K)$, each set of $b+1$ vertices has at least one black edge in the CRG they induce, otherwise $K_a + E_b \mapsto_c K$. The a-clique maps to one vertex and the b-coclique maps to the remaining b black vertices. By Turán's theorem, these edges contribute a weight of at least $(1-p^*)\left[\binom{|\mathrm{VB}(K)|}{2} - t(|\mathrm{VB}(K)|, b)\right]$.
- In the CRG induced by $\mathrm{VW}(K)$, consider a set of a vertices. If there is neither a white edge nor a spanning subgraph of black edges, then the vertices can be labeled $v_1, \ldots, v_a$ such that the $a-1$ edges incident to v_1 are all gray and $\{v_2, \ldots, v_a\}$ induces a CRG with all edges either gray or black.
  In this case, map the b-coclique to v_1 and the a vertices of the clique to $v_1, \ldots, v_a$. This exhibits the fact that $K_a + E_b \mapsto_c K$. We will apply Lemma 18 to K_{W}. As a result, these edges contribute a weight of at least

$$\begin{aligned} & p^*|\mathrm{EW}(K_{\mathrm{W}})| + (1-p^*)|\mathrm{EB}(K_{\mathrm{W}})| \\ = \; & \frac{1}{a+b-1}\left[(a-1)|\mathrm{EW}(K_{\mathrm{W}})| + b|\mathrm{EB}(K_{\mathrm{W}})|\right] \\ \geq \; & \frac{1}{a+b-1}\left[(a-1)|\mathrm{EW}(K_{\mathrm{W}})| + |\mathrm{EB}(K_{\mathrm{W}})|\right] \\ \geq \; & \frac{1}{a+b-1}\left\lceil \frac{|\mathrm{VW}(K)|}{2}(|\mathrm{VW}(K)| - a + 1)\right\rceil . \end{aligned}$$

The remaining edges of the CRG contribute the following to the weight:

$$\begin{aligned} & (1-p^*)|\mathrm{VW}(K)||\mathrm{VB}(K)| + (1-p^*)\left(\binom{|\mathrm{VB}(K)|}{2} - t\left(|\mathrm{VB}(K)|, b\right)\right) \\ \geq \; & \frac{1}{a+b-1}\left(b|\mathrm{VW}(K)||\mathrm{VB}(K)| + b\left(\binom{|\mathrm{VB}(K)|}{2} - t\left(|\mathrm{VB}(K)|, b\right)\right)\right) \\ \geq \; & \frac{1}{a+b-1}\left(b|\mathrm{VW}(K)||\mathrm{VB}(K)| + \frac{|\mathrm{VB}(K)|^2}{2} - \frac{b|\mathrm{VB}(K)|}{2}\right). \end{aligned}$$

Computing $f_K(p^*)$ gives, by definition,

$$\begin{aligned} f_K(p^*) &= \frac{1}{k^2}\left(p^*\left(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|\right) + (1-p^*)\left(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|\right)\right) \\ &\geq \frac{1}{(a+b-1)k^2}\big((a-1)|\mathrm{VW}(K)| + b|\mathrm{VB}(K)| + 2b|\mathrm{VW}(K)||\mathrm{VB}(K)| \\ &\quad +|\mathrm{VB}(K)|^2 - b|\mathrm{VB}(K)| + |\mathrm{VW}(K)|(|\mathrm{VW}(K)| - a + 1)\big) \\ &= \frac{1}{(a+b-1)k^2}\left(2b|\mathrm{VW}(K)||\mathrm{VB}(K)| + |\mathrm{VB}(K)|^2 + |\mathrm{VW}(K)|^2\right) \\ &= \frac{1}{(a+b-1)k^2}\left(k^2 + 2(b-1)|\mathrm{VW}(K)||\mathrm{VB}(K)|\right) \\ &\geq \frac{1}{a+b-1}. \end{aligned}$$

Therefore, $d^*(\mathrm{Forb}(K_a + E_b)) = \frac{1}{a+b-1}$ and $p^*(\mathrm{Forb}(K_a + E_b)) = \frac{a-1}{a+b-1}$. □

5.3 Edit distance of $K_{3,3}$

5.3.1 Upper bound

The Young tableau in Figure 2 diagrams the values of (a,c) for which $K_{3,3} \not\mapsto_c K(a,c)$ and Figure 3 gives the graph of $K(1,2)$ with the region it defines shaded.

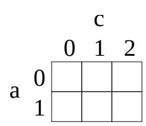

Figure 2: The Young tableau of (a,c) for which $K_{3,3} \not\mapsto_c K(a,c)$.

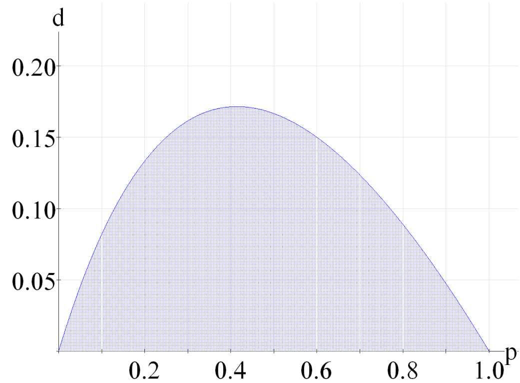

Figure 3: The graph of $g_{K(1,2)}(p)$.

Here, we choose K' to have $|\mathrm{VW}(K')| = 1$, $|\mathrm{VB}(K')| = 2$ and all edges are gray. That is, $K' = K(1,2)$ and it is easy to see that $K_{3,3} \not\mapsto_c K'$. We can use Lemma 10 to compute that $g_{K'}(p) = \frac{p(1-p)}{1+p}$. The maximum of this function on $[0,1]$ occurs at $(p^*, d^*) = \left(\sqrt{2}-1, 3-2\sqrt{2}\right)$.

5.3.2 Lower bound

Fix $p^* = \sqrt{2}-1$. Let K be any CRG for which $K_{3,3} \not\mapsto_c K$. For simplicity of notation, define K_{B} to be the CRG induced by $\mathrm{VB}(K)$. First we will give a lower bound on $p^*|\mathrm{EW}(K)| + (1-p^*)|\mathrm{EB}(K)|$.

- In the CRG induced by $\mathrm{VW}(K)$, all edges must be white, otherwise $K_{3,3} \mapsto_c K$. These edges contribute a weight of $p^*\binom{|\mathrm{VW}(K)|}{2}$.
- In the bipartite CRG induced by $(\mathrm{VW}(K), \mathrm{VB}(K))$, if there is a triangle $\{b_1, b_2, b_3\}$ in $\mathrm{VB}(K)$ that has all edges white or gray then, for every $w \in \mathrm{VW}(K)$, $\{b_i, w\}$ is white for at least one $i \in \{1,2,3\}$. Otherwise, $K_{3,3} \mapsto_c K$.
  Since there must be a white edge between every white/gray triangle in $\mathrm{VB}(K)$ and every vertex in $\mathrm{VW}(K)$, let $C \subseteq \mathrm{VB}(K)$ be a minimum-sized vertex set that contains a vertex from every triangle with no black edges in K_{B}. These edges contribute a weight of at least $p^*|\mathrm{VW}(K)||C|$.
- Let $K_{\mathrm{B}\setminus\mathrm{C}}$ denote the CRG induced by $\mathrm{VB}(K) - C$. In K_{B}, there can be no triangle with all edges gray, otherwise $K_{3,3} \mapsto_c K$. By the definition of C, in $K_{\mathrm{B}\setminus\mathrm{C}}$ there can

be no triangle with all edges white or gray. These edges contribute a weight of at least

$$\min\{p^*, 1-p^*\}\left(\binom{|\mathrm{VB}(K)|}{2} - t(|\mathrm{VB}(K)|, 2)\right)$$
$$+(1-2\min\{p^*, 1-p^*\})\left(\binom{|\mathrm{VB}(K)-C|}{2} - t(|\mathrm{VB}(K)-C|, 2)\right).$$

Since $p^* < 0.5$, $p^*|\mathrm{EW}(K)| + (1-p^*)|\mathrm{EB}(K)|$ is at least

$$p^*\binom{|\mathrm{VW}(K)|}{2} + p^*|\mathrm{VW}(K)||C| + p^*\left(\frac{|\mathrm{VB}(K)|^2 - 2|\mathrm{VB}(K)|}{4}\right)$$
$$+(1-2p^*)\left(\frac{|\mathrm{VB}(K)-C|^2 - 2|\mathrm{VB}(K)-C|}{4}\right).$$

A lower bound on $f_K(p^*)$ gives

$$\begin{aligned}
f_K(p^*)k^2 &= p^*\left(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|\right) + (1-p^*)\left(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|\right)\\
&\geq p^*|\mathrm{VW}(K)| + (1-p^*)|\mathrm{VB}(K)| + 2p^*\binom{|\mathrm{VW}(K)|}{2}\\
&\quad +2p^*|\mathrm{VW}(K)||C| + 2p^*\left(\frac{|\mathrm{VB}(K)|^2 - 2|\mathrm{VB}(K)|}{4}\right)\\
&\quad +2(1-2p^*)\left(\frac{|\mathrm{VB}(K)-C|^2 - 2|\mathrm{VB}(K)-C|}{4}\right)\\
&\geq (1-2p^*)|C| + p^*|\mathrm{VW}(K)|^2 + 2p^*|\mathrm{VW}(K)||C|\\
&\quad +\frac{1-p^*}{2}|\mathrm{VB}(K)|^2 - (1-2p^*)|\mathrm{VB}(K)||C| + \frac{1-2p^*}{2}|C|^2\\
&\geq p^*|\mathrm{VW}(K)|^2 + \frac{1-p^*}{2}|\mathrm{VB}(K)|^2\\
&\quad +\frac{1-2p^*}{2}|C|\left(|C| - 2|\mathrm{VB}(K)| + \frac{4p^*}{1-2p^*}|\mathrm{VW}(K)|\right).
\end{aligned} \tag{3}$$

All that remains is to verify that the expressions in (3) is at most $\frac{p^*(1-p^*)}{1+p^*}k^2$. We need to divide this into two cases. First, assume $|\mathrm{VB}(K)| \leq \frac{2p^*}{1-2p^*}|\mathrm{VW}(K)|$. In this case, the value of $|C|$ that minimizes (3) is $|C| = 0$,

$$\begin{aligned}
f_K(p^*)k^2 &\geq p^*|\mathrm{VW}(K)|^2 + \frac{1-p^*}{2}|\mathrm{VB}(K)|^2\\
&\geq \left(p^*\left(\frac{1-p^*}{1+p^*}\right)^2 + \frac{1-p^*}{2}\left(\frac{2p^*}{1+p^*}\right)^2\right)k^2\\
&= \frac{p^*(1-p^*)}{1+p^*}k^2 = (3-2\sqrt{2})k^2,
\end{aligned}$$

because the minimum occurs at $|\mathrm{VB}(K)| = \frac{2p^*}{1+p^*}k$. Second, assume $|\mathrm{VB}(K)| \geq \frac{2p^*}{1-2p^*}|\mathrm{VW}(K)|$, i.e, $|\mathrm{VB}(K)| \geq 2p^*k$. In this case, the value of $|C|$ that minimizes (3) is $|C| = |\mathrm{VB}(K)| - \frac{2p^*}{1-2p^*}|\mathrm{VW}(K)|$:

$$\begin{aligned} f_K(p^*)k^2 &\geq p^*|\mathrm{VW}(K)|^2 + \frac{1-p^*}{2}|\mathrm{VB}(K)|^2 - \frac{1-2p^*}{2}\left[|\mathrm{VB}(K)| - \frac{2p^*}{1-2p^*}|\mathrm{VW}(K)|\right]^2 \\ &= \left(p^* - \frac{2(p^*)^2}{1-2p^*}\right)|\mathrm{VW}(K)|^2 + \frac{p^*}{2}|\mathrm{VB}(K)|^2 + 2p^*|\mathrm{VW}(K)||\mathrm{VB}(K)| \\ &= p^*k^2 - \frac{2(p^*)^2}{1-2p^*}|\mathrm{VW}(K)|^2 - \frac{p^*}{2}|\mathrm{VB}(K)|^2. \end{aligned}$$

This expression is minimized at the endpoints of the domain of $|\mathrm{VB}(K)|$. For $|\mathrm{VB}(K)| = k$, we have $\frac{p^*}{2} \approx 0.207 > \frac{p^*(1-p^*)}{1+p^*} = 3 - 2\sqrt{2} \approx 0.172$. For the other endpoint, $|\mathrm{VB}(K)| = 2p^*k$, we have

$$f_K(p^*)k^2 \geq p^*k^2 - \frac{2(p^*)^2}{1-2p^*}(1-2p^*)^2k^2 - \frac{p^*}{2}(2p^*)^2k^2 = \left[p^* - 2(p^*)^2 + 2(p^*)^3\right]k^2.$$

This gives $f_K(p) \geq 15\sqrt{2} - 21 \approx 0.213 > \frac{p^*(1-p^*)}{1+p^*} = 3 - 2\sqrt{2} \approx 0.172$.

To summarize, $f(p^*) \geq \frac{p^*(1-p^*)}{1+p^*} = 3 - 2\sqrt{2}$. Therefore, $d^*(\mathrm{Forb}(K_{3,3})) = 3 - 2\sqrt{2}$ and $p^*(\mathrm{Forb}(K_{3,3})) = \sqrt{2} - 1$.

5.4 Edit distance of H_9

5.4.1 Upper bound

The Young tableau in Figure 4 diagrams the values of (a, c) for which $H_9 \not\mapsto_c K(a, c)$. To see this, we can exhibit the following partitions of $V(G)$:

- **3 cliques**: $\{\{0, 1, 2\}, \{3, 4, 5\}, \{6, 7, 8\}\}$
- **1 coclique, 2 cliques**: $\{\{2, 7\}, \{8, 0, 1\}, \{3, 4, 5, 6\}\}$
- **2 cocliques, 1 clique**: $\{\{1, 4, 7\}, \{2, 5, 8\}, \{0, 3, 6\}\}$
- **4 cocliques**: $\{\{1, 4, 7\}, \{0, 5\}, \{3, 8\}, \{2, 6\}\}$.

Moreover, it is easy to see that the largest clique of H_9 is 4 and the largest coclique is 3. So, $H_9 \not\mapsto_c K(0, 2), K(1, 1)$. Since there are only two cocliques of size 3, $H_9 \not\mapsto_c K(3, 0)$.

Figure 5 gives the graphs of $g_{K(0,2)}(p)$, $g_{K(3,0)}(p)$ as well as $g_{K''}(p)$ for the K'' defined in the theorem. The region they define is shaded.

Recall that K'' satisfies $|\mathrm{VW}(K'')| = 4$, $|\mathrm{VB}(K'')| = 0$, one black edge and 5 gray edges. See Figure 6. The graph H_9 has only two cocliques of order three: $\{1, 4, 7\}$ and $\{2, 5, 8\}$. The vertices that remain, $\{0, 3, 6\}$, form a clique. So, any partition of the vertices of H_9 into cocliques that uses both of these 3-cocliques, requires 5 pieces to the

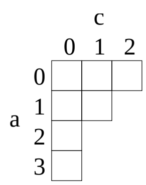

Figure 4: The Young tableau of (a, c) for which $H_9 \not\mapsto_c K(a, c)$.

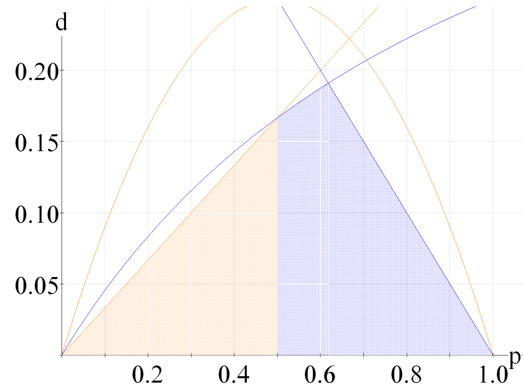

Figure 5: The graphs of the $g_K(p)$ relevant to H_9.

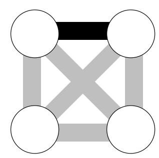

Figure 6: The colored regularity graph K''.

partition. As a result, if there were a colored-homomorphism from H_9 into K'', it would partition the vertices into one coclique of order 3 and three cocliques of order 2.

Assuming such a colored-homomorphism exists, we assume, without loss of generality, that one of the cocliques is $\{1, 4, 7\}$. The vertex 0 is only nonadjacent to 5. The vertex 3 is only nonadjacent to 8. The vertex 6 is only nonadjacent to 2. Therefore, the only partition that can witness the colored-homomorphism is $\{\{1, 4, 7\}, \{0, 5\}, \{2, 6\}, \{3, 8\}\}$. Between every pair of these cocliques is a nonedge. So, no pair of them could be mapped to endpoints of the black edge of K''. Therefore, $H_9 \not\mapsto_c K''$.

We can use Lemma 10 to conclude that $g_{K'}(p) = \frac{1-p}{2}$ and $g_{K''}(p) = \frac{p}{2(1+p)}$. The intersection is at the point $(p, d) = \left(\frac{\sqrt{5}-1}{2}, \frac{3-\sqrt{5}}{4}\right)$.

Thus, the upper bound obtained by $d^* \leq \max_{p\in[0,1]} \min_{(a,c):K(a,c)\in\mathcal{K}(\mathcal{H})} g_{K(a,c)}(p)$ would be $1/5 = 0.2$, achieved by the intersection of $g_{K(0,2)}(p) = \frac{1-p}{2}$ and $g_{K(3,0)}(p) = \frac{p}{3}$. But, as we can see by Figure 5, the value $d = \frac{3-\sqrt{5}}{4} \approx 0.191$ provides a better upper bound.

5.5 New proof of the edit distance of $\overline{P_3 + K_1}$

Alon and Stav [2] computed (p^*, d^*) for hereditary properties defined by graphs on at most 4 vertices. This paper has also done so, as a corollary of our results, with the exception of Forb $\left(\overline{P_3 + K_1}\right)$. We include a computation of the value of (p^*, d^*) as an application of our technique and to demonstrate the versatility of Lemma 18.

5.5.1 Upper bound

Here, we choose $K' = K(0,1)$ and $K'' = K(2,0)$. Recall that $\overline{P_3 + K_1}$ is a triangle with a pendant edge. It is easy to see that both $\overline{P_3 + K_1} \not\mapsto_c K'$ and $\overline{P_3 + K_1} \not\mapsto_c K''$. We can use Lemma 10 to conclude that $g_{K'}(p) = 1 - p$ and $g_{K''}(p) = \frac{p}{2}$. The intersection is at the point $(p^*, d^*) = \left(\frac{2}{3}, \frac{1}{3}\right)$. Clearly, this p^* is unique because $g(p) < 1/3$ for all $p \neq 2/3$.

5.5.2 Lower bound

Fix $p^* = \frac{2}{3}$. Let K be any CRG for which $\overline{P_3 + K_1} \not\mapsto_c K$. For simplicity of notation, define K_{W} to be the CRG induced by $\mathrm{VW}(K)$. We will give a lower bound on $p^*|\mathrm{EW}(K)| + (1 - p^*)|\mathrm{EB}(K)|$.

- In the bipartite CRG induced by $(\mathrm{VW}(K), \mathrm{VB}(K))$, no edges can be gray, otherwise $\overline{P_3 + K_1} \mapsto_c K$. This contributes $\min\{p^*, 1 - p^*\}|\mathrm{VW}(K)||\mathrm{VB}(K)|$.
- In the CRG induced by $\mathrm{VB}(K)$, no edges can be gray, otherwise $\overline{P_3 + K_1} \mapsto_c K$. This contributes $\min\{p^*, 1 - p^*\}\binom{|\mathrm{VB}(K)|}{2}$.
- In K_{W}, consider any subset of 3 vertices. If there is neither a white edge nor a pair of black edges, then it is possible to map the vertices of $\overline{P_3 + K_1}$ into those three vertices so that each vertex of the triangle is mapped to a different vertex and the pendant vertex is in the vertex incident to the two gray edges. This contributes $p^*\mathrm{EW}(K_{\mathrm{W}}) + (1 - p^*)\mathrm{EB}(K_{\mathrm{W}})$.

To summarize, using $p^* = 2/3$ and Lemma 18, we have

$$\begin{aligned}
&p^*|\mathrm{EW}(K)| + (1 - p^*)|\mathrm{EB}(K)| \\
&\geq \frac{1}{3}\left(|\mathrm{VW}(K)||\mathrm{VB}(K)| + \binom{|\mathrm{VB}(K)|}{2}\right) + \frac{1}{3}\left(2|\mathrm{EW}(K_{\mathrm{W}})| + |\mathrm{EB}(K_{\mathrm{W}})|\right) \\
&\geq \frac{1}{3}\left(|\mathrm{VW}(K)||\mathrm{VB}(K)| + \binom{|\mathrm{VB}(K)|}{2}\right) + \frac{1}{3}\left\lceil \frac{|\mathrm{VW}(K)|}{2}(|\mathrm{VW}(K)| - 2)\right\rceil \\
&= \frac{1}{3}\left(|\mathrm{VW}(K)||\mathrm{VB}(K)| + \binom{|\mathrm{VB}(K)|}{2} + \frac{|\mathrm{VW}(K)|^2 - 2|\mathrm{VW}(K)|}{2}\right).
\end{aligned}$$

Now we give a lower bound on $f_K(p^*)$:

$$\begin{aligned}
f_K(p^*)k^2 &= p^*\left(|\mathrm{VW}(K)| + 2|\mathrm{EW}(K)|\right) + (1 - p^*)\left(|\mathrm{VB}(K)| + 2|\mathrm{EB}(K)|\right) \\
&\geq \frac{2}{3}|\mathrm{VW}(K)| + \frac{1}{3}|\mathrm{VB}(K)| + \frac{2}{3}\Big(|\mathrm{VW}(K)||\mathrm{VB}(K)| \\
&\quad + \frac{|\mathrm{VB}(K)|^2 - |\mathrm{VB}(K)|}{2} + \frac{|\mathrm{VW}(K)|^2 - 2|\mathrm{VW}(K)|}{2}\Big) \\
&= \frac{1}{3}\left(|\mathrm{VW}(K)|^2 + 2|\mathrm{VW}(K)||\mathrm{VB}(K)| + |\mathrm{VB}(K)|^2\right) = \frac{1}{3}k^2.
\end{aligned}$$

So, comparing with the upper bound, $d^*(\mathrm{Forb}(\overline{P_3 + K_1})) = 1/3$ and since $g(p) \leq \min\{g_{K'}(p), g_{K''}(p)\} < 1/3$ for all $p \neq 2/3$, it is also the case that $p^*(\mathrm{Forb}(\overline{P_3 + K_1})) = 2/3$. □

6 Conclusions

6.1 Observations on functions f and g

The function $f_K(p)$ is invariant under equipartitions of $V(K)$. To see this, let $\tilde{K}$ be formed by partitioning each vertex of K into c pieces with the colors of vertices and edges be the natural coloring inherited from K. As a result, $|\mathrm{VW}(\tilde{K})| = c|\mathrm{VW}(K)|$, $|\mathrm{VB}(\tilde{K})| = c|\mathrm{VB}(K)|$, $|\mathrm{EW}(\tilde{K})| = c^2|\mathrm{EW}(K)| + \binom{c}{2}|\mathrm{VW}(K)|$ and $|\mathrm{EB}(\tilde{K})| = c^2|\mathrm{EB}(K)| + \binom{c}{2}|\mathrm{VB}(K)|$. Thus,

$$f_{\tilde{K}}(p) = \frac{1}{c^2k^2}\left[p(|\mathrm{VB}(\tilde{K})| + 2|\mathrm{EW}(\tilde{K})|) + (1-p)(|\mathrm{VW}(\tilde{K})| + 2|\mathrm{EB}(\tilde{K})|)\right] = f_K(p).$$

The same is true for $g_K(p)$ and $g_{\tilde{K}}(p)$. Any feasible solution, $\mathbf{u}$ of the quadratic program that defines $g_K(p)$ can be made into a feasible solution, $\tilde{\mathbf{u}}$ of the quadratic program that defines $g_{\tilde{K}}(p)$ by arbitrarily distributing the weight of one vertex in K to the vertices in $\tilde{K}$ to which it corresponds. It can be seen that, if $\mathbf{M}_K(p)$ is the matrix corresponding to K and $\mathbf{M}_{\tilde{K}}(p)$ is the matrix corresponding to $\tilde{K}$, then $\mathbf{u}^T\mathbf{M}_K(p)\mathbf{u} = \tilde{\mathbf{u}}^T\mathbf{M}_{\tilde{K}}(p)\tilde{\mathbf{u}}$.

The function g is more flexible, however. It is not only invariant under equipartitions but it is invariant under arbitrary partitions. To see this, construct an equivalence relation on the vertices of a CRG, K, in which vertices u and v are equivalent if u, v and $\{u,v\}$ are all the same color and, for all $w \in V(K) - \{u,v\}$, $\{u,w\}$ and $\{v,w\}$ are the same color as each other.

If K is a CRG and K_0 is the CRG induced by the equivalence relation on K, then $g_K(p) = g_{K_0}(p)$. Therefore, in the computation of $g(p)$, one may ignore CRGs which have nontrivial equivalence classes.

6.2 Open questions

- Investigating Proposition 17, is there a more convenient expression for the upper bound based only on the Young diagram (see Figures 2 and 4) of the set of CRGs $\{K(a,c) : H \not\mapsto_c K(a,c), \forall H \in \mathcal{F}(\mathcal{H})\}$?
- To compute the edit distance is hard. We do not even have a sharp result for $\mathrm{Forb}(K_{m,n})$, where $K_{m,n}$ is an arbitrary complete bipartite graph.
- The precise value for $d^*(H_9)$, where H_9 is defined in Theorem 14, is unknown, but we conjecture that the upper bound is correct and we further conjecture that $p^*(\mathrm{Forb}(H_9)) = (\sqrt{5}-1)/2$.
- Every hereditary property $\mathcal{H}$ can be expressed as $\bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$ for some family of graphs $\mathcal{F}(\mathcal{H})$. In computing edit distance, it may be that some members of $H \in \mathcal{F}(\mathcal{H})$ are unnecessary, even if they are necessary to define the family. I.e, for hereditary property $\mathcal{H}$, what are the maximal properties $\mathcal{H}' \supseteq \mathcal{H}$ such that $d^*(\mathcal{H}') = d^*(\mathcal{H})$?

  For example, the strong perfect graph theorem [11] states that perfect graphs are characterized by $\mathcal{P} = \bigcap_{k\geq 2}\left(\mathrm{Forb}(C_{2k+1}) \cap \mathrm{Forb}(\overline{C_{2k+1}})\right)$. But, it is not difficult to use Theorem 16 to show that $d^*(\mathcal{P}) = d^*(\mathrm{Forb}(C_5)) = 1/4$, as observed by Alon and Stav [1].

- Our proofs of the lower bounds for $d^*(\mathrm{Forb}(H))$ for $H = K_a + E_b$ or $H = K_{3,3}$ are cumbersome and cannot assume that the total number of vertices in each of the forbidden CRGs is bounded by any function of H. Is there a better way to compute the lower bound? Is there a function of H so that we need only to consider $g_K(p)$ for K whose order is bounded by said function?
- Finally, the edit distance is unknown for most hereditary properties. So-called unit disk graphs (UDGs), see [12], define a hereditary property but the family $\mathcal{F}$ is not known. It is easy to see that $K_{1,7}$ cannot occur as an induced subgraph in a UDG. (For some definitions of the unit disk graph, $K_{1,6}$ is forbidden also.) We believe that a small family of such forbidden induced subgraphs will be enough to determine the edit distance from the family of UDGs.

  For any graph H, both $p^*(\mathrm{Forb}(H))$ and $d^*(\mathrm{Forb}(H))$ can be considered invariants of graph H. Being able to compute these invariants even for some given fixed graph seems to be quite difficult in general.

6.3 Thanks

We thank Maria Axenovich for valuable conversations. We also thank Noga Alon for reading a draft of the paper and making several useful comments. These include the observation that the existence of the limit that defines $d^*(\mathcal{H})$ results from a simple monotonicity argument. We appreciate the careful reading of an anonymous referee, which lead to improvements in the manuscript.

References

[1] N. Alon and U. Stav, What is the furthest graph from a hereditary property?, to appear in *Random Structures Algorithms.*

[2] N. Alon and U. Stav, The maximum edit distance from hereditary graph properties, to appear in *J. Combin. Theory Ser. B.*

[3] M. Axenovich and R. Martin, Avoiding patterns in matrices via a small number of changes, *SIAM J. Discrete Math.*, **20** (2006), no. 1, 49–54.

[4] M. Axenovich, A. Kézdy and R. Martin, On the editing distance in graphs, to appear in *J. Graph Theory.*

[5] B. Bollobás, Hereditary properties of graphs: asymptotic enumeration, global structure, and colouring. Proceedings of the international Congress of Mathematicians, Vol. III (Berlin 1998). *Doc. Math.* 1998, Extra Vol. III, 333–342 (electronic).

[6] B. Bollobás, *Modern graph theory*, Graduate texts in mathematics, **184**. *Springer-Verlag, New York*, 1998. xiv+394 pp.

[7] B. Bollobás, *Random graphs.* Second edition. Cambridge Studies in Advanced Mathematics, **73**. *Cambridge University Press, Cambridge*, 2001. xviii+498 pp.

[8] B. Bollobás and A. Thomason, Projections of bodies and hereditary properties of hypergraphs. *Bull. London Math. Soc.* **27** (1995), no. 5, 417–424.

[9] B. Bollobás and A. Thomason, Hereditary and monotone properties of graphs. *The mathematics of Paul Erdős, II* 70–78, Algorithms Combin., 14, Springer, Berlin, 1997.

[10] B. Bollobás and A. Thomason, The structure of hereditary properties and colourings of random graphs. *Combinatorica* **20** (2000), no. 2, 173–202.

[11] M. Chudnovsky, N. Robertson, P. Seymour and R. Thomas, The strong perfect graph theorem. *Ann. of Math. (2)* **164** (2006), no. 1, 51–229.

[12] B.N. Clark, C.J. Colbourn and D.S. Johnson, Unit disk graphs. *Discrete Math.* **86** (1990), no. 13, 165-177.

[13] P. Erdős and A. Rényi, On random graphs I. *Publ. Math. Debrecen* **6** (1959), 290–297.

[14] P. Erdős and M. Simonovits, A limit theorem in graph theory, *Studia Sci. Math. Hungar* **1** (1966), 51–57.

[15] P. Erdős and A. Stone, On the structure of linear graphs, *Bull. Amer. Math. Soc.* **52** (1946), 1087–1091.

[16] S. Janson, T. Łuczak and A Ruciński, *Random Graphs.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[17] J. Komlós, A. Shokoufandeh, M. Simonovits and E. Szemerédi, The regularity lemma and its applications in graph theory. *Theoretical aspects of computer science (Tehran, 2000)*, 84–112, Lecture Notes in Comput. Sci., **2292**, Springer, Berlin, 2002.

[18] J. Komlós and M. Simonovits, Szemerédi's regularity lemma and its applications in graph theory. *Combinatorics, Paul Erdős is eighty, Vol. 2* (Keszthely, 1993), 295–352, Bolyai Soc. Math. Stud. 2, János Bolyai Math. Soc., Budapest, 1996.

[19] H.J. Prömel and A. Steger, Excluding induced subgraphs: quadrilaterals. *Random Structures Algorithms* **2** (1991), no. 1, 55–71.

[20] H.J. Prömel and A. Steger, Excluding induced subgraphs. II. Extremal graphs. *Discrete Appl. Math.* **44** (1993), no. 1-3, 283–294.

[21] H.J. Prömel and A. Steger, Excluding induced subgraphs. III. A general asymptotic. *Random Structures Algorithms* **3** (1992), no. 1, 19–31.

[22] W. Rudin, *Principles of mathematical analysis.* Third edition. International Series in Pure and Applied Mathematics. *McGraw-Hill Book Co., New York-Auckland-Düsseldorf*, 1976. x+342 pp.

[23] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression, *Acta Arithmetica* **27** (1975), 199–245.

[24] E. Szemerédi, Regular partitions of graphs. In *Problèmes Combinatoires et Théorie des Graphes*, 399–401, Colloq. Internat. CNRS, Univ. Orsay, Paris, 1978.

[25] P. Turán, Eine Extremalaufgabe aus der Graphentheorie. (Hungarian) *Mat. Fiz. Lapok* **48**, (1941), 436–452.