

# Editing distance of graphs

Maria Axenovich\*      André Kézdy†      Ryan Martin‡

December 17, 2003

## Abstract

Given a family of graphs  $\mathcal{G}$ , the *editing distance* of a graph  $G$  from  $\mathcal{G}$  is the number of combined deletions and additions of edges so that the resulting graph is in  $\mathcal{G}$ . In this paper, we fix a graph  $H$  and consider  $\text{Forb}(H)$ , the set of graphs that have no induced copy of  $H$ . We provide bounds for the maximum editing distance among all  $n$ -vertex graphs from  $\mathcal{G} = \text{Forb}(H)$ . In doing so, we also introduce a graph invariant which we call the *binary chromatic number*.

## 1 Introduction

The investigation of graphs not containing subgraphs with given properties is a classical problem. For example, determining the maximal number of edges in a graph with no copy of a fixed subgraph  $H$  has been studied intensively for the last 70 years [9], [14]. Very often, though, the desired task is not to determine the extremal graph without a given fixed subgraph, but rather starting with an arbitrary graph, modify it in a small number of steps such that the resulting graph does not contain a forbidden subgraph.

The problem of modifying the given graph such that the resulting graph satisfies some *global properties* has been addressed by Erdős et al.[6], [4]. Namely, it was determined how many edges is sufficient to delete from an arbitrary triangle-free graph to obtain a bipartite graph and how many edges always suffice to add to a graph to decrease its diameter.

In this paper, we investigate the graph-modification problem such that the resulting graph has a *local property* of not having a fixed induced subgraph. Starting with an arbitrary graph  $G$ , we would like to calculate the minimal number of edges needed to be added or deleted from  $G$  to obtain a graph not containing a fixed induced subgraph. Formally, let the  $\text{Dist}(G, H)$  be exactly the minimal number of edge-additions and edge-deletions necessary to perform on a graph  $G$  to obtain a graph isomorphic to  $H$ . Now, if  $\mathcal{H}$  is a class of graphs on  $n$  vertices, we define  $\text{Dist}(G, \mathcal{H}) = \min\{\text{Dist}(G, H) : H \in \mathcal{H}\}$  for a graph  $G$  on  $n$  vertices. Finally,  $\text{Dist}(n, \mathcal{H}) = \min\{\text{Dist}(G, \mathcal{H}) : |V(G)| = n\}$ . We call such distance the **editing distance** since the operations performed can be considered as editing the edge set of a graph. Our

---

\*Department of Mathematics, Iowa State University, Ames, IA 50011, axenovic@math.iastate.edu

†Department of Mathematics, University of Louisville, Louisville, KY 40292, kezdy@louisville.edu

‡Department of Mathematics, Iowa State University, Ames, IA 50011, rymartin@iastate.edu

interest here is the class  $\mathcal{H}$  of graphs on  $n$  vertices containing no copies of a given fixed graph  $H$  as an induced subgraph. We denote this class by  $\text{Forb}(n, H)$  (or simply by  $\text{Forb}(H)$  when it is clear from the context). Similarly  $\text{Forb}'(H)$  is a family of all graphs on  $n$  vertices with no subgraph isomorphic to  $H$ .

This problem has numerous applications in computer science and bioinformatics. For example, consider a metabolic network and identify genes with vertices of a graph and pairs of interacting genes with edges of a graph. Then it is a fundamental question in biology (from evolutionary and practical points of view) to find how many edge-changes in such a graph shall be performed to avoid an induced subgraph corresponding to a certain metabolic process. Another example is concerned with consensus trees. It is known that two consensus trees are comparable if there is no induced path on five vertices in a corresponding bipartite graph [15, 2, 3]. In particular, finding the smallest number of edge-changes in such a graph will determine the distance between these trees.

On the other hand, the editing problem of graphs corresponds to determining the distance between  $\{0, 1\}$ -matrices. If  $A, B$  are adjacency matrices of graphs  $G$  and  $H$  respectively, then  $\text{Dist}(G, H)$  corresponds to the number of positions where  $A$  and  $B$  differ, i.e., to the Hamming distance between  $A$  and  $B$ . Thus finding editing distance between classes of graphs provides the Hamming distance between classes of symmetric matrices with the same diagonal entries. Moreover, when the graph editing problem is restricted to bipartite graphs in which edge additions and deletions are limited to edges between partite sets, it corresponds to the problem of determining the distance between the sets of arbitrary  $\{0, 1\}$ -matrices.

Define the distance  $\text{Dist}'(n, \text{Forb}'(H))$  to be analogous to  $\text{Dist}(n, \text{Forb}(H))$ , but in this case, only permit edge-deletions. This quantity will always be equal to  $\text{Dist}(K_n, \text{Forb}'(H))$ , i.e., it is the minimal number of edges in a complement of an  $H$ -free graph. Let  $\text{ex}(n, H)$  be the maximal number of edges in a graph on  $n$  vertices with no subgraph isomorphic to  $H$ . Then

$$\text{Dist}(K_n, \text{Forb}'(H)) = \binom{n}{2} - \text{ex}(n, H). \quad (1)$$

The asymptotic behavior of  $\text{ex}(n, H)$  is provided by the following.

**Theorem 1.1 (Erdős, Stone [8])**  $\text{ex}(n, H) = (1 + o(1)) \binom{n}{2} \left(1 - \frac{1}{\chi(H)-1}\right)$ .

Thus, in particular, the distance  $\text{Dist}(n, \text{Forb}'(H))$  is asymptotically determined by the chromatic number of a graph  $H$  when it is at least 3.

Clearly, when a forbidden graph is complete or empty, finding the editing distance becomes a trivial task immediately reduced to Turán's theorem. On the other hand, perhaps the most interesting case is when the forbidden induced subgraph is self-complementary, i.e., when both operations of edge-deletions and edge-additions carry "an equal power". In this case, we derive asymptotically tight estimates for  $\text{Dist}(n, \text{Forb}(H))$ . We also give general bounds for other graphs. Our main tool in providing the lower bounds is Szemerédi's regularity lemma which allows us to express our bounds in terms of the *binary chromatic number* which we now define.

**Definition 1.2** *The binary chromatic number,  $\chi_B(G)$  is the least integer  $k + 1$  such that, for each  $c \in \{0, \dots, k + 1\}$ , there exists a partition of  $V(G)$  into  $c$  cliques and  $k + 1 - c$  cocliques.*

Next we list the main results of this paper.

**Theorem 1.3** *Let  $H$  be a graph with binary chromatic number  $k + 1$ , then*

$$\text{Dist}(n, \text{Forb}(H)) > (1 - o(1)) \frac{n^2}{4k}.$$

If  $k = \chi_B(G) - 1$ , then let  $c_{\min}$  be the least  $c$  so that  $G$  cannot be partitioned into  $c_{\min}$  cliques and  $k - c_{\min}$  cocliques. Let  $c_{\max}$  be the greatest such number. We now have an upper bound that can be expressed in terms of the binary chromatic number of  $H$  and corresponding  $c_{\min}$  and  $c_{\max}$ .

**Theorem 1.4** *Let  $H$  be a graph with binary chromatic number  $k + 1$ . Let  $c_{\min}$  and  $c_{\max}$  be defined as above. If  $c_{\min} \leq k/2 \leq c_{\max}$ , then*

$$\text{Dist}(n, \text{Forb}(H)) \leq \frac{1}{2k} \binom{n}{2}. \quad (2)$$

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

$$\text{Dist}(n, \text{Forb}(H)) \leq \left( \frac{1}{1 + 2\sqrt{\frac{c_0}{k} \left(1 - \frac{c_0}{k}\right)}} \right) \frac{1}{k} \binom{n}{2} \leq \frac{1}{k} \binom{n}{2}. \quad (3)$$

Using Theorem 1.3 and the first inequality in Theorem 1.4, we have the following asymptotically tight result:

**Corollary 1.5** *Let  $H$  be a self-complementary graph with the property that  $\chi_B(H) = k + 1$ . Then, we have*

$$\text{Dist}(n, \text{Forb}(H)) = (1 + o(1)) \frac{n^2}{4k}.$$

We need the following definitions and notations. We denote by  $K_n, E_n, C_n, P_n$  a complete graph, an empty graph, a cycle, a path on  $n$  vertices respectively. We also define  $K_p^q$  to be a complete  $p$ -partite graph with each partite set of cardinality  $q$ . We use  $\overline{G}$  to denote the complement of  $G$ . The term *coclique* will often be used in place of *independent set*. For the other definitions, we refer the reader to [17]. For two constants  $a$  and  $b$  we write that  $a \ll b$  if  $a$  is less than some function of  $b$ . This function will be defined in each specific proof. We also need the following definitions.

**Definition 1.6** *Let  $G = (A, B; E)$  be a bipartite graph. The **density** of  $(A, B)$  is denoted  $d(A, B)$  and is given by the formula*

$$d(A, B) = \frac{e(A, B)}{|A||B|}$$

where  $e(A, B) = |E|$  is the number of edges in  $G$ .

**Definition 1.7** *The bipartite graph  $G = (A, B; E)$  is  $\epsilon$ -regular if*

$$X \subset A, Y \subset B, |X| > \epsilon|A|, |Y| > \epsilon|B|$$

*imply*

$$|d(X, Y) - d(A, B)| < \epsilon,$$

*otherwise  $G$  is  $\epsilon$ -irregular.*

In the next section we state two main theorems implying the main lower bound on the editing distance and prove preliminary lemmas. Section 2 contains proofs of the main theorems. We investigate the properties of the important parameter, binary chromatic number, in Section 3. Finally, Section 4 gives several exact results.

## 2 The proofs

The proof of the main theorem, Theorem 1.3, will follow from two following theorems which basically assert that a random graph in which the edges are chosen independently with equal probability  $1/2$  requires many editing operations to “get rid” of all induced copies of  $H$ . This happens since a random graph always has a “uniform” subgraph with desired properties.

In order to state our lower bound, we need to generalize the idea of an  $\epsilon$ -regular pair.

**Definition 2.1** *An  $\epsilon$ -regular  $r$ -tuple is an  $r$ -partite graph with partite sets  $V_1, \dots, V_r$  such that  $(V_i, V_j)$  is an  $\epsilon$ -regular pair for all  $i, j, 1 \leq i < j \leq r$ .*

*We say that an  $\epsilon$ -regular  $r$ -tuple is of size  $rL$  if  $|V_1| = \dots = |V_r| = L$  and an  $\epsilon$ -regular  $r$ -tuple has  $\delta$ -bounded density if  $d(V_i, V_j) \in (\delta, 1 - \delta)$  whenever  $1 \leq i < j \leq r$ .*

*For convenience, we define an  $(\epsilon, r, L, \delta)$ -configuration to be an  $\epsilon$ -regular  $r$ -tuple of size  $rL$  that has  $\delta$ -bounded density.*

**Theorem 2.2** *Let  $r$  be a positive integer,  $\delta, 0 < \delta < 1$ , and  $\epsilon > 0$ , be given such that  $\epsilon \ll \delta$ . There is a graph  $G$  on  $n$  vertices and a constant  $M(\epsilon)$  such that if the number of edge-deletions and edge-additions performed on  $G$  is less than  $\frac{n^2}{4(r-1)}(1-2\delta)(1-\epsilon)^2$  then the resulting graph contains an  $(\epsilon, r, L, \delta)$ -configuration with  $L \geq n(1-\epsilon)/M(\epsilon)$ .*

**Theorem 2.3** *Let  $H$  be a graph with binary chromatic number  $r$ . Let another graph  $G$  have an  $(\epsilon, r, L, \delta)$ -configuration with  $0 < \epsilon \ll \delta < 1$  and  $L \geq L_0(r, \delta, \epsilon)$ . Then there is a copy  $H'$  of  $H$  in  $G$  such that the vertices of  $H'$  in each  $V_i$  of the configuration induce either a clique or a coclique.*

### 2.1 Preliminary lemmas

Our lower bound will make use of the Regularity Lemma (see [12] and [11]). We state it below for completeness.

**Lemma 2.4 (Regularity Lemma [16])** *For every positive  $\epsilon$  and positive integer  $m$ , there are positive integers  $M = M(\epsilon, m)$  and  $N = N(\epsilon, m)$  with the following property: For every graph  $G$  with at least  $N$  vertices there is a partition of the vertex set into  $\ell + 1$  classes (clusters)*

$$V = V_0 + V_1 + V_2 + \cdots + V_\ell$$

such that

1.  $m \leq \ell \leq M$ ,
2.  $|V_1| = |V_2| = \cdots = |V_\ell|$ ,
3.  $|V_0| < \epsilon n$ ,
4. at most  $\epsilon \ell^2$  of the pairs  $(V_i, V_j)$  are  $\epsilon$ -irregular.

The **reduced graph of  $G$**  is defined to be the graph on the vertex set  $\{1, \dots, \ell\}$  such that  $i \sim j$  if and only if  $(V_i, V_j)$  is an  $\epsilon$ -regular pair.

We give a name to the partition given by Lemma 2.4:

**Definition 2.5** *An  $(m, \epsilon, \ell)$ -equipartition of a vertex set  $V$  is a partition  $V = V_0 + V_1 + \dots + V_\ell$  such that the Regularity Lemma's conditions (1), (2) and (3) are satisfied (with  $M = M(\epsilon, m)$  as defined by the Lemma).*

Lemma 2.6 below establishes that a random graph behaves as expected with respect to **any** equipartition. Let  $G(n, p)$  be the archetypical random graph in which each edge from  $K_n$  is chosen independently with probability  $p$  [10]. For functions  $f = f(n)$  and  $g = g(n)$ , let  $f = \omega(g)$  be the usual asymptotic notation denoting that  $g/f \rightarrow 0$  as  $n \rightarrow \infty$ .

**Lemma 2.6** *Fix a constant  $\epsilon > 0$  and positive integer  $m$ . Let  $G = G(n, 1/2)$  and  $f(n) = \omega(n^{-1/2})$ . Let  $P$  be the probability that for each  $(m, \epsilon)$ -equipartition, all pairs of clusters  $(V_i, V_j)$ ,  $1 \leq i < j \leq \ell$ , have density in the interval  $(1/2 - f(n), 1/2 + f(n))$ . Then  $P$  approaches 1 as  $n$  goes to infinity.*

*Proof.* We just want to compute the probability that **all** pairs of disjoint sets, each of size at least  $\epsilon' n$  (where  $\epsilon' = \frac{1-\epsilon}{M(\epsilon)-1}$ ), have density  $1/2 \pm f$  for any  $f = f(n) = \omega(n^{-1/2})$ .

$$\begin{aligned} & \Pr \left\{ \bigvee_{\substack{S, T \subseteq V(G) \\ S \cap T = \emptyset \\ |S|, |T| \geq \epsilon' n}} (d(S, T) \notin (1/2 - f, 1/2 + f)) \right\} \\ & \leq 2^n 2^{2n} \Pr \{d(S, T) < 1/2 - f\} \\ & \leq 2 \cdot 4^n \exp(-2(f|S||T|)^2/(|S||T|)) \\ & \leq 2 \cdot 4^n \exp(-2f^2|S||T|) \\ & \leq 2 \cdot 4^n \exp(-2(\epsilon')^2 f^2 n^2) \rightarrow 0 \end{aligned} \tag{4}$$

Chernoff's bound (see [10]) is used to achieve inequality (4). ■

An immediate result is that there is a graph such that *every* pair of sufficiently large sets has density almost 1/2:

**Corollary 2.7** *Let  $f = f(n)$  be as in Lemma 2.6. Then there is a graph  $G$  on  $n$  vertices so that for any  $(m, \epsilon, \ell)$ -equipartition, all pairs of clusters  $(V_i, V_j)$ ,  $1 \leq i < j \leq \ell$ , are  $\epsilon$ -regular with density in the interval  $(1/2 - f, 1/2 + f)$ .*

The following lemma can be found, in a different form in the survey [12]. In that survey, it is the Key Lemma on page 11:

**Lemma 2.8 (Embedding Lemma)** *For any positive integer  $m$  and positive real constants  $\epsilon < d$ , there is a positive constant  $\epsilon_0 = (d - \epsilon)^{m-1}/(m+1)$  such that, if  $G$  is a graph containing  $m$  disjoint  $n$ -sets of vertices  $S_1, \dots, S_m$  that are pairwise  $\epsilon$ -regular with  $\epsilon \leq \epsilon_0$  and density at least  $d$  [at most  $d$ ], then  $G$  contains  $(\epsilon_0 n)^m$  copies of  $K_m$  [ $E_m$ ] with each vertex from a different  $S_i$ .*

It should be noted that the original statement of the Embedding Lemma does not ensure that each vertex is from a different set, but the proof does. In the case of  $K_m$ , this is a triviality, since we must assume no edges in any of the  $S(i)$ .

Fact 2.9 is a generalization of an intersection property defined in [12].

**Fact 2.9 (Revised Intersection Property)** *Let  $(A, B)$  be an  $\epsilon$ -regular pair with density  $d$ . Let  $Y \subset B$ . Let  $\delta = \min\{d, 1 - d\}$ . Let us be given that  $(\delta - \epsilon)^{\ell-1}|Y| > \epsilon|B|$ . Then,*

$$\# \left\{ (a_1, \dots, a_\ell) \in A^\ell : \left| Y \cap \bigcap_{i=1}^k \overline{N(a_i)} \cap \bigcap_{j=k+1}^{\ell} N(a_j) \right| < (1 - d - \epsilon)^k (d - \epsilon)^{\ell-k} |Y| \right\} \leq \ell \epsilon |A|^\ell, \quad (5)$$

and

$$\# \left\{ (a_1, \dots, a_\ell) \in A^\ell : \left| Y \cap \bigcap_{i=1}^k \overline{N(a_i)} \cap \bigcap_{j=k+1}^{\ell} N(a_j) \right| > (1 - d + \epsilon)^k (d + \epsilon)^{\ell-k} |Y| \right\} \leq \ell \epsilon |A|^\ell. \quad (6)$$

*Proof.* We proceed by induction on  $k$ . If  $k = 0$ , then this is the typical intersection property (see [12]). For  $k > 0$  there are at most  $L \times (\ell - 1) \epsilon |A|^\ell$  tuples  $(a_1, a_2, \dots, a_k, a_{k+1}, \dots, a_\ell)$  such tuples for which  $|Y'| < (1 - d - \epsilon)^{k-1} (d - \epsilon)^{k-\ell} |Y|$  where  $Y' = Y \cap \bigcap_{i=2}^k \overline{N(a_i)} \cap \bigcap_{j=k+1}^{\ell} N(a_j)$ .

Otherwise,  $|Y'| > \epsilon |B|$  and by the definition of  $\epsilon$ -regularity, there are at most  $\epsilon |A| \times |A|^{\ell-1}$  tuples  $(a_1, a_2, \dots, a_k, a_{k+1}, \dots, a_\ell)$  for which  $|Y' \cap \overline{N(a_1)}| < (1 - d - \epsilon) |Y'|$ . This gives (5).

The inequality (6) follows by taking the complement and exchanging the roles of  $k$  and  $\ell - k$ . ■

Let  $R(m, m)$  be the classical Ramsey number – the least integer  $r$  so that any two-coloring of the edges of  $K_r$  in red and blue has either a red clique on  $m$  vertices or a blue clique on  $m$  vertices.

**Lemma 2.10** *Fix a positive integer  $m$  and  $0 < \epsilon < 1/R(m, m)$ . Then, there exists a  $R = R_\epsilon(m)$  such that the following is true for any graph on  $n \geq R$  vertices: If the edges of  $K_n$  are colored red, white and blue, with the number of white edges at most  $\epsilon n^2$ , then there exists either a red  $K_m$  or a blue  $K_m$ .*

*Proof.* According to Turán’s theorem, if a graph on  $n$  vertices has fewer than  $\frac{n}{2} \left( \frac{n}{r-1} - 1 \right)$  nonedges, then it must contain a clique on  $r$  vertices.

Hence, if  $\epsilon n^2 < \frac{n}{2} \left( \frac{n}{r-1} - 1 \right)$  and  $r = R(m, m)$  then the coloring we impose will ensure either a blue clique on  $m$  vertices or a red clique on  $m$  vertices. ■

**Lemma 2.11** *Let  $G$  be a graph on  $L$  vertices and let  $M(\epsilon, m)$  be the constant given by the Regularity Lemma when applied with parameters  $\epsilon$  and  $m$ . Also, let  $n$  be a fixed, positive integer. Then  $G$  has  $n$  disjoint subsets of vertices of equal size,  $L' \geq L/M(\epsilon, R_\epsilon(n))$ , such that either*

1. all pairs are  $\epsilon$ -regular with density at least  $1/2$ , or
2. all pairs are  $\epsilon$ -regular with density less than  $1/2$ .

The constant  $R_\epsilon(n)$  is the parameter given in Lemma 2.10.

*Proof.* Simply apply the Regularity Lemma with  $\epsilon$  and  $m = R_\epsilon(n)$ , for the parameter given in Lemma 2.10. This insures  $\ell \geq R_\epsilon(n)$  clusters of equal size such that at most  $\epsilon \ell^2$  pairs are  $\epsilon$ -regular.

Color any  $\epsilon$ -irregular pair white; any  $\epsilon$ -regular pair with density less than  $1/2$ , red; and any  $\epsilon$ -regular pair with density at least  $1/2$ , blue. Lemma 2.11 allows us to conclude that the reduced graph contains either a red  $K_n$  or a blue  $K_n$ . ■

## 2.2 Proof of Theorem 2.2

Let  $G$  be a graph on  $n$  vertices as described in Corollary 2.7. Let  $G'$  be a graph with no  $(\epsilon, r, L, \delta)$ -configuration having least distance from  $G$ .

Apply the Regularity Lemma to  $G'$ , to get  $\ell + 1$  clusters  $V_0, V_1, \dots, V_\ell$  such that  $|V_1| = \dots = |V_\ell| = L$ . Furthermore, all but  $\epsilon \ell^2$  pairs  $(V_i, V_j)$ ,  $1 \leq i < j \leq \ell$  are  $\epsilon$ -regular. We will consider only pairs of clusters so that the density between them is at least  $\delta$  and at most  $1 - \delta$ .

By assumption, it is not possible for the reduced graph to have a set of  $r$  clusters such that between any two clusters, there is an  $\epsilon$ -regular pair with density at least  $\delta$  and at most

$1 - \delta$ . Thus, according to Turán's theorem, the number of pairs of clusters that either have density at most  $\delta$ , or at least  $1 - \delta$ , or are  $\epsilon$ -irregular is at least

$$(r - 1) \frac{\frac{\ell}{r-1} \left( \frac{\ell}{r-1} - 1 \right)}{2} = \frac{\ell(\ell - r + 1)}{2(r - 1)}.$$

Since the number of  $\epsilon$ -irregular pairs is at most  $\epsilon\ell^2$ , the number of pairs with density at most  $\delta$  or at least  $1 - \delta$  is at least

$$\ell^2 \left( \frac{1}{2(r - 1)} - \epsilon \right) - \frac{\ell}{2}.$$

Since  $G$  came from Corollary 2.7, if any pair in the reduced graph of  $G'$  has density at most  $\delta$  or at least  $1 - \delta$ , then at least  $(1/2 - \delta - o(1))L^2$  edges had to have been either added or deleted. Hence, the total number of edges that had to be changed is at least

$$\ell^2 L^2 \left( \frac{1}{2(r - 1)} - \epsilon - \frac{1}{\ell} \right) \left( \frac{1}{2} - \delta - o(1) \right) \leq \ell^2 L^2 \left( \frac{1}{4(r - 1)} - \frac{\delta}{2(r - 1)} \right).$$

The inequality is valid as long as  $\epsilon < \delta$ .

We can bound the total number of edges that were altered by

$$\frac{n^2}{4(r - 1)} (1 - 2\delta)(1 - \epsilon)^2.$$

■

### 2.3 Proof of Theorem 2.3

Let  $H$  have  $n$  vertices and binary chromatic number  $r = \chi_B(H)$ . We shall use constants  $\epsilon_i$ ,  $i = 1, \dots, r$ , that satisfy the following relations

$$0 < \epsilon \ll \epsilon_1 \ll \dots \ll \epsilon_r \ll \delta,$$

where  $a \ll b$  means that  $a$  is sufficiently smaller than  $b$  so as to satisfy all necessary conditions. We shall define the precise values that we need in the proof and verify, in Appendix A, that such values are sufficient to enable us to carry out the required calculations.

Let the clusters of our  $(\epsilon, r, L, \delta)$ -configuration be  $V_1, \dots, V_r$ . In order to show that the existence of such a configuration implies an induced copy of  $H$ , we need to apply the Regularity Lemma (Lemma 2.4) to each of the clusters themselves. The Regularity Lemma generally does not address the subgraphs induced by clusters. We will, however, be able to use a Ramsey-type argument to show that each cluster itself will either induce many small cliques or many small cocliques.

We do this by demonstrating that the reduced graph (as defined in the statement of the Regularity Lemma) of the graph induced by each cluster  $V_i$  has a  $K_n$  either with all sparse edges or all dense edges.

Let us apply Lemma 2.11 to each of  $V_1, \dots, V_r$  with parameter  $\epsilon_1$ . For  $i \in \{1, \dots, r\}$ , call  $V_i$  a *dense cluster* if it has  $n$  subsets that satisfy Lemma 2.11(1) and call  $V_i$  a *sparse cluster* if it has  $n$  subsets that satisfy Lemma 2.11(2).

Since  $r = \chi_B(H)$ , we can find disjoint subsets of  $V(H)$ ,  $C_1, \dots, C_r$ , so that  $C_i$  induces a clique whenever  $V_i$  is dense and  $C_i$  induces a coclique whenever  $V_i$  is sparse. Denote  $n_i := |C_i|$ . Let

$$C_i = \{c_i(1), \dots, c_i(n_i)\}.$$

With these preliminaries, we can begin the proof itself. The proof, by induction on  $i = 1, \dots, r$  is of the following statement:

For all  $j, k$  with  $j < i \leq k$ ,  $V_j$  contains vertices  $s_j(1), \dots, s_j(n_j)$  and  $V_k$  contains subsets  $S_k(1), \dots, S_k(n_k)$  such that each of the following hold:

- Q1**  $\bigcup_{j=1}^{i-1} \bigcup_{m=1}^{n_j} \{s_j(m)\}$  induces a graph isomorphic to  $H[C_1 \cup \dots \cup C_{i-1}]$ .
- Q2** Each vertex in  $S_k(m)$  is
  - (a) adjacent to  $s_j(m')$  whenever  $c_k(m) \sim c_j(m')$  and
  - (b) nonadjacent to  $s_j(m')$  whenever  $c_k(m) \not\sim c_j(m')$ .
- Q3**  $|S_k(m)| = L_i$  for all  $m = 1, \dots, n_k$ .
- Q4**  $(S_k(m), S_k(m'))$  is  $\epsilon_i$ -regular for all distinct  $m, m' \in \{1, \dots, n_k\}$ .
- Q5** If  $V_k$  is dense [sparse], then each  $(S_k(m), S_k(m'))$  has density less than  $1/2 + \epsilon_1 + \dots + \epsilon_{i-1}$  [more than  $1/2 - \epsilon_1 - \dots - \epsilon_{i-1}$ ].

The base case of the induction is  $i = 1$  and it is easy. Items **Q1** and **Q2** are null. Items **Q3**, **Q4** and **Q5** are given directly by Lemma 2.11.

Now we suppose the statement is true for  $i - 1$  and prove it for  $i$ . If  $V_i$  is dense [sparse], we will find a clique [coclique] of size  $n_i$ . Let  $s_i(\ell) \in S_i(\ell)$  for  $\ell = 1, \dots, m$ .

For an  $n_i$ -tuple  $\mathbf{s}_i = (s_i(1), \dots, s_i(m))$ , we define its *generalized neighborhood in  $S_k(m')$*  to be the vertices that are adjacent to  $s_i(m)$  if and only if  $c_k(m') \sim c_i(m)$  in  $H$ . We denote this as follows:

$$N(\mathbf{s}_i, S_k(m')) = S_k(m') \cap \bigcap_{m:c_i(m) \sim c_k(m')} N(s_i(m)) \cap \bigcap_{m:c_i(m) \not\sim c_k(m')} \overline{N(s_i(m))}$$

We say that  $\mathbf{s}_i$  is *good in  $S_k(m')$*  if  $|N(\mathbf{s}_i, S_k(m'))| \geq (\delta - \epsilon)^{n_i} L_1^{n_i}$ ; otherwise, the  $\mathbf{s}_i$  is said to be *bad in  $S_k(m')$* .

We say that  $\mathbf{s}_i$  itself is good if it is good in  $S_k(m')$  for all  $m' \in \{1, \dots, n_k\}$  and all  $k \in \{i + 1, \dots, r\}$  and it induces a clique [coclique] when  $V_i$  is dense [sparse].

We refer to the Embedding Lemma. In order to apply it to  $S_i(1), \dots, S_i(n_i)$ , we need  $\epsilon_i$  to be small enough (condition (13) in Appendix A). The Embedding Lemma gives that the number of cliques [cocliques] with one vertex in each of  $S_i(1), \dots, S_i(n_i)$  is at least

$$\left( \frac{(1/2 - \sum_{j=1}^i \epsilon_j)^{n_i-1}}{n_i + 1} \right)^{n_i} L_i^{n_i}$$

Using the fact that  $\epsilon \ll \delta$  (condition (14) in Appendix A) and the Revised Intersection Property, we have that the number of ordered  $n_1$ -tuples that are bad in a given  $S_k(m')$  for  $m' \in \{1, \dots, n_k\}$  and  $k \in \{i+1, \dots, r\}$  is at most  $n_i \epsilon L^{n_i}$ . As a result, the total number of  $n_i$ -tuples in  $V_i$  that are bad for some  $S_k(m')$  is at most

$$\left( n - \sum_{j=1}^i n_j \right) n_i \epsilon L^{n_i}.$$

In order to have at least one good clique [coclique] in  $(S_i(1), \dots, S_i(n_i))$ , we require  $\epsilon \ll L_i/L$  (condition (15)). Label its vertices  $s_i(1), \dots, s_i(n_i)$  so that  $s_i(m) \in S_i(m)$ ,  $\forall j \in \{1, \dots, n_i\}$ .

We will redefine  $S_k(m')$  to be  $N(\mathbf{s}_i, S_k(m'))$  for all  $m' \in \{1, \dots, n_k\}$  and all  $k \in \{i+1, \dots, r\}$ . In particular, we make it a subset of size exactly  $\lceil (\delta - \epsilon)^{n_i} L_i \rceil$ . According to this redefinition, items **Q1** and **Q2** are satisfied for the value of  $i+1$ . Item **Q3** is satisfied by setting  $L_{i+1} = \lceil (\delta - \epsilon)^{n_i} L_i \rceil$  (see equation (12) in Appendix A).

The Slicing Lemma is given in [12].

**Fact 2.12 (Slicing Lemma)** *Let  $(A, B)$  be an  $\epsilon$ -regular pair with density  $d$  and, for some  $\alpha > \epsilon$ , let  $A' \subseteq A$ ,  $|A'| \geq \alpha|A|$ ,  $B' \subseteq B$ ,  $|B'| \geq \alpha|B|$ . Then  $(A', B')$  is an  $\epsilon'$ -regular pair with  $\epsilon' = \max\{\epsilon/\alpha, 2\epsilon\}$ , and for its density  $d'$  we have  $|d' - d| < \epsilon$ .*

In order to apply the Slicing Lemma, we need  $\epsilon_{i+1}$  to be small enough (condition (16)).

Let  $\epsilon_{i+1} = \frac{\epsilon_i}{(\delta - \epsilon)^{n_i}}$ . Then our new  $S_k(m')$  sets being of size  $L_{i+1}$  implies that they are  $\epsilon_{i+1}$ -regular with density more than  $1/2 - \sum_{j=1}^{i-1} \epsilon_j - \epsilon_i$  [less than  $1/2 + \sum_{j=1}^{i-1} \epsilon_j + \epsilon_i$ ] if  $V_i$  is dense [sparse]. Therefore, items **Q4** and **Q5** are satisfied for  $i+1$ .

Finally, we find a clique [coclique] in  $(S_r(1), \dots, S_r(n_r))$  if  $V_i$  is dense [sparse]. The Embedding Lemma gives that, if  $\epsilon_r$  is small enough (condition (13)) the number of such cliques [cocliques] is at least

$$\left( \frac{(1/2 - \sum_{j=1}^r \epsilon_j)^{n_i-1}}{n_i + 1} \right)^{n_i} L_i^{n_i}. \quad (7)$$

The quantity in (7) must be at least 1. This is true provided  $\epsilon_r$  is small enough and  $L$  is sufficiently large relative to  $\delta$ . I.e., **Q1** holds for  $i = r+1$  and we have our desired  $H'$ , a copy of  $H$ . ■

## 2.4 Proof of Theorem 1.3

Choose a  $\delta$  arbitrarily small, and let  $G$  be the graph guaranteed by Theorem 2.2. If fewer than  $\frac{n^2}{4k}(1 - 2\delta)(1 - \epsilon)^2$  edge-operations are performed, then there are vertex-disjoint sets  $V_1, \dots, V_r$  that satisfy the conditions of Theorem 2.3. Theorem 2.3 then implies that  $G$  contains an induced  $H$ .

So, the editing distance is at least  $\frac{n^2}{4k}(1 - 2\delta)(1 - \epsilon)^2$  and, since  $\delta$  is arbitrary, we have the result. ■

## 2.5 Proof of Theorem 1.4

Lemma 2.13 emphasizes the importance of the  $c$  in the definition of  $\chi_B$ .

**Lemma 2.13** *Let  $H$  be a graph with binary chromatic number  $k + 1$ . Let  $c$  be an integer,  $0 \leq c \leq k$ , so that  $H$  cannot be covered by exactly  $c$  cliques and exactly  $k - c$  independent sets. Let  $G$  be a graph with density  $d = e(G)/\binom{n}{2}$ . As long as it is not the case that  $d = 0$  and  $c = k$  or  $d = 1$  and  $c = 0$ , then*

$$\text{Dist}(G, \text{Forb}(H)) \leq \frac{d(1-d)}{dc + (1-d)(k-c)} \binom{n}{2}. \quad (8)$$

Otherwise,  $\text{Dist}(G, \text{Forb}(H)) \leq \frac{1}{k} \binom{n}{2}$ .

*Proof.* We begin by assigning colors independently to the vertices of  $G$ :  $1, \dots, c$  each with probability  $p$  and  $c+1, \dots, k$  each with probability  $q$ , call such a coloring  $c$ . If  $c(x) = c(y) \in \{1, \dots, c\}$  and  $xy \notin E(G)$ , then add an edge  $xy$  to  $E(G)$ . If  $c(x) = c(y) \in \{c+1, \dots, k\}$ , and  $xy \in E(G)$ , then delete  $xy$  from  $E(G)$ . As a result, we obtain a graph  $G'$  with the vertex set partitioned into  $k$  subsets. The first  $c$  of these subsets induce cliques and the others induce independent sets. Since the vertices of  $H$  can not be partitioned into  $c$  cliques and  $k - c$  cocliques,  $H \not\subseteq G'$ .

The expected number of changes is

$$f(p, q) = \left( \binom{n}{2} - e(G) \right) cp^2 + e(G)(k-c)q^2 = ((1-d)cp^2 + d(k-c)q^2) \binom{n}{2}.$$

We also have the restriction

$$cp + (k-c)q = 1. \quad (9)$$

As long as we do not have the case that  $d = 0$  and  $c = k$  or the case that  $d = 1$  and  $c = 0$ , the method of Lagrange multipliers gives that the minimum of  $f(p, q)$  restricted to (9) occurs when  $p = d/(dc + (1-d)(k-c))$  and  $q = (1-d)/(dc + (1-d)(k-c))$  and is

$$\frac{d(1-d)}{dc + (1-d)(k-c)} \binom{n}{2}.$$

Since this is the expected number of changes, there is a partition of the vertices of  $G$  such that the above procedure requires only  $\frac{d(1-d)}{dc+(1-d)(k-c)} \binom{n}{2}$  changes to make the graph  $H$ -free.

If  $d = 0$  and  $c = k$ , then perform the above procedure, but fix  $p = 1$ . If  $d = 1$  and  $c = 0$ , then perform the above procedure, but fix  $q = 1$ . In both cases, an expectation of  $\frac{1}{k} \binom{n}{2}$  changes will be performed.  $\blacksquare$

In order to prove equation (2) of Theorem 1.4, we find conditions when

$$\frac{d(1-d)}{dc + (1-d)(k-c)} \leq \frac{1}{2k}. \quad (10)$$

If  $c \leq k/2$ , then (10) holds when  $d \in [0, 1/2] \cup [1 - c/k, 1]$ . If  $c \geq k/2$ , then (10) holds when  $d \in [0, 1 - c/k] \cup [1/2, 1]$ . Consider a graph  $G$  of density  $d$ . If  $d \leq 1/2$ , then choose  $c_{\min}$ ; otherwise, choose  $c_{\max}$ . As a result, for a graph  $G$  of any density,  $\text{Dist}(G, \text{Forb}(H)) \leq \frac{1}{2k} \binom{n}{2}$ .

In order to prove inequality (3) of Theorem 1.4 we need to maximize expression (8) over  $d$ . The maximum value occurs when  $d = \frac{k-c-\sqrt{c(k-c)}}{k-2c}$  and is

$$\frac{k-2\sqrt{c(k-c)}}{(k-2c)^2} \binom{n}{2} = \left( \frac{1}{1+2\sqrt{\frac{c}{k}\left(1-\frac{c}{k}\right)}} \right) \frac{1}{k} \binom{n}{2}.$$

The expression in parentheses is at most 1. ■

### 3 Binary chromatic number

It is easy to see the following

**Fact 3.1** *Let  $G$  be a graph.*

1.  $\chi_B(G) \geq \chi(G), \chi(\overline{G})$
2.  $\chi_B(G) = \chi_B(\overline{G})$ .

**Proposition 3.2** *Let  $G$  be any graph.*

$$\chi_B(G) \leq \chi(G) + \chi(\overline{G}) - 1.$$

*This bound is tight for  $G = K_p^q$ .*

*Proof.* Consider  $c$  cliques spanning a set of vertices  $A$ . We can always assume that  $c \leq \chi(\overline{G})$ . If  $c < \chi(\overline{G})$  we are done since  $\chi(G-A) \leq \chi(G)$ . Otherwise,  $c = \chi(\overline{G})$  and it is possible to partition all vertices into  $c$  cliques. We can obtain required independent sets by considering single vertices.

The second statement is a result of Proposition 3.4(3). ■

By Proposition 3.4(3), we see that the maximum binary chromatic number over all  $n$ -vertex graphs is  $n$ . This is achieved for the graphs  $K_n$  and  $\overline{K_n}$ . Proposition 3.3 gives the bounds on the smallest binary chromatic number among all  $n$ -vertex graphs.

**Proposition 3.3** *Let  $n$  be a positive integer, then*

$$\sqrt{n} \leq \min_{|V(G)|=n} \chi_B(G) \leq \sqrt{n} + (1+o(1))n^{0.2625}.$$

*Moreover there are infinitely many graphs for which the lower bound is attained.*

*Proof.* For the lower bound, we use Fact 3.1(1) and the fact that  $\chi(G)\chi(\overline{G}) \geq n$ . As a result, one of  $\chi(G), \chi(\overline{G})$  is larger than  $\sqrt{n}$ .

The lower bound is, in fact, attained by an infinite class of graphs on  $n = k^2$  vertices where  $k$  is a prime. To realize this lower bound, consider the following construction. Vertices are pairs of integers  $(i, j)$ ,  $i, j = 1, \dots, k$ . Next we create  $k + 1$  distinct partitions of  $V(G)$  into sets of cardinalities  $k$ . Let the  $i^{\text{th}}$  partition  $P_i = \{V_1^i, V_2^i, \dots, V_k^i\}$  be defined as follows for  $i = 0, \dots, k$ .  $V_j^i = \{(j, 1), (j + i, 2), (j + 2i, 3), \dots, (j + (k - 1)i, k)\}$ . Here, addition is taken modulo  $k$ .

One could think of  $P_i$  as a set of lines of slope  $i$  when the vertices of  $G$  are identified with points on a square, toroidal grid. Note that if  $xy \in V_j^i$  then  $xy \notin V_{j'}^{i'}$  where  $i \neq i'$ . Indeed, if  $x, y \in V_i^j$  and  $x = (x_1, x_2)$  then  $y = (x_1 + li, x_2 + l)$ . If  $x, y \in V_{i'}^{j'}$  then  $y = (x_1 + l'i', x_2 + l')$ . Now, since  $x_2 + l = x_2 + l'$  we have  $l = l' \pmod{k}$ . Thus  $x_1 + li = x_1 + l'i = x_1 + l'i'$ , therefore  $li = l'i'$  and  $i = i'$  if  $k$  is prime. Now, we let  $V_j^i$  induce a clique if  $i < j$  and let  $V_j^i$  induce a coclique if  $i \geq j$ . We see that  $P_0$  gives  $k$  cliques covering  $V(G)$ ,  $P_2$  gives one coclique and  $k - 1$  cliques covering  $V(G)$ , and so on. Finally,  $P_k$  gives  $k$  cocliques covering  $V(G)$ . Thus, we can always partition the vertex set into  $i$  cliques and  $k - i$  cocliques for any  $i = 0, \dots, k$ .

For arbitrary  $n$ , take the smallest  $k \geq \sqrt{n}$  such that  $k$  is a prime. Consider  $G_{k^2}$  as defined above. As we have shown,  $\chi_B(K_{k^2}) \leq k$ , which implies  $\chi_B(K_n) \leq k$ . In a paper of Baker, Hartman and Pintz [1], for  $x$  at least some  $x_0$ , there is a prime in the interval  $[x - x^{0.525}, x]$ . Thus,  $\chi_B(G) \leq k \leq \sqrt{n} + (1 + o(1))n^{0.2625}$ .  $\blacksquare$

Next we determine the binary chromatic number of some classes of graphs.

**Proposition 3.4** *Let  $\chi_B(G)$  denote the binary chromatic number of  $G$ .*

1. *If  $n \geq 5$ , then  $\chi_B(C_n) = \lceil n/2 \rceil$ .*
2. *If  $n \geq 6$ , then  $\chi_B(P_n) = \lceil n/2 \rceil$ .*
3.  *$\chi_B(K_p^q) = p + q - 1$ .*

*Proof.*

1. The lower bound follows from 1. For the upper bound, we can construct the partition of a vertex set in at most  $\lceil \frac{n}{2} \rceil$  cliques and cocliques as follows. If we need only cliques, or only cocliques, it is clear. When we need at least one clique and at least one coclique in that partition, take the largest coclique on  $\lfloor \frac{n}{2} \rfloor$  vertices. The leftover graph consists of independent vertices and, if  $n$  is odd, of one edge. Take this edge (or a single vertex when  $n$  is even) as a clique of our partition. The number of leftover vertices is  $\lceil \frac{n}{2} \rceil - 2$  and we are done.
2. This is quite similar to the case of  $C_n$ . We leave it to the reader.
3. We see that  $\chi_B(K_p^q) \geq p + q - 1$  by observing that if we require  $q - 1$  cliques in a partition of a vertex set of a graph into cliques and cocliques then  $p - 1$  cocliques is not enough. The upper bound follows from Proposition 3.2.

■

## 4 Better bounds for small graphs

The previous results are asymptotic. However, for some  $H$  we are able to determine the exact value of  $\text{Dist}(n, \text{Forb}(H))$ .

Here we shall use the fact that the extremal graphs for forbidden induced subgraphs on three vertices as well as for induced subgraphs on 4 vertices and 3 edges are known precisely [4].

**Theorem 4.1** *Let  $H \in \{K_3, \overline{K_3}, K_{1,2}, \overline{K_{1,2}}\}$ . Then  $\text{Dist}(n, \text{Forb}(H)) = \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}$ .*

*Proof.* The case of the triangle  $K_3$  and the empty graph  $\overline{K_3}$  follows immediately from 1.

Now, we consider the editing distance for  $K_{1,2}$ -free graphs. Note that the graph which contains no induced  $K_{1,2}$  is a disjoint union of cliques.

Let  $G$  be an arbitrary graph on  $n$  vertices. If  $G$  has minimal degree at least  $\lceil n/2 \rceil$  then we add all missing edges to obtain a complete graph. In this case, at most  $\binom{n}{2} - \lceil n/2 \rceil n/2 \leq \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}$  edges were added. Otherwise, delete all edges incident to a vertex  $v$  of degree at most  $\lfloor n/2 \rfloor$  and apply induction to  $G \setminus v$ . The total number of additions and deletions is at most  $\lfloor n/2 \rfloor + \binom{\lceil (n-1)/2 \rceil}{2} + \binom{\lfloor (n-1)/2 \rfloor}{2} \leq \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}$ . This provides an upper bound on  $f$ .

For the lower bound, we consider a complete bipartite graph  $H$  on  $n$  vertices with almost equal parts  $A, B$ . Let  $G$  be the disjoint union of cliques  $S_1, S_2, \dots, S_k$  on the same vertex set as  $H$ . Let  $a_i = |A \cap V(S_i)|$  and  $b_i = |B \cap V(S_i)|$ , for  $i = 1, \dots, k$ . It is clear that the number of editing operations performed on  $H$  to obtain  $G$  is

$$s = \sum_{i=1}^k k \binom{a_i}{2} + \binom{b_i}{2} + a_i(|B| - a_i).$$

This function is minimized when  $a_i = b_i$  for all, except perhaps one  $i = 1, \dots, k$  such that if  $a_i \neq b_i$ ,  $|a_i - b_i| \leq 1$ . Now,  $s \geq n^2/4 - n/2$  for even  $n$  and  $s \geq (n-1)^2/4 - (n-1)/2 + (n-1)/2$  and we are done. ■

Let  $\mathcal{Q}$  be the set of graphs on  $n$  vertices with no induced subgraphs on 4 vertices and 3 edges. In [5] it was shown that any graph in  $\mathcal{Q}$  or its complement is a disjoint union of 4-cycles and trees on at most 3 vertices.

**Theorem 4.2**  $(n^2 - 7n)/4 \leq \text{Dist}(n, \mathcal{Q}) \leq (n^2 - n)/4$ .

*Proof.* Since  $E_n, K_n \in \mathcal{Q}$ , the upper bound is trivial.

We claim that the editing problem can actually be reduced to only edge additions or only edge deletions. Indeed, consider a graph  $G$  which, after performing the smallest possible number of editing operations, is transformed into the disjoint union of trees on at most 3

vertices and 4-cycles. Consider connected components  $A_1, \dots, A_m$  of the resulting graph  $G'$ . Each  $A_i$  has at most 4 vertices. In order to obtain  $G'$ , one had to delete all edges between  $A_i$ s and no additions would have helped. Inside each  $A_i$  we could get away only with edge-deletions by considering the cases:

1.  $G[A_i] = K_4$  requires 2 edge-deletions to get  $C_4$ ,
2.  $G[A_i] = P_4$  requires either one edge-deletion or one edge-addition to get either  $P_3$  or  $C_4$ ,
3.  $G[A_i] = K_4 \setminus e$  requires two operation, both can be edge-deletions,
4. other cases similarly require only edge-deletions.

For the lower bound consider a graph  $G$  on  $(n^2 - n)/4$  edges. Assume first that the minimal number of edit operations results in a graph whose components are either 4-cycles or trees on at most 3 vertices. Then the total number of edges within these components is at most  $(6/4)n$ . Therefore, at least  $|E(G)| - (3/2)n$  edges of  $G$  had to be deleted. Consequently, the number of edit operations is at least  $(n^2 - n)/4 - (3/2)n$ . If the complement of the resulting graph has components that are all 4-cycles or trees on at most 3 vertices, then the symmetric argument with edge-additions yields the same conclusion. ■

## 5 Conclusions

The editing problem of graphs we consider in this paper can be reformulated in terms of complete edge-colored graphs, where the edges of the graphs correspond to edges of one color, say red, and the edges of the complement correspond to edges of another color, say blue. Then our editing operations are equivalent to changing the color of some edges from red to blue or from blue to red.

It is natural to consider more than two colors. Specifically for any two colorings of  $E(K_n)$  in color  $\{1, \dots, \gamma\}$ , we define the distance to the smallest number of edge-recolorings to obtain one coloring from the other. Our results for classes of graphs with forbidden induced subgraphs can be generalized for classes of multicolored graphs with forbidden color patterns.

Another generalization is to consider both  $H$ , the forbidden graph, and  $G$  to be from some restricted class of graphs and some changes not being permitted. For example, both  $H$  and  $G$  could be bipartite and no edges may be added to either partite set of  $G$ . This, in particular, will provide an answer to the editing distance question regarding  $\{0, 1\}$ -matrices that are not necessarily symmetric.

## References

- [1] R.C. Baker, G. Harman and J. Pintz, The difference between consecutive primes, II. *Proc. London Math. Soc.* **83** (2001), no. 3, 532–562.

- [2] D. Chen, O. Eulenstein, D. Fernández-Baca and M. Sanderson, Supertrees by flipping, preprint.
- [3] R.G. Downey and M.R. Fellows, *Parameterized complexity*. Springer, New York, 1999.
- [4] P. Erdős, R. Faudree, J. Pach and J. Spencer, How to make a graph bipartite. *J. Combin. Theory Ser. B* **45** (1988), no. 1, 86–98.
- [5] P. Erdős, Z. Füredi, B. L. Rothschild, and V. T. Sós, Induced subgraphs of given sizes. *Paul Erdős memorial collection*. Discrete Math. **200** (1999), no. 1-3, 61–77.
- [6] P. Erdős, A. Gyárfás and M. Ruszinkó, How to decrease the diameter of triangle-free graphs. *Combinatorica* **18** (1998), no. 4, 493–501.
- [7] P. Erdős, E. Györi and M. Simonovits, How many edges should be deleted to make a triangle-free graph bipartite? *Sets, graphs and numbers* (Budapest, 1991), 239–263, Colloq. Math. Soc. János Bolyai, **60**, North-Holland, Amsterdam, 1992.
- [8] P. Erdős and A. H. Stone, On the structure of linear graphs. *Bull. Amer. Math. Soc.* **52**, (1946), 1087–1091.
- [9] Z. Füredi, Turán type problems. *Surveys in combinatorics*, 253–300, London Math. Soc. Lecture Note Ser., **166**, Cambridge Univ. Press, Cambridge, 1991.
- [10] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.
- [11] 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.
- [12] 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.
- [13] N. Pippenger and M. C. Golumbic, The inducibility of graphs. *J. Combinatorial Theory Ser. B* **19** (1975), no. 3, 189–203.
- [14] M. Simonovits, *Extremal graph theory*, Selected topics in graph theory, **2**, 161–200, Academic Press, London, 1983.
- [15] G. Stephanopoulos, A. Aristidou and J. Nielsen, *Metabolic engineering: principles and methodologies*. Academic Press, San Diego, 1998.
- [16] E. Szemerédi, Regular partitions of graphs. In *Problèmes Combinatoires et Théorie des Graphes*, 399–401, Colloq. Internat. CNRS, Univ. Orsay, Paris, 1978.
- [17] D. West, *Introduction to Graph Theory*, second edition, Prentice Hall, 2001, pp. 588.

## A The constants for Theorem 2.3

In the proof of Theorem 2.3, there is a number of constants. This is to be expected in proofs involving the Regularity Lemma, but we would like to verify the existence of  $\epsilon, \epsilon_1, \dots, \epsilon_r$  satisfying the desired constraints. Each of  $\epsilon, \epsilon_1, \dots, \epsilon_r$  is a function of  $\delta$ .

If we are given  $\epsilon_1$ , we define  $M = M(\epsilon_1, R_{\epsilon_1}(n))$ , the Regularity constant with the generalized Ramsey number as a parameter. We set  $L_1 = \lceil L/M \rceil$  and, for  $i = 1, \dots, r-1$ ,

$$\epsilon_{i+1} = (\delta - \epsilon)^{-n_i} \epsilon_i \quad (11)$$

$$L_{i+1} = \lceil (\delta - \epsilon)^{n_i} L_i \rceil \quad (12)$$

Next, we need to specify the bounds on our constants.

$$\epsilon_i \leq \frac{(1/2 - \sum_{j=1}^i \epsilon_j)^{n_i+1}}{n_i + 1} \quad (13)$$

$$\epsilon < (\delta - \epsilon)^{n_i-1} \quad (14)$$

$$\left( n - \sum_{j=1}^i n_j \right) n_i \epsilon L^{n_i} < \left( \frac{(1/2 - \sum_{j=1}^i \epsilon_j)^{n_i-1}}{n_i + 1} \right)^{n_i} L_i^{n_i} \quad (15)$$

$$\epsilon_i < \left( \frac{1}{2} - \sum_{j=1}^i \epsilon_j \right)^{n_i} \quad (16)$$

Note that equations (11-12) for  $i = 1, \dots, r-1$  and conditions (13-16) for  $i = 1, \dots, r$  imply the inequalities  $0 < \epsilon \ll \epsilon_1 \ll \dots \ll \epsilon_r \ll \delta$ , where  $a \ll b$  means that  $a$  is sufficiently smaller than  $b$  so as to satisfy all such conditions.

Before we specify  $\epsilon$ , we only require  $\epsilon < \delta/2$  and  $\epsilon < (\delta/2)^{n-1}$ . This immediately satisfies condition (14). Now, set

$$\epsilon_1 = (\delta/2)^n \min\{1/(4r), (1/4)^n/(n+1)\}$$

Definition (11) gives  $\epsilon_1 = (\delta - \epsilon)^{n_r} \epsilon_r$ , which means we have  $\epsilon_r \leq 1/(4r)$  and  $\epsilon_r \leq (1/4)^n/(n+1)$ . Definition (11) also gives that  $\epsilon_i < \epsilon_r$  for all  $i < r$ , satisfying conditions (13) and (16). It remains to show that  $\epsilon$  satisfies the following:

$$\left( n - \sum_{j=1}^i n_j \right) n_i \epsilon L^{n_i} < \left( \frac{(1/4)^{n_i-1}}{n_i + 1} \right)^{n_i} L_i^{n_i} \quad (17)$$

Let  $M = M(\epsilon_1, R_{\epsilon_1}(n))$ , the Regularity constant with the generalized Ramsey constant as a parameter. Definition (12),  $\epsilon < \delta/2$  and  $L_1 = \lceil L/M \rceil$  gives

$$L_i \geq (\delta - \epsilon)^{\sum_{j=1}^i n_j} L_1 \geq \frac{(\delta - \epsilon)^n L}{M} \geq \frac{(\delta/2)^n L}{M}$$

As a result, as long as

$$\epsilon < \frac{4}{n^2} \left( \frac{(1/4)^{n-1} (\delta/2)^n}{M(n+1)} \right)^n,$$

inequality (17) is satisfied.