MULTICOLOR AND DIRECTED EDIT DISTANCE

MARIA AXENOVICH AND RYAN MARTIN

Abstract. The editing of a combinatorial object is the alteration of some of its elements such that the resulting object satisfies a certain fixed property. The edit problem for graphs, when the edges are added or deleted, was first studied independently by the authors and Kézdy [*J. Graph Theory* (2008) **58**(2), 123–138] and by Alon and Stav [*Random Structures Algorithms* (2008) **33**(1), 87–104]. In this paper, a generalization of graph editing is considered for multicolorings of the complete graph as well as for directed graphs. Specifically, the number of edge-recolorings sufficient to be performed on any edge-colored complete graph to satisfy a given hereditary property is investigated. The theory for computing the edit distance is extended using random structures and so-called types or colored homomorphisms of graphs.

Contents

2010 *Mathematics Subject Classification.* Primary 05C35; Secondary 05C80.

Key words and phrases. edit distance, hereditary properties, localization, split graphs, colored regularity graphs.

This author's research partially supported by NSF grant DMS-0901008 and NSA grant H-98230-09-1-0063.

This author's research partially supported by NSF grant DMS-0901008 and by an Iowa State University Faculty Professional Development grant.

1. Introduction

The combinatorial editing problem is, in general, the problem of finding the smallest number of element-changes such that the resulting combinatorial object satisfies a certain fixed property.

The simplest class of objects for which the editing problem was considered is a set of sequences. In fact, the first detailed algorithmic study of editing was motivated by bionformatics, where sequences over finite alphabets are considered and editing corresponds to changes of the elements in the sequence depicting the mutations in biomolecules. When the desired property consists of a single sequence, studying editing corresponds to investigating the Hamming distance between sequences. We prove Theorem 3 in Section 2.9. The notion of graph editing was introduced by the authors and Kézdy, [4], and independently by Alon and Stav [3]. The question considered was: "How many edges does one need to add or delete in a given graph, such that the result belongs to a given class of graphs?" Alon and Stav [3] showed that for hereditary classes of graphs, the worst case scenario is realized by a random graph. Moreover, the general bounds were given in terms of certain graph parameters.

In this paper, the generalized theory is developed for editing of edge-colored complete graphs and digraphs. The main result for edge-colored graphs, Theorem 3, is in terms of a parameter called the r-ary chromatic number and the main result for directed graphs, Theorem 16 is in terms of a parameter called the directed chromatic number. In each case, the results come from more general theorems, Theorems 7 and 21 respectively, which deal with generalizing the graph notion of types for the above combinatorial objects. The analysis is based on using a version of Szemerédi's regularity lemma, Theorem 11 (see [5] for a proof), and applying it to an Erdős-Rényi-type random edge-colored graph or random digraph, respectively. General bounds on the edit distance function are given, as well as some editing algorithms and computing methods, all of which result from Theorems 7 and 21.

The paper is structured as follows. Section 2 deals with the case of multicolorings of the edges of complete graphs. Section 3 deals with the case of directed graphs. In each of these sections we provide definitions and editing algorithms as well as some general theory and results.

2. Multicolorings of the complete graph

2.1. **Basic definitions.**

An *equipartition* of a finite set is a partition in which each pair of partite sets differ in size by at most one.

For a complete graph on vertex set V, and a finite set Q, we shall say that a Q*-coloring*, or more specifically, a Q*-edge-coloring* of this graph is a pair $G = (V, c)$, where $c : \binom{V}{2} \to Q$. Since it is sufficient to let $Q = \{1, \ldots, r\}$ for some integer r, we will refer to an $\{1, \ldots, r\}$-edge coloring of a complete graph as simply an r*-graph.* For any r-graph G, disjoint vertex sets V_i and V_j and color ρ, $\rho \in \{1, \ldots, r\}$, the expression $E_\rho(V_i)$ denotes the set of edges colored ρ with both endpoints in $G[V_i]$ and $E_\rho(V_i, V_i)$ denotes the set of edges colored ρ with one endpoint in V_i and the other in V_j. The *density vector of* V_i is an r-dimensional vector $\mathbf{p} = (p_1, \ldots, p_r)$, where $p_\rho = |E_\rho(V_i)|/\binom{|V_i|}{2}$ for $\rho = 1, \ldots, r$. The *density vector of the pair* (V_i, V_j) is an r-dimensional vector $\mathbf{p} = (p_1, \ldots, p_r)$, where $p_\rho = |E_\rho(V_i, V_j)|/(|V_i||V_j|)$ for $\rho = 1, \ldots, r$. Note that for such density vectors, $\sum_\rho p_\rho = 1$.

In this setting, a *graph property* is merely a set of r-graphs for some positive integer $r \geq 2$. If $G = (V, c)$ and $G' = (V, c')$ are r-graphs on n labeled vertices, then

$$\mathrm{dist}(G, G')$$

is the proportion of edges on which the colors differ, i.e., the number of edges on which the colors in G and G' differ, divided by $\binom{n}{2}$. We may call this the *normalized edit distance* between G and G'.

For any property $\mathcal{H}$, a coloring G, an integer n, we define $\mathrm{dist}(G,\mathcal{H})$, $\mathrm{dist}(n,\mathcal{H})$, and $\mathrm{dist}(\mathcal{H})$ as follows:

$$\begin{aligned}
\mathrm{dist}(G,\mathcal{H}) &:= \min\left\{\mathrm{dist}(G,G') : V(G') = V(G), G' \in \mathcal{H}\right\},\\
\mathrm{dist}(n,\mathcal{H}) &:= \max\{\mathrm{dist}(G,\mathcal{H}) : |V(G)| = n\},\\
\mathrm{dist}(\mathcal{H}) &:= \lim_{n\to\infty} \mathrm{dist}(n,\mathcal{H}).
\end{aligned}$$

Note that $\mathrm{dist}(G,G'), \mathrm{dist}(G,\mathcal{H}), \mathrm{dist}(n,\mathcal{H}), \mathrm{dist}(\mathcal{H}) \in [0,1]$.

The last parameter $\mathrm{dist}(\mathcal{H})$ is the limit of the largest proportion of the edges necessary to be changed in a coloring of a complete graph bring it to a property $\mathcal{H}$; we show the existence of this limit later.

A *hereditary property of r-graphs* (or, simply, *hereditary property*, where the context is understood) is a set of r-graphs that is closed under vertex-deletion and isomorphisms. Let an r-graph G' be an *induced coloring* of an r-graph G if G' can be obtained from G by vertex-deletion.

For an r-graph, H, of a complete graph, the family $\mathrm{Forb}(H)$ consists of all r-graphs that have no (induced) copies of H. For every hereditary property, $\mathcal{H}$, there is a family, $\mathcal{F}(\mathcal{H})$, of r-graphs such that $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$. If $\mathcal{F}$ is a family of r-graphs, then we use $\mathrm{Forb}(\mathcal{F})$ to denote $\bigcap_{H\in\mathcal{F}} \mathrm{Forb}(H)$.

2.2. **The r-ary chromatic number.**

Definition 1. *For a hereditary property $\mathcal{H} = \cap_{H\in\mathcal{F}(\mathcal{H})}\mathrm{Forb}(H)$ of r-graphs, a* ***good tuple*** *$(a_1,\dots,a_r)$ is an r-tuple of non-negative integers such that for some $H \in \mathcal{F}(\mathcal{H})$, the vertex set $V(H)$ can be partitioned into sets $S_1,\dots,S_r$ such that, for each $i \in \{1,\dots,r\}$ with $a_i \neq 0$, the partition can be further refined $S_i = V_{i,1} \cup \cdots \cup V_{i,a_i}$ such that each $V_{i,j} \in S_i$ does not induce an edge of color i. The* ***clique spectrum*** *of $\mathcal{H}$ is the set of all tuples $(a_1,\dots,a_r)$ that are NOT good. The* ***r-ary chromatic number*** *of $\mathcal{H}$, $\chi_r(\mathcal{H})$, is the maximum $\ell+1$ such that for some non-negative integers $a_1,\dots,a_r$ with $a_1+\cdots+a_r = \ell$, the tuple $(a_1,\dots,a_r)$ is in the clique spectrum of $\mathcal{H}$.*

Note that if $(a_1,\dots,a_r)$ is in the clique spectrum of $\mathcal{H}$ and G is an r-graph partitioned into a_1 cliques with no edge of color 1, a_2 cliques with no edge of color 2, etc., then $G \in \mathcal{H}$ since G contains no forbidden subgraphs from $\mathcal{F}(\mathcal{H})$.

In the case of $r=2$ and $\mathcal{H} = \mathrm{Forb}(H)$, $\chi_2(H)$ corresponds exactly to the *binary chromatic number of H*, introduced in [4]. This is also called the "colouring number" in related literature such as Bollobás and Thomason [16, 17]. In the case where $\mathcal{H}$ is a principal hereditary property, i.e., $\mathcal{H} = \mathrm{Forb}(H)$ for some H, then we denote the r-ary chromatic number of $\mathcal{H}$ to be, simply $\chi_r(H)$. Note that there is a monotonicity to the clique spectrum. Precisely: if $(a_1,\dots,a_r)$ is in a clique spectrum and $a'_i \le a_i$, $\forall i \in \{1,\dots,r\}$, then $(a'_1,\dots,a'_r)$ is in that clique spectrum. Similarly, if $(a'_1,\dots,a'_r)$ is a good r-tuple and $a'_i \le a_i$, $\forall i \in \{1,\dots,r\}$, then $(a_1,\dots,a_r)$ is also a good r-tuple. Note also that the zero vector is always in the clique spectrum.

2.2.1. *Examples illustrating the r-ary chromatic number of a hereditary family.*

(1) Let $r=3$ and $\mathcal{H}$ be a family of $\{1,2,3\}$-colored complete graphs not containing a triangle H_1 with colors $1,1,2$ on its edges and not containing a triangle H_2 with colors $2,2,3$ on its edges. So, $\mathcal{F}(\mathcal{H}) = \{H_1,H_2\}$. Since $r=3$, and $\mathcal{F}(\mathcal{H})$ contains a triangle, any 3-tuple (a_1,a_2,a_3) with $a_1+a_2+a_3 \ge 3$ must be good. Indeed, each of H_1 and H_2 can be vertex-partitioned into three parts such that each part is a single vertex, thus not inducing edges of any colors.

Thus, it is sufficient to consider the tuples with $a_1+a_2+a_3 \le 2$. The tuple $(1,0,0)$ is good since we can partition the vertex set of H_2 in one part not containing edges of color 1. Similarly, $(0,0,1)$ is good. By monotonicity, all tuples (a_1,a_2,a_3) with $a_1 \ge 1$ or $a_3 \ge 1$ are good.

However $(0,1,0)$ is not good because both H_1 and H_2 contain edges of color 2. But $(0,2,0)$ is good because H_1 can be vertex-partitioned in two parts not containing edges of color 2.

Thus the clique spectrum of $\mathcal{H}$ is $\{(0,1,0),(0,0,0)\}$. The 3-ary chromatic number of $\mathcal{H}$, $\chi_3(\mathcal{H})$, is therefore equal to 2.

(2) Let $r=3$ and $\mathcal{H}$ be a family of $\{1,2,3\}$-colored complete graphs not containing a triangle H_1 with colors $1,1,2$ on its edges. So, $\mathcal{F}(\mathcal{H})=\{H_1\}$. As in the previous example, we need to consider only tuples (a_1,a_2,a_3) such that $a_1+a_2+a_3\le 2$. The tuple $(0,0,1)$ is good because H_1 does not have edge of color 3. By monotonicity, all tuples (a_1,a_2,a_3) with $a_3\ge 1$ are good.

Let us consider the tuples with $a_3=0$. The tuples $(2,0,0)$, $(0,2,0)$ are good because one can partition the vertices of H_1 into two parts not containing edges of color 1, color 2, and color 3, respectively. The tuple $(1,1,0)$ is good because it is possible to partition the vertices of H_1 in two parts, one not containing an edge of color 1 and another not containing an edge of color 2. However, neither $(1,0,0)$ nor $(0,1,0)$ are not good tuples since H_1 has edges of colors 1 and 2. Thus the clique spectrum of $\mathcal{H}$ is $\{(1,0,0),(0,1,0),(0,0,0)\}$ and $\chi_3(\mathcal{H})=2$.

(3) Let $r=2$, which we can consider to be the graph case. Let H be a K_5 colored with edges colored with colors 1 and 2 such that each color class is a 5-cycle. Let $\mathcal{H}$ be a family of colorings not containing H, i.e., $\mathcal{F}(\mathcal{H})=\{H\}$.

We need only consider 2-tuples (i.e., pairs) (a_1,a_2) with $a_1+a_2\le 4$. It is relatively easy to see that $(2,1)$, $(1,2)$, $(3,0)$ and $(0,3)$ are good. The pairs $(2,0)$, $(0,2)$ and $(1,1)$ are not good since H has no monochromatic clique on more than 2 vertices, but has a total of 5 vertices. By monotonicity, $(1,0)$ and $(0,1)$ are also not good. Thus the clique spectrum of $\mathcal{H}$ is $\{(2,0),(1,0),(1,1),(0,2),(0,1),(0,0)\}$, and $\chi_2(\mathcal{H})=3$.

2.3. **A simple editing algorithm.**

Let $\mathcal{H}$ be a hereditary property of r-graphs, such that $\mathcal{H}=\bigcap_{H\in\mathcal{F}(\mathcal{H})}\mathrm{Forb}(H)$. Further, let $\ell=\chi_r(\mathcal{H})-1$ and $(a_1,\dots,a_r)$ be in the clique spectrum and $\sum_{i=1}^r a_i=\ell$.

Partition V into r sets $S_1,\dots,S_r$ and further refine the partition such that $S_i=V_{i,1}\cup\cdots\cup V_{i,a_i}$, for $i=1,\dots,r$ and then recolor the edges in each $V_{i,j}$ by recoloring the edges of color i with some other arbitrary color. This new coloring does not contain any $H\in\mathcal{F}(\mathcal{H})$, otherwise the tuple $(a_1,\dots,a_r)$ would be good for some H.

If the sizes of the $V_{i,j}$-s differ by at most one; i.e., $\lfloor n/\ell\rfloor\le|V_{i,j}|\le\lceil n/\ell\rceil$, then the number of changes provided by this algorithm is at most $\ell\binom{\lceil n/\ell\rceil}{2}$. Thus,

$$\mathrm{dist}(\mathcal{H})\le\lim_{n\to\infty}\frac{\ell\binom{\lceil n/\ell\rceil}{2}}{\binom{n}{2}}=\frac{1}{\ell}=\frac{1}{\chi_r(\mathcal{H})-1}.$$

2.4. **Previous results and new main results.**

In [4], the authors and Kézdy provide a general bound for $\mathrm{dist}(\mathcal{H})$ in the 2-color case.

Theorem 2 ([4]). *For any hereditary property of graphs, $\mathcal{H}$, with binary chromatic number $\chi_2\ge 2$,*

$$\frac{1}{2(\chi_2-1)}\le\mathrm{dist}(\mathcal{H})\le\frac{1}{\chi_2-1}.$$

Furthermore, if $\mathcal{H}=\mathrm{Forb}(H)$ such that H is self-complementary, then $\mathrm{dist}(\mathcal{H})=\frac{1}{2(\chi_2(H)-1)}$.

Here, we show a similar result in the general case.

Theorem 3. *Let $\mathcal{H}$ be a hereditary property of $\{1,\ldots,r\}$-edge-colorings of complete graphs. Let $\chi_r = \chi_r(\mathcal{H}) \geq 2$ be the r-ary chromatic number of $\mathcal{H}$. Then,*

$$\frac{1}{r(\chi_r - 1)} \leq \mathrm{dist}(\mathcal{H}) \leq \frac{1}{\chi_r - 1}.$$

Furthermore, if $\mathcal{H} = \mathrm{Forb}(H)$ such that all color classes of H are isomorphic, then $\mathrm{dist}(\mathcal{H}) = \frac{1}{r(\chi_r(H)-1)}$.

We prove Theorem 3 in Section 2.9. The upper bound is found in the simple editing algorithm, but to get the lower bound, we need a more general theory. This is Theorem 7 which is stated in Section 2.6. We also prove the result for symmetric colorings in Corollary 9. Theorem 7 gives the basic results that deal with computing the edit distance for given hereditary properties. To state these results, we need to provide some preliminary material.

2.5. **The edit distance function.**

2.5.1. *Preliminary definitions.* For an r-graph, $G = (V, c)$, and some color $\rho \in \{1,\ldots,r\}$, let $E_\rho(G)$ denote the graph on vertex set V corresponding to the edges with color ρ in c. For a positive integer r, recall that a density vector $\mathbf{p} = (p_1,\ldots,p_r)$ (we also refer to it as a *probability vector*) is a nonnegative real vector with the property that $\sum_{\rho=1}^{r} p_i = 1$. For any density vector $\mathbf{p} = (p_1,\ldots,p_r)$, and integer n, we denote[1]

$$\mathrm{dist}_n(\mathbf{p},\mathcal{H}) = \max\left\{\mathrm{dist}(G,\mathcal{H}) : |V(G)| = n \text{ and } |E_\rho(G)| = p_\rho\tbinom{n}{2}, \rho = 1,\ldots,r\right\}.$$

In Theorem 7, we show that the following limit exits, which we call the *edit distance function*:

$$\mathrm{dist}(\mathbf{p},\mathcal{H}) = \lim_{n\to\infty} \mathrm{dist}_n(\mathbf{p},\mathcal{H}).$$

Having the edit distance function at our disposal, we may also define $\mathrm{dist}(\mathcal{H}) = \max_{\mathbf{p}} \mathrm{dist}(\mathbf{p},\mathcal{H})$, where the maximum is taken over all density vectors.

2.5.2. *Types of colorings.* In Section 2.6.1, we define two functions which are described in terms of types of colorings, which allow us to compute the edit distance function. In Section 2.6, we shall provide an algorithm to do such computation. We define a notion which was called a colored regularity graph (CRG) by Alon and Stav [3], but earlier called a *type* by Bollobás and Thomason [16]. We adopt latter terminology.

Definition 4. *An r-**type** (or just, **type**, where the context is understood), K, is a pair (U,ϕ), where U is a finite set of vertices and $\phi : U \times U \to 2^{\{1,\ldots,r\}} \setminus \emptyset$, such that $\phi(x,y) = \phi(y,x)$ and $\phi(x,x) \neq \{1,\ldots,r\}$, for all $x, y \in U$. Informally, we will view an r-type as a complete graph with a coloring of both vertices and edges using nonempty subsets of $\{1,\ldots,r\}$, where the whole set is a forbidden color on the vertices. The **sub-r-type** of K induced by $W \subseteq U$ is the r-type achieved by deleting the vertices $U - W$ from K.*

*We say that an r-graph $H = (V, c)$ **embeds in type** $K = (U,\phi)$ if there is a map $\gamma : V \to U$ such that $c(\{v,v'\}) = c_0$ implies $c_0 \in \phi(\gamma(v),\gamma(v'))$. In other words, there is a mapping γ that brings each edge of color c_0 to a vertex or an edge containing c_0 in its color set. If H embeds in type K, we write $H \mapsto K$, otherwise we write $H \not\mapsto K$. For every hereditary property $\mathcal{H}$, we let $\mathcal{K}(\mathcal{H})$ be the set of all r-types such that none of $\mathcal{F}(\mathcal{H})$ embeds in that type, i.e.,*

$$\mathcal{K}(\mathcal{H}) = \{K : K \text{ is an } r\text{-type and } H \not\mapsto K, \forall H \in \mathcal{F}(\mathcal{H})\}.$$

[1] Formally, the sizes of the partitions of the edge set should be integral, so we can take the floor function for the sizes of, say E_ρ for $\rho = 1,\ldots,r-1$ and the size of E_r is what remains. Since we fix p_ρ for $\rho = 1,\ldots,r$ and let n approach infinity, this will make no appreciable difference.

We say that an r-graph $G' = (V, c)$ ***has type*** *$K = (U, \phi)$* ***if*** *G' embeds into K with mapping $\gamma : V \to U$ and γ is surjective.*

Fact 5 generalizes the ideas underlying the simple editing algorithm in Section 2.3.

Fact 5. *If K is an r-type, G' is of type K and H does not embed into K, then $H \not\subseteq G'$.*

2.6. Editing algorithm using types.

Let $\mathbf{p} = (p_1, \ldots, p_r)$ and $\mathbf{w} = (w_1, \ldots, w_k)$ be density vectors; i.e., their entries are nonnegative and sum to 1. They play different roles, however. The vector $\mathbf{p}$ will represent a vector of densities, p_ρ. That is, the graph G has $p_\rho\binom{n}{2}$ edges of color ρ. The vector $\mathbf{w}$ will represent a vector of weights, $w_1, \ldots, w_k$, assigned to the vertices of an r-type with vertices $u_1, \ldots, u_k$, respectively.

Let $G = (V, c)$ be an r-graph with edges having densities according to the vector $\mathbf{p} = (p_1, \ldots, p_r)$, and $\mathcal{H}$ be a hereditary property. In order to find an upper bound on $\mathrm{dist}(G, \mathcal{H})$, it is sufficient to change G to an r-graph, G', such that, for all $H \in \mathcal{F}(\mathcal{H})$, H does not embed into the new coloring. In particular, if the resulting coloring has type $K \in \mathcal{K}(\mathcal{H})$, then G' is in $\mathcal{H}$.

Algorithm 6. *Fix a $K = (U, \phi) \in \mathcal{K}(\mathcal{H})$ and bring G to a coloring of type K by edge-recoloring. Let $U = \{u_1, \ldots, u_k\}$. Partition the vertices of G randomly into sets $V_1, \ldots, V_k$ such that the probability of a vertex to be in a part V_i is w_i. Consider an edge $\{x, y\}$ of G, let $x \in V_i$, $y \in V_j$, for $i, j \in \{1, \ldots, k\}$. If $c(\{x, y\}) \notin \phi(\{u_i, u_j\})$, recolor $\{x, y\}$ with a color from $\phi((u_i, u_j))$. This gives the new r-graph G' which, according to Fact 5, does not admit an embedding of any $H \in \mathcal{F}(\mathcal{H})$, thus $G' \in \mathcal{H}$.*

Note that this generalizes the simple algorithm in Section 2.3. In that algorithm, the type had restricted colorings only on the vertices (possibly of different colors) but each edge receives the color $2^{\{1,\ldots,r\}}$.

2.6.1. *Analysis of the editing algorithm.* Consider Algorithm 6 applied with type K. Let G be a graph such that the number of edges of color ρ are p_ρ for $\rho = 1, \ldots, r$. The expected number of changes is

$$\begin{aligned}
\mathbf{E}[\#\text{ changes}] &= \binom{n}{2} - \sum_{x,y\in V,\ x\neq y} \Pr(\{x, y\} \text{ is not changed}) \\
&= \binom{n}{2} - \sum_{x,y\in V,\ x\neq y} \sum_{1\leq i,j\leq k} \Pr(x \in V_i, y \in V_j)\mathbf{1}_{c(\{x,y\})\in\phi(u_i,u_j)} \\
&= \binom{n}{2} - \sum_{1\leq i,j\leq k} w_i w_j \sum_{x,y\in V,\ x\neq y} \mathbf{1}_{c(\{x,y\})\in\phi(u_i,u_j)} \\
&= \binom{n}{2} - \sum_{1\leq i,j\leq k} w_i w_j \sum_{\rho\in\phi(u_i,u_j)} p_\rho \binom{n}{2}
\end{aligned}$$

Let $\mathbf{M}_K(\mathbf{p})$ be a $k \times k$ matrix such that the (i, j)-th entry, $\mathbf{M}_K(\mathbf{p})(i, j)$ is $1 - \sum_{\rho\in\phi(u_i,u_j)} p_\rho$. Thus, if $\mathbf{w} = (w_1, \ldots, w_k)$, then

$$\mathbf{E}[\#\text{ changes}] = \mathbf{w}^T \mathbf{M}_K(\mathbf{p}) \mathbf{w} \binom{n}{2}.$$

Finally, we define two functions in terms of the matrix $\mathbf{M}_K(\mathbf{p})$:

$$f_K(\mathbf{p}) = \left(\frac{1}{k}\mathbf{1}\right)^T \mathbf{M}_K(\mathbf{p}) \left(\frac{1}{k}\mathbf{1}\right) \qquad \text{and} \qquad g_K(\mathbf{p}) = \left\{ \begin{array}{lrcl} \min & \mathbf{w}^T \mathbf{M}_K(\mathbf{p}) \mathbf{w} & & \\ \text{s.t.} & \mathbf{w}^T \mathbf{1} & = & 1 \\ & \mathbf{w} & \geq & \mathbf{0} \end{array} \right.$$

The f and g functions can be interpreted as follows: If the vertices of an r-graph, G, are assigned randomly to parts corresponding to the vertices of K, then $f_K(\mathbf{p})$ and $g_K(\mathbf{p})$ represent the expectation of the proportion of times that the color of an edge does not map the set of colors in a corresponding vertex or an edge of K. The function $f_k(\mathbf{p})$ is obtained from the uniform distribution, and $g_k(\mathbf{p})$ is obtained using the optimal distribution $(w_1, \ldots, w_k)$ of the proportion of sizes of parts. Although the g function provides a better bound for $\mathrm{dist}(\mathbf{p}, \mathcal{H})$, the linearity of the f function helps prove results of $\mathrm{dist}(\mathbf{p}, \mathcal{H})$.

2.7. **Basic results on r-graphs.**

Theorem 7 summarizes some facts about the edit distance function that generalize easily from results in both and [6] or [12]. The proof is in Section 2.9.2. Fix a density vector $\mathbf{p} = (p_1, \ldots, p_r)$. Formally, the *random r-graph of density* $\mathbf{p}$, or *random r-graph* where the context is clear, is denoted $G(n, \mathbf{p})$. It is a random variable that is an $\{1, \ldots, r\}$-coloring of the edges of a labeled K_n in which each edge, e, is colored independently such that e receives color ρ with probability p_ρ.

Theorem 7. *Let $\mathcal{H}$ be a hereditary property of r-graphs. Fix an r-dimensional density vector $\mathbf{p}$. Then the limit $\mathrm{dist}(\mathbf{p}, \mathcal{H}) := \lim_{n\to\infty} \mathrm{dist}_n(\mathbf{p}, \mathcal{H})$ exists. Moreover,*

1. $\mathrm{dist}(\mathbf{p}, \mathcal{H}) = \inf_{K\in\mathcal{K}(\mathcal{H})} f_K(\mathbf{p}) = \inf_{K\in\mathcal{K}(\mathcal{H})} g_K(\mathbf{p})$;
2. *for a fixed $\epsilon > 0$, then with probability approaching 1 as $n \to \infty$,*
$$\mathrm{dist}(\mathbf{p}, \mathcal{H}) - \epsilon \le \mathrm{dist}(G(n, \mathbf{p}), \mathcal{H}) \le \mathrm{dist}(\mathbf{p}, \mathcal{H});$$
3. $\mathrm{dist}(\mathbf{p}, \mathcal{H}) = \lim_{n\to\infty} \mathbf{E}[\mathrm{dist}(G(n, \mathbf{p}), \mathcal{H})]$;
4. $\mathrm{dist}(\mathbf{p}, \mathcal{H})$ *is continuous over the domain of r-dimensional density vectors and is concave down;*[2]
5. $\mathrm{dist}(\mathbf{p}, \mathcal{H})$ *achieves its maximum,* $\mathrm{dist}(\mathcal{H})$*, at some density vector $\mathbf{p}^*_{\mathcal{H}}$ (in fact, denote the set of all such vectors $\mathbf{p}^*_{\mathcal{H}}$) and so,*
$$\mathrm{dist}(\mathcal{H}) = \lim_{n\to\infty} \mathbf{E}[\mathrm{dist}(G(n, \mathbf{p}^*_{\mathcal{H}}), \mathcal{H})];\ \textit{and}$$
6. *Both $\mathbf{p}^*_{\mathcal{H}}$ and $\mathrm{dist}(\mathcal{H})$ exist and $\mathbf{p}^*_{\mathcal{H}}$ is a convex and closed set in $[0, 1]^{r-1}$.*

Remark 8. *Note that $\mathbf{p}^*_{\mathcal{H}}$ typically consists of a single vector, but we abuse notation by denoting the set of such vectors as $\mathbf{p}^*_{\mathcal{H}}$ when the vector at which the maximum is obtained is not unique.*

Corollary 9. *Let $\mathcal{H}$ be a symmetric hereditary property; that is, one that has the property such that if the r-tuple $(a_1, \ldots, a_r)$ is in the clique spectrum of $\mathcal{H}$, then for any permutation φ of $\{1, \ldots, r\}$, the r-tuple $(a_{\varphi(1)}, \ldots, a_{\varphi(r)})$ is also in the clique spectrum. Then,*

$$\mathrm{dist}(\mathcal{H}) \le r^{-1} \left(\sum_{i=1}^{r} a_i \right)^{-1}.$$

In particular, if $\mathcal{H} = \mathrm{Forb}(H)$ such that all color classes of H are isomorphic, then $\mathrm{dist}(\mathcal{H}) \le \frac{1}{r(\chi_r(H)-1)}$.

Proof of Corollary 9. Consider an arbitrary density vector $\mathbf{p} = (p_1, \ldots, p_r)$ and without loss of generality assume that $p_1 \le \cdots \le p_r$. Choose a permutation of the a_i-s such that $a_1 \ge \cdots \ge a_r$. Let $K = (U, \phi)$ be a r-type on $\ell = \sum_{i=1}^{r}$ vertices such that $\phi(u_i, u_j) = \{1, \ldots, r\}$ if $i \ne j$ and there are exactly a_j vertices u such that $\phi(u, u) = \{1, \ldots, r\} - \{j\}$.

[2] A function $\psi(\mathbf{p})$ being concave down means for every pair of density vectors $\mathbf{p}_1, \mathbf{p}_2$ and every real number $t \in [0, 1]$, $t\mathbf{p}_1 + (1-t)\mathbf{p}_2$ is a density vector and $\psi(t\mathbf{p}_1 + (1-t)\mathbf{p}_2) \ge t\psi(\mathbf{p}_1) + (1-t)\psi(\mathbf{p}_2)$.

The off-diagonal entries of $\mathbf{M}_K(\mathbf{p})$ are zero and so it is easy to see that $f_K(p) = \ell^{-2}\sum_{i=1}^r a_i p_i$. We can use a correlation inequality such as FKG [8] to see that

$$f_K(p) = \ell^{-2}\sum_{i=1}^r a_i p_i \leq \ell^{-2} r^{-1}\left(\sum_{i=1}^r a_i\right)\left(\sum_{i=1}^r p_i\right) = r^{-1}\ell^{-1}.$$

To finish the proof observe that, in the case of $\mathcal{H} = \mathrm{Forb}(H)$, $\ell = \sum_{i=1}^r a_i = \chi_r(H) - 1$. □

2.8. Example: triangles.

Theorem 10 gives some basic results on examples of hereditary properties of r-graphs defined by triangles. The proof is in Section 2.9.3.

Theorem 10. *Let $r = 3$ and consider hereditary properties of r-graphs.*

(1) *If $\mathcal{F}$ is a family of that consists of a single monochromatic triangle, then* $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) = 1/2$.
(2) *If $\mathcal{F}$ is a family that consists of a single triangle with two edges colored 1 and one edge colored 2, then* $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) = 1/2$.
(3) *If $\mathcal{F}$ is a family that consists of two monochromatic triangles of different colors, then* $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) = 1/2$.
(4) *If $\mathcal{F}$ is a family that consists of all six bi-chromatic triangles, then* $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) = 2/3$.
(5) *If $\mathcal{F}$ is a family that consists of a single rainbow triangle, then* $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) = 1/3$.

2.9. Proofs.

2.9.1. *Proof of Theorem 3.*

The upper bound for this theorem is proven by the simple editing algorithm from Section 2.3.

For the lower bound, we apply part (1) of Theorem 7, which states that $\mathrm{dist}(\mathbf{p},\mathcal{H}) = \inf_{K\in\mathcal{K}(\mathcal{H})} f_K(\mathbf{p})$. Consider an arbitrary $K = (U,\phi) \in \mathcal{K}(\mathcal{H})$, an r-type on k vertices. Let $\tilde{K}$ be an auxiliary graph with vertex set U such that u and u' are adjacent in $\tilde{K}$ if and only if $\phi(u,u') = \{1,\ldots,r\}$. We observe that $\tilde{K}$ has no clique on χ_r vertices, otherwise for some $H \in \mathcal{F}(\mathcal{H})$, $H \mapsto K$. Using Turán's theorem, the number of edges of $\tilde{K}$ is at most $\frac{\chi_r - 2}{\chi_r - 1}\cdot\frac{k^2}{2}$.

Let $\mathbf{p} = \frac{1}{r}\mathbf{1}$. Consider the matrix $\mathbf{M}_K(\mathbf{p})$ and observe that every entry is either zero or is a positive integer multiple of $1/r$. The zero entries correspond exactly to pairs with ϕ value equal to $\{1,\ldots,r\}$. Thus, this matrix $\mathbf{M}_K(\mathbf{p})$ has at least $k^2 - 2\left(\frac{\chi_r-2}{\chi_r-1}\cdot\frac{k^2}{2}\right) \geq \frac{k^2}{\chi_r-1}$ entries with value at least $1/r$. Therefore, $f_K(\mathbf{p}) = \frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(\mathbf{p})\mathbf{1}$ is at least $\frac{1}{r(\chi_r-1)}$. Since K was arbitrary, this gives a lower bound for $\mathrm{dist}(\mathbf{p},\mathcal{H})$.

2.9.2. *Proof of Theorem 7.*

Let $f(\mathbf{p}) = \inf_{K\in K(\mathcal{H})} f_K(\mathbf{p})$ and let $g(\mathbf{p}) = \inf_{K\in K(\mathcal{H})} g_K(\mathbf{p})$.

A: Upper bound on $\mathrm{dist}(\mathbf{p},\mathcal{H})$.
Let G be an r-graph with the density of its i-th color class be p_ρ for $\rho = 1,\ldots,r$. Let $K \in \mathcal{K}(\mathcal{H})$. Apply the editing algorithm in Section 2.6 to G using K. The analysis of the algorithm in Section 2.6.1 gives that the expected number of changes is $f_K(\mathbf{p})\binom{n}{2}$ and so $\mathrm{dist}_n(\mathbf{p},\mathcal{H}) \leq f(\mathbf{p})\binom{n}{2}$.

B: Equality of f and g.
By the definition of $g_K(\mathbf{p})$, it is easy to see that $g_K(\mathbf{p}) \leq f_K(\mathbf{p})$ for every density vector $\mathbf{p}$. Therefore, $g(\mathbf{p}) \leq f(\mathbf{p})$. For the other direction, we will use K and its optimal weight vector $\mathbf{w}^* = \{w_1,\ldots,w_k\}$,

where w_i corresponds to $v_i \in V(K)$ in order to construct a sequence of CRGs, $\{K_\ell\}$ such that $\lim_{\ell\to\infty} f_{K_\ell}(\mathbf{p}) = g_K(\mathbf{p})$.

First, choose ℓ large enough to ensure that $w_i\ell \geq 2$ for $i = 1, \ldots, k$. Then, for each vertex $u_i \in V(K)$, create $\lfloor w_i\ell \rfloor$ copies of u_i in the following sense: Let u_i' and u_j'' be copies of u_i and u_j, respectively, where $u_i, u_j \in V(K)$. Let ϕ be the coloring function of K and ϕ' be the coloring function of K_ℓ. If $i \neq j$, then $\phi'(u_i', u_j'') = \phi(v_i, v_j)$. If $i = j$ and $v_i' \neq v_i''$, then $\phi'(u_i', u_i'') = \phi(v_i, v_i)$. Finally, $\phi'(v_i', v_i') = \phi(v_i, v_i)$.

The (i,j)-th block is a $\lfloor w_i\ell \rfloor \times \lfloor w_j\ell \rfloor$ matrix and each entry of the (i,j)-th block is the same as the (i,j)-th entry of $\mathbf{M}_K(\mathbf{p})$.

If we denote the (i,j)-th entry of $\mathbf{M}_K(\mathbf{p})$ by m_{ij}, then

$$\begin{aligned}
f_{K_\ell}(\mathbf{p}) &= \frac{1}{|V(K)|^2}\mathbf{1}^T\mathbf{M}_{K_\ell}(\mathbf{p})\mathbf{1} &&= \left(\sum_i \lfloor w_i\ell \rfloor\right)^{-2} \sum_{i,j} m_{ij}\lfloor w_i\ell \rfloor \lfloor w_j\ell \rfloor \\
&\leq \ell^2 \left(\sum_i \lfloor w_i\ell \rfloor\right)^{-2} \sum_{i,j} m_{ij} w_i w_j &&= \ell^2 \left(\sum_i \lfloor w_i\ell \rfloor\right)^{-2} g_K(\mathbf{p}) \\
&\leq \ell^2 \left(\sum_i (w_i\ell - 1)\right)^{-2} g_K(\mathbf{p}) &&= \frac{\ell^2}{(\ell-k)^2} g_K(\mathbf{p}).
\end{aligned}$$

Taking $\ell \to \infty$, we see that $\lim_{\ell\to\infty} f_{K_\ell}(\mathbf{p}) \leq g_K(\mathbf{p})$. Consequently, for any $K \in \mathcal{K}(\mathcal{H})$,

$$f(\mathbf{p}) = \inf_{\tilde{K}\in\mathcal{K}(\mathcal{H})} f_{\tilde{K}}(\mathbf{p}) \leq \lim_{\ell\to\infty} f_{K_\ell}(\mathbf{p}) \leq g_K(\mathbf{p}).$$

Take the infimum over all $K \in \mathcal{K}(\mathcal{H})$, and we have that $f(\mathbf{p}) \leq g(\mathbf{p})$.

C: Lower bound on $\mathrm{dist}(\mathbf{p}, \mathcal{H})$ using the random graph.
We apply Theorem 12, which is given in [5]. Theorem 12 is a corollary of Theorem 11, a relatively straightforward generalization to r-graphs and digraphs of a theorem by Alon, Fischer, Krivelevich and M. Szegedy [1], which is suitable for induced graphs.

In an r-graph, the *density vector of a pair of disjoint sets of vertices* (V_i, V_j) is simply $\mathbf{d}(V_i, V_j) := (d_1(V_i, V_j), \ldots, d_r(V_i, V_j))$. So we can state the general version of the regularity lemma. For all definitions of regularity, see [1].

Theorem 11 (Alon, et al. [1])**.** *Fix $r \geq 2$. For every m and function $\mathcal{E}$ with $\mathcal{E} : \mathbb{N} \to (0,1)$, there exist $S = S_{11}(r, m, \mathcal{E})$ and $\delta = \delta_{11}(r, m, \mathcal{E})$ with the following property:*

If G is a graph [r-graph, digraph] with $n \geq S$ vertices then there exist an equipartition $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$ of G and an induced subgraph [induced r-subgraph, induced subdigraph] G' of G, with an equipartition $\mathcal{A}' = \{V_i' : 1 \leq i \leq k\}$ of the vertices of G' that satisfy:

- *$S \geq k \geq m$.*
- *$V_i' \subset V_i$ for all $i \geq 1$, and $|V_i'| \geq \delta n$.*
- *In the equipartition $\mathcal{A}'$, **all** pairs are $\mathcal{E}(k)$-regular.*
- *All but at most $\mathcal{E}(0)\binom{k}{2}$ of the pairs $1 \leq i < i' \leq k$ are such that $\|\mathbf{d}(V_i, V_{i'}) - \mathbf{d}(V_i', V_{i'}')\|_\infty < \mathcal{E}(0)$.*

We use Theorem 11 in order to prove Theorem 12, which is the result that we need.

Theorem 12 ([5])**.** *Let G' be an r-graph in hereditary property $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$ and $\mathbf{p} = (p_1, \ldots, p_r)$ be a density vector. Then, there exists an r-type $K \in \mathcal{K}(\mathcal{H})$ such that $H \not\mapsto K$ for all $H \in \mathcal{F}(\mathcal{H})$ and with probability going to 1 as $n \to \infty$, $\mathrm{dist}(G_{n,\mathbf{p}}, \mathcal{H}) \geq f_K(\mathbf{p})\binom{n}{2} - o(n^2)$.*

The proof of Theorem 12 from Theorem 11 is straightforward and the details are given in [5]. We begin with G distributed according to $G(n,\mathbf{p})$ and typical in the sense that any Szemerédi partition will have every pair $n^{-0.4}$-regular. Let G' be the graph of smallest distance from G and apply Theorem 11. The resulting partition $\mathcal{A}'$ describes a type K which must be in $\mathcal{K}(\mathcal{H})$. Furthermore, the number of changes required to ensure that G' has partition $\mathcal{A}$ is very close to $f_K(\mathbf{p})$ because almost every pair in $\mathcal{A}$ has the same density as in $\mathcal{A}'$.

Using part A, we see that for any $\epsilon > 0$, with probability approaching 1 as $n \to \infty$,

$$f(\mathbf{p}) - \epsilon/2 \leq \mathrm{dist}(G(n,\mathbf{p}),\mathcal{H}) \leq \mathrm{dist}(\mathbf{p},\mathcal{H}) \leq f(\mathbf{p}). \tag{1}$$

We can now combine A, B and C. Take the limit of (1) as $n \to \infty$, and we obtain that for all $\epsilon > 0$, $f(\mathbf{p}) - \epsilon/2 \leq \mathrm{dist}(\mathbf{p},\mathcal{H}) \leq f(\mathbf{p})$. Hence, $\mathrm{dist}(\mathbf{p},\mathcal{H}) = f(\mathbf{p}) = g(\mathbf{p})$. Moreover, we can replace the second term with $\mathbf{E}[\mathrm{dist}(G(n,\mathbf{p}),\mathcal{H})]$ because that random variable is bounded (in $[0,1]$) and so (1) occurring with high probability implies that the random variable is concentrated around its mean, which approaches $\mathrm{dist}(\mathbf{p},\mathcal{H})$. This verifies parts (1), (2) and (3) of the theorem.

D: Continuity of f.
Because the set of r-types is countable, we can linearly order $\mathcal{K}(\mathcal{H})$ to be $K_1, K_2, \ldots$. For every density vector $\mathbf{p}$, set $m_\ell(\mathbf{p}) = \min_{i\leq\ell} f_{K_i}(\mathbf{p})$.

We want to show that each function m_ℓ is Lipschitz with coefficient 1 with respect to the L^1 metric. Let $\mathbf{p} = (p_1,\ldots,p_\rho)$ and $\mathbf{q} = (q_1,\ldots,q_\rho)$ be density vectors and define r-types $K_\mathbf{p}, K_\mathbf{q} \in \{K_1,\ldots,K_\ell\}$ on $k_\mathbf{p}, k_\mathbf{q}$ vertices, respectively, such that $m_\ell(\mathbf{p}) = f_{K_\mathbf{p}}(\mathbf{p})$ and $m_\ell(\mathbf{q}) = f_{K_\mathbf{q}}(\mathbf{q})$. Then, using the matrix definition of f and the definition of m_ℓ as a minimum of linear functions,

$$\begin{aligned} f_{K_\mathbf{p}}(\mathbf{p}) - f_{K_\mathbf{p}}(\mathbf{q}) \leq \quad & f_{K_\mathbf{p}}(\mathbf{p}) - f_{K_\mathbf{q}}(\mathbf{q}) && \leq f_{K_\mathbf{q}}(\mathbf{p}) - f_{K_\mathbf{q}}(\mathbf{q}) \\ \left(\frac{1}{k_\mathbf{p}}\mathbf{1}\right)^T \mathbf{M}_{K_\mathbf{p}}(\mathbf{p}-\mathbf{q})\left(\frac{1}{k_\mathbf{p}}\mathbf{1}\right) \leq \quad & f_{K_\mathbf{p}}(\mathbf{p}) - f_{K_\mathbf{q}}(\mathbf{q}) && \leq \left(\frac{1}{k_\mathbf{q}}\mathbf{1}\right)^T \mathbf{M}_{K_\mathbf{q}}(\mathbf{p}-\mathbf{q})\left(\frac{1}{k_\mathbf{q}}\mathbf{1}\right) \end{aligned}$$

Since each of the entries in matrices $\mathbf{M}_{K_\mathbf{p}}$ and $\mathbf{M}_{K_\mathbf{q}}$ is between zero and one, and the number of entries in these matrices is $k_\mathbf{p}^2$ and $k_\mathbf{q}^2$, respectively, it is the case that

$$\left|f_{K_\mathbf{p}}(\mathbf{p}) - f_{K_\mathbf{q}}(\mathbf{q})\right| \leq \|\mathbf{p}-\mathbf{q}\|_1.$$

Since $\{m_\ell\}_{\ell\geq1}$ is Lipschitz, Definition 7.22 from Rudin [15] says that the sequence of functions is equicontinuous. The sequence is also pointwise bounded above by $f_{K_1}(\mathbf{p})$ and below by 0. By Theorem 7.25(b) from [15] the sequence $\{m_\ell\}_{\ell\geq1}$ has a uniformly convergent subsequence. Since $\{m_\ell\}_{\ell\geq1}$ is an equicontinuous, each member is itself continuous. Theorem 7.12 from [15] gives that the aforementioned uniformly convergent subsequence has a continuous limit. The monotonicity of $\{m_\ell\}_{\ell\geq1}$ gives that the limit of any subsequence is the same as the pointwise limit of the sequence itself, namely $\lim_{\ell\to\infty} m_\ell = \inf_{K\in\mathcal{K}(\mathcal{H})} f_K = \mathrm{dist}(\mathcal{H})$.

E: Concavity.
Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be density vectors and $t \in [0,1]$ be a real number. Observe that $t\mathbf{p}_1 + (1-t)\mathbf{p}_2$ is still

a density vector and, hence, in the domain. Furthermore,

$$\begin{aligned} f(t\mathbf{p}_1+(1-t)\mathbf{p}_2) &= \inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(t\mathbf{p}_1+(1-t)\mathbf{p}_2)\} \\ &= \inf_{K\in\mathcal{K}(\mathcal{H})}\{tf_K(\mathbf{p}_1)+(1-t)f_K(\mathbf{p}_2)\} \\ &\geq t\left(\inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(\mathbf{p}_1)\}\right)+(1-t)\left(\inf_{K\in\mathcal{K}(\mathcal{H})}\{f_K(\mathbf{p}_2)\}\right) \\ &= tf(\mathbf{p}_1)+(1-t)f(\mathbf{p}_2). \end{aligned}$$

This gives concavity.

Using D and E, we obtain part (4) directly and the fact that $g_{\mathcal{H}}$ achieves its maximum follows from continuity (and compactness) and Theorem 4.16 from [15]. Let S be the set of density vectors $\mathbf{p}$ such that $\mathrm{dist}(\mathbf{p},\mathcal{H})=\mathrm{dist}(\mathcal{H})$. The set S must be convex set, because if $\mathrm{dist}(\mathbf{p}_1,\mathcal{H})=\mathrm{dist}(\mathbf{p}_2,\mathcal{H})=\mathrm{dist}(\mathcal{H})$, then by continuity and concavity, the line segment that connects $\mathbf{p}_1$ and $\mathbf{p}_2$ must consist of vectors in S. The set S must be closed because a corollary to Theorem 4.8 from [15] says that, under a continuous mapping, the inverse image of a closed set is closed. Since $\mathrm{dist}(\mathbf{p},\mathcal{H})$ is a continuous function and S is the inverse image of the closed set, $\{\mathrm{dist}(\mathcal{H})\}$, then S is closed. This verifies parts (4), (5) and (6) of the theorem and concludes the proof. □

2.9.3. *Proof of Theorem 10.*

(1) In order to destroy all copies of a monochromatic 1-colored triangle in an arbitrary coloring of K_n, it is sufficient to split the vertex set into two parts and recolor all edges within these parts in color 2. This requires at most $\frac{1}{2}\binom{n}{2}$ changes. To see the lower bound, consider K_n with all edges colored 1. After all editing is done, in the resulting coloring color class 1 is triangle-free, having at most $\frac{n^2}{4}$ edges. Thus, at least $\frac{n^2}{4}=\frac{1}{2}\binom{n}{2}+o(n^2)$ edges must have been changed.

(2) In order to destroy all such triangles, it suffices to equipartition the vertex set into two parts and recolor all edges within these parts to color 3. This requires at most $\frac{1}{2}\binom{n}{2}$ changes. To see the lower bound, consider K_n on vertex set with equipartition $V_1\cup V_2$. Let all edges between V_1 and V_2 be colored 1 and let all edges within parts V_i, $i=1,2$ be colored 2. We may assume that the only editing operations are recoloring an edge of color 1 into color 3 and recoloring an edge of color 2 into color 3 because this editing will never create a forbidden triangle. Let c be such a recoloring not containing triangles with two edges of color 1 and one edge of color 2. Let G be an auxiliary graph corresponding to edges of color 3 in this coloring. The complement of G can not have any triangles with vertices in both V_1 and V_2. It is easy to prove by induction on n that a graph with satisfying such a condition could have at most $\frac{1}{2}\binom{n}{2}$ edges. Therefore G has at least $\frac{1}{2}\binom{n}{2}$ edges, and this corresponds to the number of changes made.

(3) Assume that $\mathcal{F}$ consists of a triangle with all edges colored 1 and of a triangle with all edges colored 2. In order to destroy both of these triangles in an any coloring, as in the previous case, it is sufficient to equipartition the vertex set into two parts and recolor all edges within these parts in color 3. This requires at most $\frac{1}{2}\binom{n}{2}$ changes.

As to the lower bound, fix $\mathbf{p}=(1/2,1/2,0)$ and consider a 3-type, $K\in\mathcal{K}(\mathcal{H})$, on k vertices. Each of the vertices must have color 3. By Turán's theorem, at least $\binom{k}{2}-\lfloor k^2/4\rfloor=\lceil (k^2-2k)/4\rceil$ edges cannot have color 1 and at least $\lceil (k^2-2k)/4\rceil$ edges cannot have color 2. Hence, if we consider the off-diagonal entries of $\mathbf{M}_K(\mathbf{p})$, the sum is at least $\frac{1}{2}\lceil (k^2-2k)/4\rceil+\frac{1}{2}\lceil (k^2-2k)/4\rceil$. So, for any such

K,

$$f_K(p) \geq \frac{1}{k^2}\left[k+2\left\lceil\frac{k^2-2k}{4}\right\rceil\right] \geq \frac{1}{2}.$$

As a result, $\inf_{K\in\mathcal{K}(\mathcal{H})} f_K(\mathbf{p}) \geq 1/2$.

(4) It is suffices to recolor edges of colors 1 or 2 into color 3. As a result, all forbidden colored triangles will be destroyed via at most $\frac{2}{3}\binom{n}{2}$ changes. In fact, for fixed $\mathbf{p} = (p_1, p_2, p_3)$, at most $(1-\max\{p_1,p_2,p_3\})\binom{n}{2}$ changes suffice.

To see the lower bound, consider a 3-type $K \in \mathcal{K}(\mathcal{H})$ on k vertices. The vertices must be monochromatic and, in addition, the edges incident to a vertex must share the color of that vertex. Otherwise, there would be a bichromatic triangle H with $H \mapsto K$. This implies, however, that K must be entirely monochromatic. Hence, $g_K(\mathbf{p}) \geq 1-\max\{p_1,p_2,p_3\}$.

Note the this determines not only $\mathrm{dist}(\mathcal{H})$, but the entire function $\mathrm{dist}(\mathbf{p},\mathcal{H}) = 1-\max\{p_1,p_2,p_3\}$.

(5) Observe that in order to destroy all rainbow triangles using colors 1, 2 and 3, it is sufficient to edit the smallest of these color classes, thus performing at most a $\min\{p_1,p_2,p_3\}$ proportion of changes.

For the lower bound, simply observe that no edge in any $K \in \mathcal{K}(\mathcal{H})$ can be trichromatic. Otherwise, that edge, together with any vertex to which it is incident admits a mapping of a rainbow triangle. Hence, each entry of $M_K(p)$ is at least $\min\{p_1,p_2,p_3\}$ and so $f_K(p) \geq \min\{p_1,p_2,p_3\}$. Hence $\mathrm{dist}(\mathbf{p},\mathcal{H}) = \min\{p_1,p_2,p_3\}$ and $\mathrm{dist}(\mathcal{H}) = 1/3$. □

3. Directed graphs

3.1. Basic definitions.

We give a number of definitions that are similar to the case of r-graphs, however, there are some important distinctions.

Definition 13. *A* ***simple directed graph*** *or* ***digraph*** *is defined to be a pair (V,E) where V is a labeled vertex set, $E \subseteq (V)_2$ and $(V)_2$ denotes the set $V \times V - \{(v,v) : v \in V\}$. We will also view this as a coloring; that is, a digraph is a pair (V,c) where $c : (V)_2 \to \{\bigcirc, -, \leftarrow, \rightarrow\}$ is a function which has the property that, for distinct v, w,*

- *$c(v,w) = c(w,v)$ if and only if $c(v,w) \in \{\bigcirc, -\}$ and*
- *$c(v,w) =\rightarrow$ if and only if $c(w,v) =\leftarrow$.*

Let $\overleftrightarrow{\mathcal{A}} := \{\bigcirc, -, \leftarrow, \rightarrow\}$. Here we interpret the color $c(v,w) = \bigcirc$ to mean that neither (v,w) nor (w,v) are in E, the color $c(v,w) = -$ to mean that both (v,w) and (w,v) are in E and the color $c(v,w) =\rightarrow$ to mean that $(v,w) \in E$ and $(w,v) \notin E$.

For any digraph G on fixed vertex set $\{v_1,\ldots,v_n\}$, disjoint vertex sets V_i and V_j and color ρ, $\rho \in \overleftrightarrow{\mathcal{A}}$, the expression $E_\rho(V_i)$ denotes the set of pairs $\{v_i, v_i'\}$ with $v_i, v_i' \in V_i$, $i < i'$ and $c(v_i, v_i') = \rho$. The expression $E_\rho(V_i,V_j)$ denotes the set of pairs $\{v_i, v_j\}$ with $v_i \in V_i$, $v_j \in V_J$ and $c(v_i,v_j) = \rho$. Hence, $E_\leftarrow(V_i,V_j) = E_\rightarrow(V_j,V_i)$. As it happens, we will be able to assume, as in the proof of Theorem 7, that our graphs are random. We will also be able to assume that, among the pairs that have directed edges, a $\leftarrow$ is as likely as $\rightarrow$. Hence, we will postpone the definition of a density vector for directed graphs.

Definition 14. *We say that $\mathcal{P} \subseteq \overleftrightarrow{\mathcal{A}}$ is a* ***palette*** *if either none or both of "$\rightarrow$" and "$\leftarrow$" are in $\mathcal{P}$. There are 5 possible nontrivial palettes:*

(0) $\mathcal{P}_0 = \overleftrightarrow{\mathcal{A}}$ *is the most general case.*
(1) $\mathcal{P}_{\rm compl} = \{-, \leftarrow, \rightarrow\}$ *is the case of simple digraphs such that every pair of vertices has at least one arc between them.*
(2) $\mathcal{P}_{\rm orien} = \{\bigcirc, \leftarrow, \rightarrow\}$ *is the case of oriented graphs; that is, no pair of vertices has two arcs between them.*
(3) $\mathcal{P}_{\rm undir} = \{\bigcirc, -\}$ *is the case of simple, undirected graphs.*
(4) $\mathcal{P}_{\rm tourn} = \{\leftarrow, \rightarrow\}$ *is the case of tournaments.*

The palette is the universe in which the editing takes place. That is, if $\bigcirc$ is not in the palette, then no pair (v, w) can be changed to color $\bigcirc$ in the editing process.

If $\mathcal{P}$ is a fixed palette and $G = (V, c)$ and $G' = (V, c')$ are digraphs with colors in $\mathcal{P}$, then $\mathrm{dist}(G, G')$ is the proportion of edges on which the colors differ; i.e., the number of edges on which the colors differ, divided by $\binom{n}{2}$.[3] We may call this the *normalized edit distance* between G and G'. For any property $\mathcal{H}$, a simple digraph G with all edge-colors in palette $\mathcal{P}$, an integer n, we define $\mathrm{dist}(G, \mathcal{H})$, $\mathrm{dist}(n, \mathcal{H})$, and $\mathrm{dist}(\mathcal{H})$ similarly to the multicolor case.

A *hereditary property of digraphs with respect to palette* $\mathcal{P}$ (or, simply, *hereditary property*, where the context is understood) is a set of digraphs with all edge-colors in $\mathcal{P}$ that is closed under vertex-deletion and isomorphisms. Let a digraph G' be an *induced digraph* of G if G' can be obtained from G by vertex-deletion. For a fixed palette, $\mathcal{P}$ and a digraph, H, the family $\mathrm{Forb}(H)$ (the palette will be understood) consists of all digraphs with edge-colors in $\mathcal{P}$ that have no (induced) copies of H. For every palette $\mathcal{P}$ and every hereditary property $\mathcal{H}$ with respect to $\mathcal{P}$, there is a family, $\mathcal{F}(\mathcal{H})$, of digraphs such that $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$. If $\mathcal{F}$ is a family of digraphs, then we use $\mathrm{Forb}(\mathcal{F})$ to denote $\bigcap_{H \in \mathcal{F}} \mathrm{Forb}(H)$.

3.2. **The directed chromatic number.**

Definition 15. *For a hereditary property $\mathcal{H} = \cap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$ and a palette $\mathcal{P}$, a **good triple** (a_0, a_1, a_2) is a triple of non-negative integers such that for some $H \in \mathcal{F}(\mathcal{H})$, the vertex set $V(H)$ can be partitioned into sets S_0, S_1, S_2 such that, for each $i \in \{0, 1, 2\}$ with $a_i \neq 0$, the partition can be further refined $S_i = V_{i,1} \cup \cdots \cup V_{i,a_i}$ and*

- *each $V_{0,j}$ does not induce a nonedge (i.e., does not induce an edge of color $\bigcirc$),*
- *each $V_{1,j}$ ensures that the directed edges induced by $V_{1,j}$ form an acyclic digraph, and*
- *each $V_{2,j}$ does not induce a bidirectional edge (i.e., does not induce an edge of color $-$).*

*The **clique spectrum** of $\mathcal{H}$ with respect to a palette $\mathcal{P}$ is the set of all triples (a_0, a_1, a_2) that are NOT good and such that $a_0 = 0$ if $\bigcirc \notin \mathcal{P}$, $a_1 = 0$ if $\{\rightarrow, \leftarrow\} \cap \mathcal{P} = \emptyset$, $a_2 = 0$ if $- \notin \mathcal{P}$. The **directed chromatic number**, $\chi_{\rm dir}^{\mathcal{P}}(\mathcal{H})$, of $\mathcal{H}$ with respect to a palette $\mathcal{P}$ is the maximum $\ell + 1$ such that for some non-negative integers a_0, a_1, a_2, with $a_0 + a_1 + a_2 = \ell$, the triple (a_0, a_1, a_2) is in the clique spectrum of $\mathcal{H}$. We merely use $\chi_{\rm dir}(\mathcal{H})$ for the directed chromatic number if the palette is understood.*

If the palette is $\mathcal{P}_{\rm undir} = \{\bigcirc, -\}$, then $\chi_{\rm dir}(\mathcal{H})$ is merely the binary chromatic number of hereditary property $\mathcal{H}$. If $\mathcal{H} = \mathrm{Forb}(H)$ and the palette is $\mathcal{P}_{\rm tourn} = \{\leftarrow, \rightarrow\}$, the case of tournaments, then $\chi_{\rm dir}(H)$ is the fewest number of transitive subtournaments into which $V(H)$ can be partitioned. Note that monotonicity holds for the clique spectrum here as well. If (a_0, a_1, a_2) is in a clique spectrum and $a_i' \leq a_i$ for $i = 0, 1, 2$, then (a_0', a_1', a_2') is in that clique spectrum.

3.3. **A simple editing algorithm.**

Let $\mathcal{P}$ be a palette and let $\mathcal{H}$ be a hereditary property of digraphs such that $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$

[3] Here, we can talk about pairs because the color of the pair (v, w) determines the color of the pair (w, v).

and each edge of each $H \in \mathcal{F}(\mathcal{H})$ has a color in $\mathcal{P}$. Further, let $\ell = \chi_{\mathrm{dir}}(\mathcal{H}) - 1$ and (a_0, a_1, a_2) be in the clique spectrum and $\sum_{i=0}^{2} a_i = \ell$. Recall that if a color is not in the palette, then its corresponding a_i value must be set to zero.

Partition V into 3 sets, S_0, S_1, S_2 and further refine the partition such that $S_i = V_{i,1} \cup \cdots \cup V_{i,a_i}$, for $i = 0, 1, 2$ and then recolor the edges induced by each $V_{i,j}$ as follows:

- If $i = 0$, then recolor the edges colored $\bigcirc$ into some other arbitrary color in the palette.
- If $i = 1$, then recolor the edges $\leftarrow$ and $\rightarrow$ so that there are no directed cycles among those directed edges.
- If $i = 2$, then recolor the edges colored $-$ into some other arbitrary color in the palette.

This new coloring does not contain any $H \in \mathcal{F}(\mathcal{H})$, otherwise the triple (a_0, a_1, a_2) would be good for some H. As in the multicolor case, if the partition into sets $V_{i,j}$ is an equipartition, then

$$\mathrm{dist}(\mathcal{H}) \leq \frac{1}{\ell} = \frac{1}{\chi_{\mathrm{dir}}(\mathcal{H}) - 1}.$$

3.4. **Main results.**

In Section 2, we have seen a general bound in the r-graph case. Here, we show a similar result in the directed case.

Theorem 16. *Let $\mathcal{H}$ be a hereditary property of digraphs. Let $\chi_{\mathrm{dir}}^{\mathcal{P}} = \chi_{\mathrm{dir}}^{\mathcal{P}}(\mathcal{H})$ be the directed chromatic number of $\mathcal{H}$. Then,*

(0) $\dfrac{1}{4(\chi_{\mathrm{dir}}^{\mathcal{P}} - 1)} \leq \mathrm{dist}(\mathcal{H}) = \dfrac{1}{\chi_{\mathrm{dir}}^{\mathcal{P}} - 1}$, *if* $\mathcal{P} = \mathcal{P}_0 = \{\bigcirc, \leftarrow, \rightarrow, -\}$.

(1) $\dfrac{1}{3(\chi_{\mathrm{dir}}^{\mathcal{P}} - 1)} \leq \mathrm{dist}(\mathcal{H}) = \dfrac{1}{\chi_{\mathrm{dir}}^{\mathcal{P}} - 1}$, *if* $\mathcal{P} = \mathcal{P}_{\mathrm{compl}} = \{-, \leftarrow, \rightarrow\}$.

(2) $\dfrac{1}{3(\chi_{\mathrm{dir}}^{\mathcal{P}} - 1)} \leq \mathrm{dist}(\mathcal{H}) = \dfrac{1}{\chi_{\mathrm{dir}}^{\mathcal{P}} - 1}$, *if* $\mathcal{P} = \mathcal{P}_{\mathrm{orien}} = \{\bigcirc, \leftarrow, \rightarrow\}$.

(3) $\dfrac{1}{2(\chi_{\mathrm{dir}}^{\mathcal{P}} - 1)} \leq \mathrm{dist}(\mathcal{H}) = \dfrac{1}{\chi_{\mathrm{dir}}^{\mathcal{P}} - 1}$, *if* $\mathcal{P} = \mathcal{P}_{\mathrm{undir}} = \{\bigcirc, -\}$.

(4) $\mathrm{dist}(\mathcal{H}) = \dfrac{1}{2(\chi_{\mathrm{dir}}^{\mathcal{P}} - 1)}$, *if* $\mathcal{P} = \mathcal{P}_{\mathrm{tourn}} = \{\leftarrow, \rightarrow\}$.

We prove Theorem 16 in Section 3.10. As in the multicolor case, the upper bound is a consequence of the simple editing algorithm. The lower bound comes from Theorem 21, stated below, is the digraph version of Theorem 7 and deals with computing the edit distance for given hereditary properties of digraphs. In order to do so, we need to investigate the so-called edit distance function, which computes the edit distance of a digraph such that nonedges, directed edges and undirected edges having a specified density.

3.5. **The edit distance function.**

3.5.1. *Preliminary definitions.* For a digraph, $G = (V, c)$ with $c : (V)_2 \to \overleftrightarrow{\mathcal{A}}$ and c having the required symmetries as in Definition 13, partition $(V)_2$ as follows:

- $E_{\bigcirc}(G)$ is the set of all unordered pairs $\{v, w\}$ such that $c(v, w) = \bigcirc$,
- $E_{\leftarrow}(G)$ is the set of all ordered pairs (v, w) such that $c(v, w) = \leftarrow$,
- $E_{\rightarrow}(G)$ is the set of all ordered pairs (v, w) such that $c(v, w) = \rightarrow$,
- $E_{-}(G)$ is the set of all unordered pairs $\{v, w\}$ such that $c(v, w) = -$,

The definition of a density vector in the r-graph case does not translate well to the directed case because of the asymmetry that results from directed edges, so we have a new definition.

Given a palette, $\mathcal{P}$, A *directed density vector* (p,q) *with respect to* $\mathcal{P}$ (or, simply, *density vector* or *probability vector* where the context is understood) is a nonnegative real vector with the property that $p + 2q \leq 1$. Furthermore,

(1) If $\mathcal{P} = \mathcal{P}_{\mathrm{compl}} = \{-, \leftarrow, \rightarrow\}$, then $p + 2q = 1$.
(2) If $\mathcal{P} = \mathcal{P}_{\mathrm{orien}} = \{\bigcirc, \leftarrow, \rightarrow\}$, then $p = 0$ and $q \leq 1/2$.
(3) If $\mathcal{P} = \mathcal{P}_{\mathrm{undir}} = \{\bigcirc, -\}$, then $q = 0$ and $p \leq 1$. This is the r-graph case where $r = 2$ or simply the case of undirected graphs. See [3] and [4].
(4) If $\mathcal{P} = \mathcal{P}_{\mathrm{tourn}} = \{\leftarrow, \rightarrow\}$, then $p = 0$ and $1 - p - 2q = 0$, so $q = 1/2$.

For any density vector (p,q), and an integer n, we denote[4]

$$\mathrm{dist}_n((p,q),\mathcal{H}) = \max\left\{\mathrm{dist}(G,\mathcal{H}) : \begin{array}{l} |V(G)| = n, |E_-(G)| = p\binom{n}{2}, |E_\rightarrow(G)| = q\binom{n}{2}, \\ |E_\leftarrow(G)| = q\binom{n}{2} \text{ and } |E_\bigcirc(G)| = (1-p-2q)\binom{n}{2} \end{array}\right\}.$$

Observe that there are, in fact, four densities here; two are equal and all sum to one. Thus, we only need two parameters and we choose to only use two, in part because the case of $q = 0$ gives the classical case of undirected graphs, as we see below. Later in the paper, we show that the following limit exits, which we call the *edit distance function*:

$$\mathrm{dist}((p,q),\mathcal{H}) = \lim_{n\to\infty} \mathrm{dist}_n((p,q),\mathcal{H}).$$

Having the edit distance function, we see that $\mathrm{dist}(\mathcal{H}) = \max_{(p,q)} \mathrm{dist}((p,q),\mathcal{H})$, where the maximum is taken over all density vectors that are valid under the conditions imposed by the palette.

3.5.2. *Types of colorings.* In Section 3.6.1, we define two functions which are described in terms of dir-types, which allow us to compute the edit distance function. Later in the paper, we shall provide algorithms for such computing.

Definition 17. *For a palette $\mathcal{P}$, a $\mathcal{P}$-**dir-type** (or **dir-type** or **type**, where the context and the palette are understood), K, is a pair (U,ϕ), where U is a finite set of vertices and $\phi : U \times U \to 2^{\mathcal{P}} \setminus \emptyset$, such that*

- *for distinct x, y and $a \in \{\bigcirc, -\}$, $\phi(x,y) \ni a$ if and only if $\phi(y,x) \ni a$ and*
- *for distinct x, y, $\phi(x,y) \ni\rightarrow$ if and only if $\phi(y,x) \ni\leftarrow$ and*
- *$\phi(x,x) \neq \mathcal{P}$.* [5]

*The **sub-dir-type** of K induced by $W \subseteq U$ is the dir-type achieved by deleting the vertices $U - W$ from K.*

*We say that a digraph $H = (V,c)$ **embeds in type** $K = (U,\phi)$ if there is a map $\gamma : V \to U$ such that for all vertices $v \neq v'$,*

- *if $\gamma(v) \neq \gamma(v')$, then $c(v,v') \in \phi(\gamma(v),\gamma(v'))$,*
- *if $c_0 \in \{\bigcirc, -\}$ and $c_0 \notin \phi(u,u)$, then $\gamma^{-1}(u)$ has no pair with color c_0,*
- *if $\{\leftarrow,\rightarrow\} \cap \phi(u,u) = \emptyset$, then $\gamma^{-1}(u)$ has no directed edge, and*
- *if $|\{\leftarrow,\rightarrow\} \cap \phi(u,u)| = 1$, then $\gamma^{-1}(u)$ has no directed cycle.*

In other words, there is a mapping γ that brings each edge of color c_0 to a vertex or an edge containing c_0 in its color set, except that if a vertex contains exactly one of $\{\leftarrow,\rightarrow\}$ then the pre-image of that vertex can be ordered transitively with respect to the oriented edges. If H embeds in type K, we write

[4] Formally, the sizes of the partitions of the edge set should be integral, so we can take the floor function for the sizes of, say E_-, $E_\leftarrow$, $E_\rightarrow$ and the size of $E_\bigcirc$ is what remains. Since we fix p and q and let n approach infinity, this will make no appreciable difference.

[5] Note that it is possible that $|\{\leftarrow,\rightarrow\} \cap \phi(x,x)| = 1$.

$H \mapsto K$, otherwise we write $H \not\mapsto K$. For every hereditary property $\mathcal{H}$, we let $\mathcal{K}(\mathcal{H})$ be the set of all dir-types such that none of $\mathcal{F}(\mathcal{H})$ embeds in that type, i.e.,

$$\mathcal{K}(\mathcal{H}) = \{K : K \text{ is a dir-type}, H \not\mapsto K, \forall H \in \mathcal{F}(\mathcal{H})\}\,.$$

*We say that an digraph $G' = (V, c)$ **has type** $K = (U, \phi)$ **if** G' embeds into K with mapping $\gamma : V \to U$ and γ is surjective.*

We have Fact 18, also similar to the r-graph case, which generalizes the ideas underlying the simple editing algorithm in Section 3.3.

Fact 18. *If K is a dir-type, G' is of type K and H does not embed into K, then $H \not\subseteq G'$.*

3.6. **Editing algorithm using types.**

Let $\mathbf{w} = (w_1, \ldots, w_k)$ be a density vector and let $(p_{\bigcirc}, p_{\leftarrow}, p_{\rightarrow}, p_{-})$ be a density vector. This latter vector will represent a vector of densities. The number of ordered pairs (x, y) with color "$-$" will be $p_{-}(n)_2$ and the number of ordered pairs with color "$\bigcirc$" will be $p_{\bigcirc}(n)_2$. The number of ordered pairs with color "$\leftarrow$" is $p_{\leftarrow}(n)_2$ and the number of *ordered* pairs with color "$\rightarrow$" is $p_{\rightarrow}(n)_2$. Consequently, $p_{-} + p_{\bigcirc} + p_{\leftarrow} + p_{\rightarrow} = 1$.

The vector $\mathbf{w}$ will represent a vector of weights, assigned to the vertices of an dir-type with vertices $u_1, \ldots, u_k$, respectively.

Let $\mathcal{P} \subseteq \{\bigcirc, \leftarrow, \rightarrow, -\}$ be a palette, $\mathcal{H}$ be a hereditary property and $G = (V, c)$ be a digraph in $\mathcal{P}$ such that the density vector is $(p_{\bigcirc}, p_{\leftarrow}, p_{\rightarrow}, p_{-})$. In order to find an upper bound on $\operatorname{dist}(G, \mathcal{H})$, it is sufficient to change G to a digraph such that, for all $H \in \mathcal{F}(\mathcal{H})$, H does not embed into the new coloring. In particular, if the resulting coloring has type $K \in \mathcal{K}(\mathcal{H})$, then this coloring is in $\mathcal{H}$.

Algorithm 19. *Fix such a $K = (U, \phi) \in \mathcal{K}(\mathcal{H})$ and try to bring G to a coloring of type K by edge-recoloring. Let $U = \{u_1, \ldots, u_k\}$. Partition the vertices of G randomly into sets $V_1, \ldots, V_k$ such that the probability of a vertex to be in a part V_i is w_i. With an ordering of the vertices of G and vertices $x < y$, consider an edge (x, y) of G, let $x \in V_i$, $y \in V_j$, for $i, j \in \{1, \ldots, k\}$. If $i \neq j$ and $c(x, y) \notin \phi(u_i, u_j)$, recolor (x, y) with a color from $\phi(u_i, u_j)$.*

Next, consider the edges in V_i. If $\phi(u_i, u_i)$ contains exactly one of $\{\leftarrow, \rightarrow\}$, then consider a random order of the vertices of V_i, call it σ. Let $x < y$ and both in V_i. If $c(x, y) = \leftarrow$, then recolor (x, y) if and only if $\sigma(x) < \sigma(y)$. If $c(x, y) = \rightarrow$, then recolor (x, y) if and only if $\sigma(x) > \sigma(y)$. Note that this forces V_i to have no directed cycles. If $\phi(u_i, u_i) \not\ni a$ for some $a \in \{\bigcirc, -\}$, then recolor any edge with color a to a color in $\phi(u_i, u_i)$. This concludes the algorithm.

Algorithm 19 is simply a directed graph version of Algorithm 6. We only needed to address the editing of oriented edges.

3.6.1. *Analysis of the editing algorithm.*

Let us first consider a pair (x, y). If $c(x, y) \in \{\bigcirc, -\}$, then the probability that the color of (x, y) is unchanged is

$$\sum_{1 \leq i, j \leq k} w_i w_j \mathbf{1}_{c(x,y) \in \phi(u_i, u_j)}.$$

If $c(x, y) \in \{\leftarrow, \rightarrow\}$, then the probability that the color of (x, y) is unchanged is

$$\sum_{1 \leq i < j \leq k} w_i w_j \mathbf{1}_{\rightarrow \in \phi(u_i, u_j)} + \sum_{i=1}^{k} w_i^2 \frac{|\{\leftarrow, \rightarrow\} \cap \phi(u_i, u_i)|}{2} = \sum_{1 \leq i, j \leq k} w_i w_j \frac{1}{2} \left|\{\leftarrow, \rightarrow\} \cap \phi(u_i, u_j)\right|.$$

It doesn't matter whether we consider the pair (u_i, u_j) or (u_j, u_i) in the last term because $|\{\leftarrow, \rightarrow\} \cap \phi(u_i, u_j)|$ is invariant whether $i < j$ or $i > j$.

Now, the expected number of changes is

$$\begin{aligned}\mathbf{E}[\#\text{ changes}] &= \binom{n}{2} - \sum_{x,y\in V,\ x<y} \Pr((x,y)\text{ is not changed})\\ &= \binom{n}{2} - \sum_{1\le i,j\le k} w_iw_jp_{\bigcirc}\binom{n}{2}\mathbf{1}_{\bigcirc\in\phi(u_i,u_j)} - \sum_{1\le i,j\le k} w_iw_jp_{-}\binom{n}{2}\mathbf{1}_{-\in\phi(u_i,u_j)}\\ &\quad - \sum_{1\le i,j\le k} w_iw_j\frac{p_\leftarrow+p_\to}{2}\binom{n}{2}\left|\{\leftarrow,\to\}\cap\phi(u_i,u_j)\right|\end{aligned}$$

Let $p=p_-$, $q=\frac{p_\leftarrow+p_\to}{2}$ and so $1-p-2q=p_{\bigcirc}$. For $K=(U,c)$, and $\rho\in\{\bigcirc,-\}$, the matrix $\mathbf{A}_\rho$ is such that the $(i,j)^{\text{th}}$ entry is 1 if $c(u_i,u_j)\ni\rho$ and zero otherwise. The matrix $\mathbf{A}_\to$ is a $\{0,1\}$-matrix with the property that

$$(\mathbf{A}_\to)_{ij} = \left|\{\leftarrow,\to\}\cap c(u_i,u_j)\right|.$$

With $\mathbf{J}$ denoting the $k\times k$ all-ones matrix, then we define

$$\mathbf{M}_K(\mathbf{p}) = \mathbf{J}-(1-p-2q)\mathbf{A}_{\bigcirc}-p\mathbf{A}_- - q\mathbf{A}_\to.$$

Consequently, if $\mathbf{w}=(w_1,\ldots,w_k)$, then $\mathbf{E}[\#\text{ changes}]=\mathbf{w}^T\mathbf{M}_K(\mathbf{p})\mathbf{w}\binom{n}{2}$.

As in the r-graph case, we define two functions in terms of the matrix $\mathbf{M}_K(\mathbf{p})$:

- $f_K(\mathbf{p})=\left(\frac{1}{k}\mathbf{1}\right)^T\mathbf{M}_K(\mathbf{p})\left(\frac{1}{k}\mathbf{1}\right)$ and
- $g_K(\mathbf{p})=\min\left\{\mathbf{w}^T\mathbf{M}_K(\mathbf{p})\mathbf{w} : \mathbf{w}^T\mathbf{1}=1, \mathbf{w}\ge\mathbf{0}\right\}$.

Note 20. *In the directed case, each ordered pair can receive one of 4 directions, but the density vectors only have two entries rather than three. This is because the above computation shows that an upper bound on editing any digraph is determined not by the pair $(p_\leftarrow,p_\to)$ but only by $q=(p_\leftarrow+p_\to)/2$. It is straightforward, by the same arguments as in the proof of Theorem 7, to see that the lower bound for the maximum edit distance is asymptotically achieved by a random graph in which the probability of a forward arc is equal to the probability of a backward arc.*

3.7. Basic results on digraphs.

Theorem 21 is a parallel to Theorem 7 and summarizes some facts about the edit distance function. Recall that, depending on the palette, there may be further restrictions on the density vector other than the necessary $p+2q\le 1$. The dimension, r, of the palette, $\mathcal{P}$, is the number of members of $\{\bigcirc,\to,-\}$ that $\mathcal{P}$ has.

Theorem 21. *Let $\mathcal{H}$ be a hereditary property of digraphs and $\mathcal{P}$ a palette. Fix a density vector with respect to $\mathcal{P}$, $\mathbf{p}=(p,q)$. The limit $\text{dist}(\mathbf{p},\mathcal{H}):=\lim_{n\to\infty}\text{dist}_n(\mathbf{p},\mathcal{H})$ exists. Moreover,*

1. $\text{dist}(\mathbf{p},\mathcal{H})=\inf_{K\in\mathcal{K}(\mathcal{H})}f_K(\mathbf{p})=\inf_{K\in\mathcal{K}(\mathcal{H})}g_K(\mathbf{p})$*;*
2. *Fix $\epsilon>0$, then with probability approaching 1 as $n\to\infty$,*
$$\text{dist}(\mathbf{p},\mathcal{H})-\epsilon\le\text{dist}(G(n,\mathbf{p}),\mathcal{H})\le\text{dist}(\mathbf{p},\mathcal{H});$$
3. $\text{dist}(\mathbf{p},\mathcal{H})=\lim_{n\to\infty}\mathbf{E}[\text{dist}(G(n,\mathbf{p}),\mathcal{H})]$*;*
4. $\text{dist}(\mathbf{p},\mathcal{H})$ *is continuous over the domain of density vectors with respect to $\mathcal{P}$ and is concave down;*
5. $\text{dist}(\mathbf{p},\mathcal{H})$ *achieves its maximum,* $\text{dist}(\mathcal{H})$*, at some density vector $\mathbf{p}^*_{\mathcal{H}}$ (in fact, denote the set of all such vectors $\mathbf{p}^*_{\mathcal{H}}$) and so,*
$$\text{dist}(\mathcal{H})=\lim_{n\to\infty}\mathbf{E}[\text{dist}(G(n,\mathbf{p}^*_{\mathcal{H}}),\mathcal{H})]; \textit{ and}$$
6. *Both $\mathbf{p}^*_{\mathcal{H}}$ and* $\text{dist}(\mathcal{H})$ *exist and $\mathbf{p}^*_{\mathcal{H}}$ is a convex and closed set in $[0,1]^{r-1}$.*

Note 22. *Again, we abuse notation so that $\mathbf{p}_{\mathcal{H}}^*$ can be a single vector or a set.*

3.8. **Example: tournaments.**

The case of tournaments is relatively straightforward. Because in tournaments, there are no edges labeled ◯ or $-$, there is only one density vector, $\mathbf{p} = (0, 1/2)$. This means that we only need to consider tournaments that are random, that each arc is forward independently with probability 1/2. This leads to a rather simple expression for the edit distance:

Theorem 23. *Let $\mathcal{H}$ be a nontrivial hereditary property of tournaments and let $\mathcal{P} = \mathcal{P}_{\text{tourn}} = \{\leftarrow, \rightarrow\}$. Then,*

$$\text{dist}(\mathcal{H}) = \frac{1}{2(\chi_{\text{dir}}^{\mathcal{P}}(\mathcal{H}) - 1)}.$$

Note that in the case of tournaments, the directed chromatic number of tournament H, $\chi_{\text{dir}}^{\mathcal{P}_{\text{tourn}}}(H)$ is the smallest number of transitive subtournaments into which H can be partitioned. We prove Theorem 23 in Section 3.10.3.

3.9. **Example: triangles.**

Theorem 24 gives some basic results on examples of hereditary properties of digraphs defined by triangles. The proof is in Section 3.10.4.

Theorem 24. *Consider hereditary properties of digraphs.*

(1) *If $\mathcal{F}$ is a family that consists of a single directed triangle, then, regardless of the palette,* $\text{dist}(\text{Forb}(\mathcal{F})) = 1/2$.

(2) *If $\mathcal{F}$ is a family that consists of a single transitive triangle and $\mathcal{P} = \mathcal{P}_{\text{tourn}}$, the palette of tournaments, then* $\text{Forb}(\mathcal{F})$ *is a trivial hereditary property.*

(3) *If $\mathcal{F}$ is a family of that consists of a single transitive triangle, then, if $\mathcal{P}$ is any palette other than $\mathcal{P}_{\text{tourn}}$, then* $\text{dist}(\text{Forb}(\mathcal{F})) = 1/2$.

(4) *If $\mathcal{F}$ is a family that consists of both a transitive and a directed triangle, and $\mathcal{P}$ is any palette other than $\mathcal{P}_{\text{tourn}}$, then* $\text{dist}(\text{Forb}(\mathcal{F})) = 1/2$.

3.10. **Proofs.**

3.10.1. *Proof of Theorem 16.* The upper bound for this theorem is proven by the simple editing algorithm from Section 3.3.

Let $r = |\mathcal{P}|$. For the lower bound, we apply part 1 of Theorem 21, which states that $\text{dist}(\mathbf{p}, \mathcal{H}) = \inf_{K \in \mathcal{K}(\mathcal{H})} f_K(\mathbf{p})$. Consider an arbitrary $K = (V, \phi) \in \mathcal{K}(\mathcal{H})$, a $\mathcal{P}$-dir-type on k vertices. Let $\tilde{K}$ be a graph with vertex set V such that v and v' are adjacent in $\tilde{K}$ if and only if $\phi(v, v') = \mathcal{P}$. We observe that $\tilde{K}$ has no clique on $\chi_{\text{dir}}^{\mathcal{P}}$ vertices, otherwise for some $H \in \mathcal{F}(\mathcal{H})$, $H \mapsto K$. Using Turán's theorem, the number of edges of $\tilde{K}$ is at most $\frac{\chi_{\text{dir}}^{\mathcal{P}} - 2}{\chi_{\text{dir}}^{\mathcal{P}} - 1} \cdot \frac{k^2}{2}$. Let $\mathbf{p} = \frac{1}{r}\mathbf{1}$. Consider the matrix $\mathbf{M}_K(\mathbf{p})$ and observe that every entry is either zero or is a positive integer multiple of $1/r$. The zero entries correspond exactly to pairs with ϕ value equal to $\mathcal{P}$. Thus, this matrix $\mathbf{M}_K(\mathbf{p})$ has at least $k^2 - 2\left(\frac{\chi_{\text{dir}}^{\mathcal{P}} - 2}{\chi_{\text{dir}}^{\mathcal{P}} - 1} \cdot \frac{k^2}{2}\right) \geq \frac{k^2}{\chi_{\text{dir}}^{\mathcal{P}} - 1}$ entries with value at least $1/r$. Therefore, $f_K(\mathbf{p}) = \frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(\mathbf{p})\mathbf{1}$ is at least $1/r(\chi_{\text{dir}}^{\mathcal{P}} - 1)$. Since K was arbitrary, this gives a lower bound for $\text{dist}(\mathbf{p}, \mathcal{H})$.

3.10.2. *Proof of Theorem 21.* The proof of most of this theorem is identical to that of Theorem 7, which is found in Section 2.9.2. The only significant wrinkle is the upper bound. That is, if G is a digraph with $p\binom{n}{2}$ edges with color $-$ and $(1-p-2q)\binom{n}{2}$ edges with color $\bigcirc$, then, with $\mathbf{p}=(p,q)$,

$$\mathrm{dist}(G,\mathcal{H})/\binom{n}{2} \leq \inf_{K\in\mathcal{K}(\mathcal{H})} f_K(\mathbf{p}).$$

This follows directly from the analysis of the editing algorithm using types from Section 3.6.

3.10.3. *Proof of Theorem 23.*
In this case, $\mathbf{p}=(0,1/2)$. Let $\mathcal{H}$ be a hereditary property of tournaments and $\chi_{\rm dir}=\chi_{\rm dir}^{\mathcal{P}_{\rm tourn}}(\mathcal{H})$. In any type K on k vertices, the vertices have color "$\rightarrow$" and the edges either have one direction or both. By the definition of the directed chromatic number, $H\mapsto K$ if K has a clique of order $\chi_{\rm dir}$ such that every edge of K has color set $\{\leftarrow,\rightarrow\}$.

Similar to the argument in Section 3.10.1, we can use Turán's theorem to find a lower bound for $f_K(\mathbf{p})$. The bilinear form $\mathbf{1}^T\mathbf{M}_K(\mathbf{p})\mathbf{1}$ counts $\frac{1}{2}|V(K)|+\frac{1}{2}|E_{\leftarrow}(K)|+\frac{1}{2}|E_{\rightarrow}(K)|$, where $E_\rho(K)$ is the set of *ordered* pairs with color ρ. Since $|E_{\{\leftarrow,\rightarrow\}}(K)|+|E_{\leftarrow}(K)|+|E_{\rightarrow}(K)|=k(k-1)$, Turán's theorem gives that $|E_{\{\leftarrow,\rightarrow\}}(K)|\leq\frac{\chi_{\rm dir}-2}{\chi_{\rm dir}-1}k^2$. Consequently,

$$f_K(\mathbf{p})=\frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(\mathbf{p})\mathbf{1}=\frac{1}{k^2}\left[\frac{1}{2}k+\frac{1}{2}k(k-1)-\frac{\chi_{\rm dir}-2}{\chi_{\rm dir}-1}k^2\right]=\frac{1}{2(\chi_{\rm dir}-1)}.$$

This concludes the proof of Theorem 23.

3.10.4. *Proof of Theorem 24.*

(1) As to the upper bound, linearly order the vertices so that the number of backward edges (i.e., pairs $\{v_i,v_j\}$ such that $i<j$ and $c(v_i,v_j)=\leftarrow$) is minimized. A greedy ordering results in at most half of such edges being present. Reorient such edges so that they become forward edges, hence $\mathrm{dist}(\mathrm{Forb}(\mathcal{F}))\leq 1/2$. Note that this corresponds to a K that consists of a single vertex which has color $\rightarrow$.

For the lower bound, consider an arbitrary $K\in\mathcal{K}(\mathcal{H})$ with vertex set $\{u_1,\ldots,u_k\}$ and $\mathbf{p}=(0,1/2)$. This means that $\mathbf{M}_K(\mathbf{p})=\mathbf{J}-\frac{1}{2}\mathbf{A}_{\rightarrow}$. I.e., $(\mathbf{M}_K(\mathbf{p}))_{i,j}=1-\frac{1}{2}|c(u_i,u_j)\cap\{\leftarrow,\rightarrow\}|$.

Here we use an approach due to Sidorenko [18]. See also [12, 19, 20, 21]. For the optimal solution, $\mathbf{w}^*$ to the quadratic program $g_K(\mathbf{p})=\min\left\{\mathbf{w}^T\mathbf{M}_K(\mathbf{p})\mathbf{w}:\mathbf{w}^T\mathbf{1}=1,\mathbf{w}\geq\mathbf{0}\right\}$, the vector $\mathbf{M}_K(\mathbf{p})\mathbf{w}^*$ is a constant vector, equal to $g_K(\mathbf{p})\mathbf{1}$.

Observe that there can be no entry $(\mathbf{M}_K(\mathbf{p}))_{ii}=0$ because that means, for the corresponding vertex u_i, $c(u_i,u_i)\supseteq\{\leftarrow,\rightarrow\}$ and a directed triangle maps to such a vertex. Suppose there is some entry $(\mathbf{M}_K(\mathbf{p}))_{ij}=0$. This implies that there are a pair of vertices, u_i and u_j such that $c(u_i,u_j)\supseteq\{\leftarrow,\rightarrow\}$. We observe that $|c(u_i,u_i)\cap\{\leftarrow,\rightarrow\}|=|c(u_i,u_i)\cap\{\leftarrow,\rightarrow\}|=0$, otherwise the directed triangle would map to these two vertices of K. Moreover, for every $\ell\in\{1,\ldots,k\}-\{i,j\}$, we have $(\mathbf{M}_K(\mathbf{p}))_{i\ell}+(\mathbf{M}_K(\mathbf{p}))_{j\ell}\geq 1$. If not, then without loss of generality, we have a triangle $\{u_i,u_j,u_\ell\}$ in K such that two edges contain $\{\leftarrow,\rightarrow\}$ and the third contains one of $\{\leftarrow,\rightarrow\}$. It is easy to see that a directed triangle maps to three such vertices. But then,

$$\sum_\ell(\mathbf{M}_K(\mathbf{p}))_{i\ell}w_\ell+\sum_\ell(\mathbf{M}_K(\mathbf{p}))_{j\ell}w_\ell\geq(1-w_i-w_j)+(\mathbf{M}_K(\mathbf{p}))_{i\ell}(w_i+w_j)+(\mathbf{M}_K(\mathbf{p}))_{ii}w_i+(\mathbf{M}_K(\mathbf{p}))_{jj}w_j=1.$$

Since both sums on the left hand side must be equal to $g_K(\mathbf{p})$, it must be that $g_K(\mathbf{p})\geq 1/2$.

Finally, if there is no zero entry in the i-th row of $\mathbf{M}_K(\mathbf{p})$, then $\sum_\ell(\mathbf{M}_K(\mathbf{p}))_{i\ell}w_\ell\geq 1/2$. Thus, in all cases, $g_K(\mathbf{p})\geq 1/2$.

(2) Here we make the easily verified observation that any tournament with at least 4 vertices has a transitive subtournament of size 3. So, the hereditary property consists of no tournaments of size 4 or more.

(3) As to the upper bound, equipartition the vertex set arbitrarily and recolor an edge inside each part to have a color other than one in $\{\leftarrow, \rightarrow\}$. Hence, $\mathrm{dist}(\mathrm{Forb}(\mathcal{F})) \leq 1/2$. Note that this corresponds to a K that consists of two vertices colored with some nonempty subset of $\{\bigcirc, -\}$ and an edge colored $\{\leftarrow, \rightarrow\}$.

For the lower bound, simply let G be a transitive tournament. After editing G to make G', there can be no triangles from G that remain and so Mantel's theorem gives that

$$\mathrm{dist}(G, G') \geq \frac{1}{\binom{n}{2}} \left(\binom{n}{2} - \left\lfloor \frac{n^2}{4} \right\rfloor \right) = \frac{1}{2} - O\left(\frac{1}{n}\right).$$

(4) Here we can use the trivial fact that if $\mathcal{H}$ is the hereditary property that forbids both directed and transitive triangles and $\mathcal{H}'$ is the larger hereditary property in (3) which forbids only the transitive triangle, then $\mathrm{dist}(\mathcal{H}) \geq \mathrm{dist}(\mathcal{H}') = 1/2$. But the example above of a type K that consists of two vertices with none of $\{\leftarrow, \rightarrow\}$ in its color set is in $\mathcal{K}(\mathcal{H})$ in this case as well. Hence, $f_K(\mathbf{p}) = 1/2$ and so $\mathrm{dist}(\mathcal{H}) = 1/2$.

References

[1] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, *Combinatorica* **20**(4) (2000), no. 4, 451–476.

[2] N. Alon and J. Spencer, *The probabilistic method.* Third edition. With an appendix on the life and work of Paul Erdős. Wiley-Interscience Series in Discrete Mathematics and Optimization. *John Wiley & Sons, Inc., Hoboken, NJ*, 2008.

[3] N. Alon and A. Stav, What is the furthest graph from a hereditary property? *Random Structures Algorithms* **33** (2008), no. 1, pp. 87–104.

[4] M. Axenovich, A. Kézdy and R. Martin, On the editing distance of graphs, *J. Graph Theory* **58** (2008), no. 2, 123–138.

[5] M. Axenovich and R. Martin, A version of Szemerédi's regularity lemma for multicolored graphs and directed graphs that is suitable for induced graphs, manuscript.

[6] J. Balogh and R. Martin, Edit distance and its computation, *Electron. J. Combin.* **16**(1) (2009), Research Paper 109, 16pp. (electronic).

[7] W.G. Brown, On graphs that do not contain a Thomsen graph, *Canad. Math. Bull.*, 1966, **9**, 281-285.

[8] C.M. Fortuin, P.W. Kasteleyn and J. Ginibre, "Correlation inequalities on some partially ordered sets," *Communications in Mathematical Physics*, 1971, **22**, 89–103.

[9] D. E. Knuth, "Permutations, matrices, and generalized Young tableaux," *Pacific Journal of Mathematics*, **34**(3), (1970), 709–727.

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

[11] E. Marchant, Ph.D. thesis, University of Cambridge, (2010).

[12] E. Marchant and A. Thomason, Extremal graphs and multigraphs with two weighted colours, in *Fete oif Combinatorics and Computer Science* (eds. G.O.H. Katona, A. Schrijver and T. Szönyi), *Bolyai Soc. Math. Stud.* **20**, Springer, Berlin, in press.

[13] R. Martin, Edit distance and localization, preprint.

[14] R. Martin and T. McKay, On the edit distance from $K_{2,t}$-free graphs I: The case of $t = 3, 4$, preprint.

[15] W. Rudin, *Principles of Mathematical Analysis.* Third edition. International Series in Pure and Applied Mathematics. *McGraw-Hill Book Co., New York-Auckland-Düsseldorf*, 1976.

[16] B. Bollobás and A. Thomason, Hereditary and monotone properties of graphs, in *The Mathematics of Paul Erdős II* (R.L. Graham and J. Nešetřil, eds.) *Algorithms and Combinatorics* **14**, Springer-Verlag (1997), 70–78.

[17] B. Bollobás and A. Thomason, The structure of hereditary properties and colourings of random graphs, *Combinatorica* **20** (2000), 173–202.

[18] A.F. Sidorenko, Boundedness of optimal matrices in extremal multigraph and digraph problems, *Combinatorica* **13** (1993), 109–120.

[19] R. Martin, Edit distance and localization, submitted, `arXiv:1007.1897v3`.

[20] R. Martin and T. McKay, On the edit distance from $K_{2,t}$-free graphs, submitted, `arXiv:1012.0800`.

[21] R. Martin, On the computation of edit distance functions, submitted, `arXiv:1012.1237`.

Department of Mathematics, Iowa State University, Ames, Iowa 50011
E-mail address: `axenovic@iastate.edu`

Department of Mathematics, Iowa State University, Ames, Iowa 50011
E-mail address: `rymartin@iastate.edu`