The edit distance in graphs: methods, results and generalizations

Ryan R. Martin

Abstract The edit distance is a very simple and natural metric on the space of graphs. In the edit distance problem, we fix a hereditary property of graphs and compute the asymptotically largest edit distance of a graph from the property. This quantity is very difficult to compute directly but in many cases, it can be derived as the maximum of the edit distance function. Szemerédi's regularity lemma, strongly-regular graphs, constructions related to the Zarankiewicz problem – all these play a role in the computing of edit distance functions. The most powerful tool is derived from symmetrization, which we use to optimize quadratic programs that define the edit distance function. In this paper, we describe some of the most common tools used for computing the edit distance function, summarize the major current results, outline generalizations to other combinatorial structures, and pose some open problems.

1 Introduction

The edit distance in graphs was originally studied to answer two different and independent problems: one to answer questions on property-testing [7] and the other, to answer a question regarding consensus trees from evolutionary biology [8]. In metabolic networks, the presence or absence of edges in a certain graph correspond to pairs of genes which activate or deactivate one another. In evolutionary theory, avoiding forbidden induced subgraphs [23] is studied, which is equivalent to a similar edit problem of bipartite graphs or matrices. Edit distance problems with respect to more general classes of graphs are important in the algorithmic aspects of property testing [4, 5, 3, 7] and in the techniques involved in computing the speed of dense graph properties [46, 19].

Ryan R. Martin
Iowa State University, Ames, Iowa 50011-2064, e-mail: rymartin@iastate.edu

The (normalized) edit metric is a metric on the set of simple, labeled n-vertex graphs. The distance between two graphs is the symmetric difference of the edge sets divided by the total number of possible edges. If $\mathrm{dist}(G,G')$ denotes the edit distance between G and G' on the same labeled vertex set, then

$$\mathrm{dist}(G,G') = |E(G) \triangle E(G')|/\binom{n}{2}.$$

As with any metric, we may take a *property* of graphs $\mathscr{H}$ (that is, a set of graphs), and compute the distance of a graph from that property:

$$\mathrm{dist}(G,\mathscr{H}) = \min\left\{\mathrm{dist}(G,G') : V(G') = V(G)\right\}. \tag{1}$$

The properties that we study in this paper are *hereditary properties*. A property of graphs is hereditary if is closed under isomorphism and deletion of vertices. Alon and Stav [7] suggest that "In fact, almost all interesting graph properties are hereditary." Planarity, having chromatic number at most k or not having a given H as an induced subgraph all are commonly-studied hereditary properties. The property of having no graph H as an induced subgraph is called a *principal hereditary property* and we denote it by $\mathrm{Forb}(H)$. For every hereditary property $\mathscr{H}$ there exists a family of graphs $\mathscr{F}(\mathscr{H})$ ("forbidden graphs") such that $\mathscr{H} = \bigcap_{H\in\mathscr{F}(H)} \mathrm{Forb}(H)$. A hereditary property is said to be *nontrivial* if there is an infinite sequence of graphs that are in the property.

In the seminal papers by Alon and Stav [6, 7] and by Axenovich, Kézdy and Martin [8], the fundamental question was the maximum distance of a graph G on n vertices from hereditary property $\mathscr{H}$. In fact, the maximum distance is asymptotically the same as that of the Erdős-Rényi random graph $G(n,p)$, for some value of p.

Theorem 1 (Alon-Stav [7]) *Let $\mathscr{H}$ be an arbitrary graph property. There exists $p^* = p^*_{\mathscr{H}} \in [0,1]$ such that*

$$\max\{\mathrm{dist}(G,\mathscr{H}) : |V(G)| = n\} = \mathbb{E}[\mathrm{dist}(G(n,p^*),\mathscr{H})] + o(1). \tag{2}$$

We denote the limit of the quantity in (2) by $d^*_{\mathscr{H}}$. This is, asymptotically, the maximum distance of a graph from $\mathscr{H}$. Although $d^*_{\mathscr{H}}$ is the quantity in which we are most interested, determining its value is most often done by generalizing the result in Theorem 1. We do so by instead finding the maximum edit distance of a density-p graph from $\mathscr{H}$, for **all** values of p.

Balogh and Martin [16] introduced the *edit distance function* of a hereditary property.

Definition 2 *Let $\mathscr{H}$ be a nontrivial hereditary property of graphs. The* edit distance function *of $\mathscr{H}$ is*

$$\mathrm{ed}_{\mathscr{H}}(p) := \lim_{n\to\infty} \max\left\{\mathrm{dist}(G,\mathscr{H}) : |V(G)| = n, |E(G)| = \left\lfloor p\binom{n}{2}\right\rfloor\right\}. \tag{3}$$

The existence of the limit in (3) was proven in [16].[1]

Theorem 3 (Balogh-Martin [16]) *Let $\mathscr{H}$ be an arbitrary nontrivial graph property. Then*

$$\mathrm{ed}_{\mathscr{H}}(p) = \lim_{n\to\infty} \mathbb{E}[\mathrm{dist}(G(n,p),\mathscr{H})].$$

Theorems 1 and 3 make use of Szemerédi's regularity lemma [48] but in a way that detects induced subgraphs. The idea of applying Szemerédi's regularity lemma to hereditary properties has been studied in a number of contexts, including pioneering work by Prömel and Steger [40, 41, 42], Scheinerman and Zito [46], and Bollobás and Thomason [18, 19, 20]. The essential technique is to apply the regularity lemma twice – once to the graph itself and a second time to each of the graphs induced by the non-exceptional clusters. More directly, one can use a variant of Szemerédi's regularity lemma due to Alon, et al. [5] that has been used in a number of papers, including the edit distance papers [7, 16].

The edit distance function is symmetric with respect to complementation. It is easy to see that $\mathrm{ed}_{\mathrm{Forb}(H)}(p) = \mathrm{ed}_{\mathrm{Forb}(\overline{H})}(1-p)$ and, in fact, a general case is true.

Proposition 4 *Let $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$ be a nontrivial hereditary property and let $\mathscr{H}^* = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(\overline{H})$. Then $\mathrm{ed}_{\mathscr{H}}(p) = \mathrm{ed}_{\mathscr{H}^*}(1-p)$.*

A very similar setting to the edit distance problem was studied by Richer [43] and as further investigated byMarchant and Thomason [31, 32] regarding the two-coloring of the edges of the complete graph. Many of the most vital results for solving the edit distance problem come from this setting. In solving the problems they pose on a hereditary property $\mathscr{H}$, they obtain the function $1-\mathrm{ed}_{\mathscr{H}}(p)$. The connection between the two settings is addressed in [31] as well as by Thomason [49] in a survey.

2 Colored regularity graphs

The key observation in computing the edit distance is that a graph can be approximated by a graph-like structure in which the clusters either behave like cliques or independent sets and the ε-regular pairs either behave like complete bipartite graphs, empty bipartite graphs or random graphs with density bounded away from both 0 and 1.

Alon and Stav [7] defined a *colored regularity graph (CRG) K* to be a simple complete graph, together with a partition of the vertices into white and black $V(K) =$

[1] It should be noted that early papers on the edit distance do not normalize the distance. That is, the distance is merely $|E(G)\triangle E(G')|$. Normalization, however, is required in order to define the edit distance function and it seems most natural to put the normalization in the metric itself, rather than doing so in order to define $\mathrm{ed}_{\mathscr{H}}$.

$\mathrm{VW}(K) \cup \mathrm{VB}(K)$ and a partition of the edges into white, gray and black, $E(K) = \mathrm{EW}(K) \cup \mathrm{EG}(K) \cup \mathrm{EB}(K)$.[2]

We say that a graph H embeds in K, writing $H \mapsto K$, if there is a function $\varphi : V(H) \to V(K)$ so that the following occurs:

- If $h_1h_2 \in E(H)$, then either $\varphi(h_1) = \varphi(h_2) \in \mathrm{VB}(K)$ or $\varphi(h_1)\varphi(h_2) \in \mathrm{EB}(K) \cup \mathrm{EG}(K)$.
- If $h_1h_2 \notin E(H)$, then either $\varphi(h_1) = \varphi(h_2) \in \mathrm{VW}(K)$ or $\varphi(h_1)\varphi(h_2) \in \mathrm{EW}(K) \cup \mathrm{EG}(K)$.

A CRG K' is said to be a *sub-CRG* of K if K' can be obtained by deleting vertices of K and is a *proper sub-CRG* if $K' \neq K$.

If a graph H embeds in CRG K then a large enough graph that is approximated by K will have an induced copy of H. This is stated and proven more formally in Section 4 of [16]. However, the main idea is that for any large graph in a hereditary property $\mathscr{H} = \bigcap_{H \in \mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$, the CRG K that approximates the graph satisfies the property that $H \not\mapsto K$ for all $H \in \mathscr{F}(\mathscr{H})$. We denote $\mathscr{K}(\mathscr{H})$ to be the subset of CRGs K such that no forbidden graph maps into K. Formally, $\mathscr{K}(\mathscr{H}) = \{K : H \not\mapsto K, \forall H \in \mathscr{F}(\mathscr{H})\}$.

2.1 The f and g functions

There is a matrix associated with a CRG called $\mathbf{M}_K(p)$ that plays a role similar to the role the adjacency matrix does for graphs. We can use this matrix to help define the functions f_K and g_K, that are essential for understanding edit distance.

Definition 5 *Let K be a CRG on vertex set $\{v_1, \ldots, v_k\}$ with* VW *and* VB *denoting the white and black vertices, respectively, and* EW, EG *and* EB *denoting the white, gray and black edges, respectively. Let $\mathbf{M}_K(p)$ denote the matrix with entries defined as follows:*

$$m_K(p)_{ij} = \begin{cases} p, & \text{if } i \neq j \text{ and } v_iv_j \in \mathrm{EW} \text{ or } i = j \text{ and } v_i \in \mathrm{VW}; \\ 0, & \text{if } i \neq j \text{ and } v_iv_j \in \mathrm{EG}; \\ 1-p, & \text{if } i \neq j \text{ and } v_iv_j \in \mathrm{EB} \text{ or } i = j \text{ and } v_i \in \mathrm{VB}. \end{cases} \tag{4}$$

The functions f_K and g_K are defined as follows:

$$f_K(p) = \frac{1}{k^2}\left[p(|\mathrm{VW}| + 2|\mathrm{EW}|) + (1-p)(|\mathrm{VB}| + 2|\mathrm{EB}|)\right] = \frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(p)\mathbf{1} \tag{5}$$

$$g_K(p) = \min\left\{\mathbf{x}^T\mathbf{M}_K(p)\mathbf{x} : \mathbf{x}^T\mathbf{1} = 1, \mathbf{x} \geq \mathbf{0}\right\}. \tag{6}$$

The vector $\mathbf{0}$ is the all-zeroes vector, $\mathbf{1}$ is the all-ones vector, and vector inequalities are coordinatewise.

[2] Papers by Bollobás and Thomason [18, 19, 20] and others such as [49] use the term “type” rather than CRG.

Clearly, for any CRG K and any $p \in [0,1]$, we have $g_K(p) \leq f_K(p)$. Although the linearity of f_K makes proving general results about $\mathrm{ed}_{\mathscr{H}}$ possible, the g function is more useful in computing the edit distance function. In fact, if an optimal vector of (6) has a zero entry, we may obtain K' by deleting the corresponding entry and achieve $g_{K'}(p) = g_K(p)$. We say that a CRG, K is *p-core* if, for any proper sub-CRG K' of K, we have $g_{K'}(p) > g_K(p)$.

The edit distance function can be defined in terms of the f and g functions:

Theorem 6 *Let $\mathscr{H}$ be a nontrivial hereditary property. For any $p \in [0,1]$,*

$$\mathrm{ed}_{\mathscr{H}}(p) = \inf\{f_K(p) : K \in \mathscr{K}(\mathscr{H})\} = \inf\{g_K(p) : K \in \mathscr{K}(\mathscr{H})\} \qquad (7)$$

$$= \min\{g_K(p) : K \in \mathscr{K}(\mathscr{H})\}. \qquad (8)$$

Equation (7) is due to Balogh and Martin [16]. Equation (8) is from the results of Marchant and Thomason [31] and gives rise to the question as to whether only a finite set of p-core CRGs is sufficient to define the edit distance function for any nontrivial hereditary property and all $p \in [0,1]$.

There is some evidence (see Theorem 31(d) and Theorem 30(b) below) that for some hereditary properties, determining the edit distance function requires knowledge of an infinite sequence of CRGs. Nonetheless, we believe that the bulk of the edit distance function can be determined from a finite number of CRGs. That is, for any $\varepsilon > 0$, we believe a finite set of CRGs can simultaneously define $\mathrm{ed}_{\mathscr{H}}$ for all $p \in [\varepsilon, 1-\varepsilon]$. This is Conjecture 1 in Section 8.2.

2.2 Clique spectrum

Certain colored regularity graphs play a key role in the computation of the edit distance. A *gray-edge CRG* is the CRG, K with all $\binom{|V(K)|}{2}$ edges gray. The gray-edge CRG with r white vertices and s black vertices is denoted $K(r,s)$. The *clique spectrum* of $\mathscr{H}$ is the set

$$\Gamma(\mathscr{H}) \stackrel{\mathrm{def}}{=} \{(r,s) : H \not\mapsto K(r,s), \forall H \in \mathscr{F}(\mathscr{H})\}.$$

For example, if $\mathscr{H} = \mathrm{Forb}(H)$ is a hereditary property, then pairs (r,s) are in the clique spectrum of $\mathrm{Forb}(H)$ if and only if H **cannot** be partitioned into r independent sets and s cliques.

The clique spectrum has a number of useful properties. For example, it is monotone in the sense that if $(r,s) \in \Gamma(\mathscr{H})$ and $0 \leq r' \leq r$ and $0 \leq s' \leq s$, then $(r',s') \in \Gamma(\mathscr{H})$. As a result, the clique spectrum of a hereditary property can be expressed as a Ferrers diagram. An *extreme point* of the clique spectrum Γ is a pair $(r,s) \in \Gamma$ for which both $(r+1,s) \notin \Gamma$ and $(r,s+1) \notin \Gamma$. Figure 1 shows the graph H_9, and Figure 2 shows its clique spectrum, expressed as a Ferrers diagram.

Since the matrix $\mathbf{M}_{K(r,s)}(p)$ is a diagonal matrix with r entries of value p and s entries with value $1-p$, it is easy to compute that for all $p \in (0,1)$

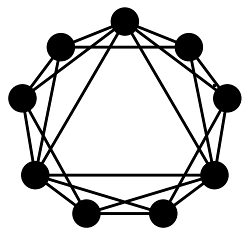

Fig. 1 A graph H_9 on 9 vertices.

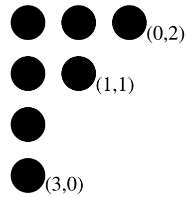

Fig. 2 The Ferrers diagram of the clique spectrum of H_9 with the extreme points labeled.

$$g_{K(r,s)}(p) = \left(\frac{r}{p} + \frac{s}{1-p}\right)^{-1} = \frac{p(1-p)}{r(1-p)+sp}. \tag{9}$$

We have the natural convention that if $r = 0$ then $g_{K(r,s)}(0) = 1$ and if $s = 0$ then $g_{K(r,s)}(1) = 1$.

In fact, we have a more general way of computing the edit distance function if the matrix $\mathbf{M}_K(p)$ is block diagonal matrix, where the blocks correspond to a CRG notion of components.

Definition 7 *A sub-CRG, K', of a CRG K is a* component *if it is maximal with respect to the property that, for all $v, w \in V(K')$, there exists a path consisting of white and black edges entirely within K'.*

More simply, components of K are the components of the graph G with vertex set $V(K)$ and the nonedges of G the gray edges of K. This leads to the generalization of (9).

Proposition 8 (Martin [33]) *Let K be a CRG with components $K_{(1)}, \ldots, K_{(\ell)}$. Then*

$$(g_K(p))^{-1} = \sum_{i=1}^{\ell} (g_{K_{(i)}}(p))^{-1}$$

2.3 Characterization of p-core CRGs

Marchant and Thomason [31] gave a characterization of all p-core CRGs.

Theorem 9 (Marchant-Thomason [31]) *Let K be a p-core CRG.*

(a) If $p \leq 1/2$ then there are no black edges and the white edges are only incident to black vertices.
(b) If $p \geq 1/2$ then there are no white edges and the black edges are only incident to white vertices.

Consequently, if $p = 1/2$ then all edges are gray.

Theorem 9 is an essential tool and is used in most results on the edit distance function as we shall see below.

3 Estimating the edit distance function

Although it is difficult to compute the edit distance function for general hereditary properties, we can estimate the function through a variety of techniques. First, we can use the clique spectrum and (9) to construct an upper bound.

3.1 Upper bound via the clique spectrum

We begin with a trio of results with elementary proofs, followed by a result concerning the nature of the edit distance function.

Theorem 10 *Let $\mathscr{H}$ be a nontrivial hereditary property and let $\Gamma(\mathscr{H})$ denote the clique spectrum of $\mathscr{H}$. If we define*

$$\gamma_{\mathscr{H}}(p) := \min_{(r,s)\in\Gamma(\mathscr{H})} g_{K(r,s)}(p) = \min_{(r,s)\in\Gamma(\mathscr{H})} \frac{p(1-p)}{r(1-p)+sp}, \tag{10}$$

Then $\mathrm{ed}_{\mathscr{H}}(p) \leq \gamma_{\mathscr{H}}(p)$.

There are three (not necessarily distinct) extreme points of a clique spectrum that are of particular interest. First, if $(r,0)\in\Gamma(\mathscr{H})$ but $(r+1,0)\notin\Gamma(\mathscr{H})$, then $r+1$ is the *chromatic number* of $\mathscr{H}$, denoted $\chi(\mathscr{H})$ or just χ, when the hereditary property is understood. Second, if $(0,s)\in\Gamma(\mathscr{H})$ but $(0,s+1)\notin\Gamma(\mathscr{H})$, then $s+1$ is the *complementary chromatic number* of $\mathscr{H}$, denoted $\overline{\chi}(\mathscr{H})$ or just $\overline{\chi}$. Note that if $\mathscr{H}=\mathrm{Forb}(H)$ for some graph H then $\chi(\mathscr{H})=\chi(H)$ and $\overline{\chi}(\mathscr{H})=\chi(\overline{H})$.

We observe that if $\chi(\mathscr{H})\geq 2$ then $(\chi-1,0)\in\Gamma(\mathscr{H})$ and if $\overline{\chi}(\mathscr{H})\geq 2$ then $(0,\overline{\chi}-1)\in\Gamma(\mathscr{H})$. Therefore, we have the following corollary.

Corollary 11 *Let $\mathscr{H}$ be a nontrivial hereditary property with chromatic number χ and complementary chromatic number $\overline{\chi}$.*

(a) If $\chi\geq 2$, then $\mathrm{ed}_{\mathscr{H}}(p)\leq p/(\chi-1)$.
(b) If $\overline{\chi}\geq 2$, then $\mathrm{ed}_{\mathscr{H}}(p)\leq (1-p)/(\overline{\chi}-1)$.

The chromatic and complementary chromatic numbers of a hereditary property $\mathscr{H}$ can be defined in terms of $\mathscr{F}(\mathscr{H})$ as in Proposition 12.

Proposition 12 *Let $\mathscr{H}=\bigcap_{H\in\mathscr{F}}\mathrm{Forb}(H)$ be a nontrivial hereditary property. Then,*

(a) $\chi(\mathscr{H})=\min\{\chi(H):H\in\mathscr{F}\}$ *and*
(b) $\overline{\chi}(\mathscr{H})=\min\{\chi(\overline{H}):H\in\mathscr{F}\}$.

The third extreme point we address is evaluated as follows: the largest value of $r+s+1$ such that $(r,s)\in\Gamma(\mathscr{H})$ is called the *binary chromatic number* of $\mathscr{H}$ and is denoted $\chi_B(\mathscr{H})$ or just χ_B. This quantity has appeared in the literature previously. Prömel and Steger [40, 41, 42] called χ_B-1 simply τ. Bollobás and Thomason [18, 20] called χ_B-1 the *colouring number*.

Since Theorem 9 establishes that every $1/2$-core CRG is a gray edge CRG (that is, of the form $K(r,s)$) we can compute $\mathrm{ed}_{\mathscr{H}}(1/2)$ in terms of $\chi_B(\mathscr{H})$. Combining this with other basic facts, we obtain Theorem 13.

Theorem 13 *Let $\mathscr{H}$ be a nontrivial hereditary property.*

(a) $\mathrm{ed}_{\mathscr{H}}(p)$ *is continuous.*
(b) $\mathrm{ed}_{\mathscr{H}}(p)$ *is concave down.*
(c) $\mathrm{ed}_{\mathscr{H}}(1/2) = \frac{1}{2(\chi_B(\mathscr{H})-1)}$.

Theorem 13(a) was established by Marchant and Thomason [31]. We note that a different, analysis-based proof of this is in [16]. Theorem 13(b) was proven in [16]. Theorem 13(c) was proven in [8] in the case where $\mathscr{H}$ is a principal hereditary property. More sophisticated knowledge of the edit distance function has made it a simple corollary.

Using only Corollary 11 and Theorem 13 we can already find edit distance functions for some important hereditary properties. If P_4 denotes the path on 4 vertices, then $\mathrm{ed}_{\mathrm{Forb}(P_4)}(p) = \min\{p, 1-p\}$. If C_5 denotes the cycle on 5 vertices, then $\mathrm{ed}_{\mathrm{Forb}(C_5)}(p) = \frac{1}{2}\min\{p, 1-p\}$. More about hereditary properties forbidding self-complementary graphs is below in Corollary 15.

Because $\mathrm{ed}_{\mathscr{H}}(p)$ is continuous and concave down, the function achieves its maximum on the interval $[0,1]$. Thus, both $d^*_{\mathscr{H}}$ and $p^*_{\mathscr{H}}$ are well-defined and the coordinate $(p^*_{\mathscr{H}}, d^*_{\mathscr{H}})$ is the point at which $\mathrm{ed}_{\mathscr{H}}$ achieves its maximum value.

Note: Although $p^*_{\mathscr{H}}$ is formally defined to be a closed interval, in all but a few (very) interesting cases[3] the interval is degenerate. That is, $p^*_{\mathscr{H}}$ is usually a single value. We will often abuse notation and terminology by referring to $p^*_{\mathscr{H}} = p$, rather than $p^*_{\mathscr{H}} = [p,p]$ and instead indicate explicitly where the interval is not degenerate.

3.2 Upper bound using $\chi_B(\mathscr{H})$

If $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ is a hereditary property such that $\mathscr{F}$ contains no complete graph and no empty graph, then it is trivial that $\mathrm{ed}_{\mathscr{H}}(0) = \mathrm{ed}_{\mathscr{H}}(1) = 0$. Indeed, Proposition 12 gives that $\chi(\mathscr{H}) \geq 2$ and $\overline{\chi}(\mathscr{H}) \geq 2$. The statement then follows from the simple bounds given by Corollary 11.

Using only the $\gamma_{\mathscr{H}}$ function, we may narrow down the possible values for $p^*_{\mathscr{H}}$ and for $d^*_{\mathscr{H}}$.

Theorem 14 *Let $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ with $(r,s) \in \Gamma(\mathscr{H})$ such that $r+s = \chi_B(\mathscr{H}) - 1$.*

(a) $\mathrm{ed}_{\mathscr{H}}(p) \leq \gamma_{\mathscr{H}}(p) \leq \frac{p(1-p)}{r(1-p)+sp}$ *for all* $p \in [0,1]$.
(b) $d^*_{\mathscr{H}} \leq \frac{1}{r+s+2\sqrt{rs}}$.
(c) $\mathrm{ed}_{\mathscr{H}}(p) \geq \min\left\{\frac{p}{\chi_B(\mathscr{H})-1}, \frac{1-p}{\chi_B(\mathscr{H})-1}\right\}$.

[3] See, e.g. Section 5.5.2.

(d) If $r \leq s$ then $p^*_{\mathscr{H}} \in \left[\frac{r}{r+s}, \frac{1}{2}\right]$.
(e) If $s \leq r$ then $p^*_{\mathscr{H}} \in \left[\frac{1}{2}, \frac{r}{r+s}\right]$.

Theorem 14(a) comes from Theorem 10. Theorem 14(b) is simply the maximum value of $g_{K(r,s)}(p)$. Theorem 14(c) follows from Theorem 13 – continuity, concavity, and the value of $\mathrm{ed}_{\mathscr{H}}(1/2)$ – and the fact that $\mathrm{ed}_{\mathscr{H}}(0) = \mathrm{ed}_{\mathscr{H}}(1) = 0$. Theorem 14(d) and (e) follow from the fact that these are the intervals over which $g_{K(r,s)}(p) \geq 1/(2(\chi_B(\mathscr{H}) - 1))$.

Corollary 15 gives the values of $p^*_{\mathscr{H}}$ and of $d^*_{\mathscr{H}}$ if $\mathscr{H} = \mathrm{Forb}(H)$ for a self-complementary graph H.

Corollary 15 *Let $\mathscr{H} = \bigcap_{H \in \mathscr{F}} \mathrm{Forb}(H)$ with $(r_1, s_1) \in \Gamma(\mathscr{H})$ and $(r_2, s_2) \in \Gamma(\mathscr{H})$ (not necessarily distinct) such that $r_1 + s_1 = r_2 + s_2 = \chi_B(\mathscr{H}) - 1$, $r_1 \leq s_1$, and $r_2 \geq s_2$. Then*

$$p^*_{\mathscr{H}} = 1/2 \qquad \text{and} \qquad d^*_{\mathscr{H}} = 1/(2(\chi_B(\mathscr{H}) - 1)).$$

In particular, if $\mathscr{H} = \mathrm{Forb}(H)$, where H is a self-complementary graph, then $p^*_{\mathscr{H}} = 1/2$ and $d^*_{\mathscr{H}} = 1/(2(\chi_B(H) - 1))$.

4 Symmetrization

The most powerful tool for determining the edit distance function of a hereditary property is called symmetrization. This is a term Pikhurko [38] used for a method due to Sidorenko [47]. In fact, symmetrization can be traced back to Zykov [50] and his proof of Turán's theorem. Our version of symmetrization comes directly from the matrix defined by a CRG.

Theorem 16 (Martin [33]) *Let $p \in [0,1]$ and let K be a p-core CRG with associated matrix $\mathbf{M}_K(p)$, as defined in (4). If $\mathbf{x}^*$ is an optimal solution of the quadratic program from (6), namely that $\mathbf{x}^* \geq \mathbf{0}$, $\mathbf{x}^*\mathbf{1} = 1$ and $g_K(p) = (\mathbf{x}^*)^T \mathbf{M}_K(p) \mathbf{x}^*$, then*

$$\mathbf{M}_K(p) \cdot \mathbf{x}^* = g_K(p)\mathbf{1}.$$

In addition, by virtue of K being p-core, the vector $\mathbf{x}^*$ has no zero entries and $\mathbf{x}^*$ is unique for any fixed labeling of the vertices of K.

4.1 The weighted gray degree of a vertex

In order to interpret Theorem 16, we define the *white neighborhood* of vertex v in CRG K to be $N_W(v) := \{v' \in V(K) : vv' \in \mathrm{EW}(K)\} \cup \{v : \text{if } v \in \mathrm{VW}(K)\}$. The *black neighborhood* of v is $N_B(v) := \{v' \in V(K) : vv' \in \mathrm{EB}(K)\} \cup \{v : \text{if } v \in \mathrm{VB}(K)\}$. The *gray neighborhood* of v is $N_G(v) := \{v' \in V(K) : vv' \in \mathrm{EG}(K)\}$.

If $\mathbf{x}$ is the optimum weight vector in the quadratic program (6) that defines $g_K(p)$, then the *weighted white degree* of vertex $v \in V(K)$ is $\mathrm{d}_W(v) := \sum_{v' \in N_W(v)} \mathbf{x}(v')$. The *weighted black degree* of vertex $v \in V(K)$ is $\mathrm{d}_B(v) := \sum_{v' \in N_B(v)} \mathbf{x}(v')$. The *weighted gray degree* of vertex $v \in V(K)$ is $\mathrm{d}_G(v) := \sum_{v' \in N_G(v)} \mathbf{x}(v')$.

Theorem 16 gives that, for any $v \in \mathrm{VW}(K)$,

$$p\mathrm{d}_W(v) + (1-p)\mathrm{d}_B(v) = g_K(p). \tag{11}$$

Using the characterization of p-core CRGs from Theorem 9, we can apply (11) to compute the gray degree of each vertex.

Theorem 17 (Martin [33]) *Let $p \in (0,1)$ and K be a p-core CRG with optimum weight vector* $\mathbf{x}$.

(a) If $p \leq 1/2$ then $\mathbf{x}(v) = g_K(p)/p$ for all $v \in \mathrm{VW}(K)$ and

$$\mathrm{d}_G(v) = \frac{p - g_K(p)}{p} + \frac{1-2p}{p}\mathbf{x}(v), \qquad \textit{for all } v \in \mathrm{VB}(K).$$

(b) If $p \geq 1/2$ then $\mathbf{x}(v) = g_K(p)/(1-p)$ for all $v \in \mathrm{VB}(K)$ and

$$\mathrm{d}_G(v) = \frac{1 - p - g_K(p)}{1-p} + \frac{2p-1}{1-p}\mathbf{x}(v), \qquad \textit{for all } v \in \mathrm{VW}(K).$$

Most of the results below use Theorem 17 as a primary tool. Intuitively, if $g_K(p)$ is small, then $\mathrm{d}_G(v)$ is large for each vertex and so K has a large amount of gray. However, if K has too much gray, then some $H \in \mathscr{F}(\mathscr{H})$ would map to K, which contradicts the choice of $K \in \mathscr{K}(\mathscr{H})$. This general paradigm is made more precise by knowing more about the structure of the CRGs $K \in \mathscr{K}(\mathscr{H})$.

4.2 Basic structural facts of p-core CRGs

We can use Theorem 17 to obtain some basic helpful results on certain types of CRGs:

Corollary 18 *Let $t \geq 2$ and $k \geq 2$ be integers.*

(a) Let $p \leq 1/2$ and let K be a p-core CRG on k black vertices.

  (i) If K has no gray edges, then $g_K(p) = \frac{1}{k}[1 + (k-2)p]$.
  (ii) If K has no gray clique of order t, then $g_K(p) > p/(t-1)$.

(b) Let $p \geq 1/2$ and let K be a p-core CRG on k white vertices.

  (i) If K has no gray edges, then $g_K(p) = \frac{1}{k}[1 + (k-2)(1-p)]$.
  (ii) If K has no gray clique of order t, then $g_K(p) > (1-p)/(t-1)$.

Proof. By symmetry, it is sufficient to prove (a). For (a)(i) we observe that, by Theorem 9(a), all edges are white and it is easy to see that the optimum weight vector in equation (6) is constant. Thus, all vertices have the same weight and the result follows.

For (a)(ii), we use a well-worn trick, used e.g. in [35]. Let the maximum-sized clique of K (in terms of the number of vertices) be on vertex set $\{v_1,\ldots,v_c\}$ where $c \geq 2$. For every $w \notin \{v_1,\ldots,v_c\}$ we know that wv_i is a gray edge for at most $c-1$ values of i. Using Theorem 17(a), we have

$$\sum_{i=1}^{c}\left(\frac{p-g_K(p)}{p}+\frac{1-2p}{p}\mathbf{x}(v_i)\right) \leq (c-1)\left(1-\sum_{i=1}^{c}\mathbf{x}(v_i)\right)$$
$$(c-3+1/p)\sum_{i=1}^{c}\mathbf{x}(v_i)+1 \leq \frac{c}{p}g_K(p).$$

Since $c-3+1/p \geq c-1 > 0$, we can conclude that $g_K(p) > p/c \geq p/(t-1)$, which concludes the proof. □

Remark 19 *The bound in Corollary 18(a)(ii) can be approached by a CRG on black vertices where the gray edges induce a blow-up of* K_{t-1}.

5 Known results

5.1 Hereditary properties that forbid either a complete or an empty graph

If $\mathscr{H} \subseteq \mathrm{Forb}(K_h)$ then $\mathrm{ed}_{\mathscr{H}}(1) > 0$ and, by Proposition 4, if $\mathscr{H} \subseteq \mathrm{Forb}(\overline{K_h})$ then $\mathrm{ed}_{\mathscr{H}}(0) > 0$. We can produce bounds on the edit distance function for such properties.

Theorem 20 (Martin [33]) *Let* $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})}\mathrm{Forb}(H)$ *be a nontrivial hereditary property with* $\mathscr{H} \subseteq \mathrm{Forb}(K_h)$ *for some* $h \geq 2$ *such that*

- *m is the least positive integer such that $\mathscr{F}(\mathscr{H})$ contains a complete multipartite graph with m parts, and*
- *χ is the chromatic number of $\mathscr{H}$.*

Note $\chi \geq 2$ because $\mathscr{H}$ is nontrivial and $\chi \leq m \leq h$. Then

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p}{\chi-1}$, *for all* $p \in [0,1/2]$, *and*
(b) $\frac{1-p}{\chi-1}+\frac{2p-1}{m-1} \leq \mathrm{ed}_{\mathscr{H}}(p) \leq \min\left\{1-p+\frac{2p-1}{m-1},\frac{p}{\chi-1}\right\}$, *for all* $p \in [1/2,1]$.

In particular, if $\mathscr{H} = \mathrm{Forb}(K_h)$ *then* $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p}{h-1}$.

Remark 21 *The bound* $\mathrm{ed}_{\mathscr{H}}(p) \leq p/(\chi-1)$ *in Theorem 20 was not expressed explicitly in [33] but follows directly from the concavity of the edit distance function. By Proposition 4, there are similar bounds for* $\mathscr{H}$ *where* $\mathscr{H} \subseteq \mathrm{Forb}(\overline{K_h})$. *Consequently,* $\mathrm{ed}_{\mathrm{Forb}(\overline{K_h})}(p) = (1-p)/(h-1)$.

5.2 C_6^* and H_9

5.2.1 $\mathrm{Forb}(C_6^*)$

In [31], Marchant and Thomason address the graph C_6^*, which is the 6-cycle with a diagonal. The extreme points of the clique spectrum of $\mathrm{Forb}(C_6^*)$ are $(1,1)$ and $(0,2)$. Thus, if $\mathscr{H} = \mathrm{Forb}(C_6^*)$, then $\gamma_{\mathscr{H}}(p) = \min\{p(1-p),(1-p)/2\}$.

In fact, the edit distance function has a smaller value for $p \in (0,1)$.

Theorem 22 (Marchant-Thomason [31]) *Let* $\mathscr{H} = \mathrm{Forb}(C_6^*)$*, where* C_6^* *is the* 6-*cycle with a diagonal.*

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{1+2p}, \frac{1-p}{2}\right\}$*, for* $p \in [0,1]$.
(b) $p^*_{\mathscr{H}} = 1/2$ *and* $d^*_{\mathscr{H}} = 1/4$.

The CRG that corresponds to the $p/(1+2p)$ part of the function has 1 white vertex, 2 black vertices, one white edge between the black vertices and two gray edges incident to the white vertex. See Figure 3. Although the edit distance function cannot be determined by the clique spectrum, the values of $p^*_{\mathscr{H}} = 1/2$ and $d^*_{\mathscr{H}} = 1/4$ can be computed by knowing only the clique spectrum.

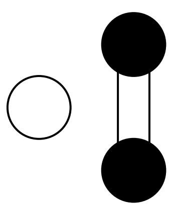

Fig. 3 The 3-vertex CRG that gives $p/(1+2p)$ in Theorem 22. The white edge is indicated, the two gray edges are not.

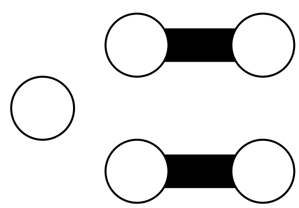

Fig. 4 The 5-vertex CRG that gives $p/(1+4p)$ in Theorem 23. The two black edges are indicated, the eight gray edges are not.

Using the tools in Section 4, the proof of Theorem 22 is much easier than the original proof.

Proof. By Theorem 13, we may use continuity, concavity and the knowledge of the value at $p = 1/2$ to conclude that $\mathrm{ed}_{\mathscr{H}}(p) = (1-p)/2$ for all $p \in [1/2,1]$. Let $p \in [0,1/2)$ and K be a p-core CRG. If $C_6^* \not\mapsto K$ and K has no white vertices, then it has no gray triangle and Corollary 18(a)(ii) gives that $g_K(p) > p/2$. If $C_6^* \not\mapsto K$ and K has one white vertex, then there are at most two black vertices, which cannot have a gray edge. Corollary 18(a)(ii) and Proposition 8 give that $g_K(p) \geq \left(p^{-1} + (1/2)^{-1}\right)^{-1} = p/(1+2p)$, as required. □

5.2.2 $\mathrm{Forb}(H_9)$

Balogh and Martin [16] introduced the graph H_9, which is drawn in Figure 1. For $\mathscr{H} = \mathrm{Forb}(H_9)$, the values of $p^*_{\mathscr{H}}$ and $d^*_{\mathscr{H}}$ cannot be determined by the clique spectrum and this was established in [16]. Later, the author determined the edit distance function completely.

Theorem 23 (Martin [34]) *Let H_9 be the graph drawn in Figure 1 and let $\mathscr{H} = \mathrm{Forb}(H_9)$.*

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p}{1+4p}, \frac{1-p}{2}\right\}$ *for $p \in [0,1]$.*
(b) $p^*_{\mathscr{H}} = \frac{1}{8}(1+\sqrt{17})$ *and* $d^*_{\mathscr{H}} = \frac{1}{8}(7-\sqrt{17})$.

The CRG that corresponds to the $p/(1+4p)$ part of the function has 5 white vertices, 2 nonadjacent white edges and the remaining 8 edges gray. See Figure 4.

5.3 Cycles

The case of $\mathrm{Forb}(C_h)$, where C_h is a cycle on $h \geq 3$ vertices, has been widely investigated. Theorem 20 gives immediately that $\mathrm{ed}_{\mathrm{Forb}(C_3)}(p) = \mathrm{ed}_{\mathrm{Forb}(K_3)}(p) = p/2$.

In her Master's thesis, Peck almost completely settled the edit distance function for hereditary properties that forbid a cycle. Utilizing techniques inspired by the cycle arguments of Pósa [39], she determined the edit distance function for $\mathrm{Forb}(C_h)$ for odd $h \geq 5$. For even $h \geq 4$, she was able to determine enough of the function to find the maximum.

Theorem 24 (Peck [37]) *Let $\mathscr{H} = \mathrm{Forb}(C_h)$ where C_h is the cycle on $h \geq 4$ vertices.*

(a) If h is odd, then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{2}, \frac{p(1-p)}{1-p+(\lceil h/3\rceil-1)p}, \frac{1-p}{\lceil h/2\rceil-1}\right\}$ *for all $p \in [0,1]$.*
(b) If h is even, then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{1-p+(\lceil h/3\rceil-1)p}, \frac{1-p}{\lceil h/2\rceil-1}\right\}$ *for all $p \in [\lceil h/3\rceil^{-1}, 1]$.*

Marchant and Thomason [31] first proved the case of $h = 4$ and, in fact, proved

Theorem 25 (Marchant-Thomason [31]) $\mathrm{ed}_{\mathrm{Forb}(C_4)}(p) = p(1-p)$ *for all $p \in [0,1]$.*

Marchant [32] proved the case for $h = 5, 7$. The cases of $h = 6, 8, 9, 10$ were first proven in [33] and, in fact, a larger range of p was proven in [33] for small even h.

Theorem 26 (Martin [33]) *Let $\mathscr{H} = \mathrm{Forb}(C_h)$, where C_h is the cycle on $h \geq 4$ vertices.*

(a) *If* $h=6$ *then* $\mathrm{ed}_{\mathscr{H}}(p)=\min\left\{p(1-p),\frac{1-p}{2}\right\}$ *for all* $p\in[0,1]$.

(b) *If* $h=8$ *then* $\mathrm{ed}_{\mathscr{H}}(p)=\min\left\{\frac{p(1-p)}{1+p},\frac{1-p}{3}\right\}$ *for all* $p\in[0,1]$.

(c) *If* $h=10$ *then* $\mathrm{ed}_{\mathscr{H}}(p)=\min\left\{\frac{p(1-p)}{1+2p},\frac{1-p}{4}\right\}$ *for all* $p\in[1/7,1]$.

Corollary 27 *Let* $\mathscr{H}=\mathrm{Forb}(C_h)$ *where* C_h *is the cycle on* $h\geq 4$ *vertices.*

(a) *If* $h\notin\{4,7,8,10,16\}$ *then*

$$p^*_{\mathscr{H}}=\frac{1}{\lceil h/2\rceil-\lceil h/3\rceil+1}\qquad\text{and}\qquad d^*_{\mathscr{H}}=\frac{\lceil h/2\rceil-\lceil h/3\rceil}{(\lceil h/2\rceil-1)(\lceil h/2\rceil-\lceil h/3\rceil+1)}.$$

(b) *If* $h\in\{4,7,8,10,16\}$ *then*

$$p^*_{\mathscr{H}}=\frac{1}{1+\sqrt{\lceil h/3\rceil-1}}\qquad\text{and}\qquad d^*_{\mathscr{H}}=\frac{1}{\lceil h/3\rceil+2\sqrt{\lceil h/3\rceil-1}}.$$

It is interesting that $p^*_{\mathrm{Forb}(C_h)}$ and $d^*_{\mathrm{Forb}(C_h)}$ are both rational and result from the intersection of the g functions of two p-core CRGs, except in the cases $h\in\{7,8,10,16\}$.

5.4 Powers of cycles

A natural extension of hereditary properties defined by forbidding certain cycles are hereditary properties defined by forbidding certain powers of cycles. For $h\geq 2t+1$, we define C_h^t to be the graph with vertex set $\{1,\ldots,h\}$ and $ij\in E(C_h^t)$ if and only if $|i-j|\leq t \pmod h$. We consider the case for $t=2$, that is, the case of the squared cycle.

For $h=5$, C_5^2 is complete and from Theorem 20 we see that

$$\mathrm{ed}_{\mathrm{Forb}(C_5^2)}(p)=\mathrm{ed}_{\mathrm{Forb}(K_5)}(p)=p/4.$$

In the case of C_6^2, the complement is a perfect matching and we can use Proposition 4 if we know the edit distance function for $\mathrm{Forb}(M_6)$, where M_6 is the perfect matching on 6 vertices. It is easy to see that the CRGs $K\in\mathscr{F}(\mathrm{Forb}(M_6))$ have at most 2 black vertices and no pair of white vertices can have a gray edge between them, otherwise $M_6\mapsto K$. By Theorem 9(a), we can conclude that $\mathrm{ed}_{\mathrm{Forb}(M_6)}=\frac{p(1-p)}{1+p}$ for $p\in[0,1/2]$. Some more work verifies that $\mathrm{ed}_{\mathrm{Forb}(M_6)}=\frac{p(1-p)}{1+p}$ for $p\in[1/2,1]$ also. Hence,

$$\mathrm{ed}_{\mathrm{Forb}(C_6^2)}(p)=\mathrm{ed}_{\mathrm{Forb}(M_6)}(1-p)=\frac{p(1-p)}{2-p}.$$

The complement of C_7^2 is simply C_7 and so Proposition 4 gives

$$\mathrm{ed}_{\mathrm{Forb}(C_7^2)}(p) = \mathrm{ed}_{\mathrm{Forb}(C_7)}(1-p) = \min\left\{\tfrac{p}{3}, \tfrac{p(1-p)}{2-p}, \tfrac{1-p}{2}\right\}.$$

Peck established some more values of $\mathrm{ed}_{\mathrm{Forb}(C_h^2)}(p)$ for $h \in \{8,9,10\}$, which we give in Theorem 28.

Theorem 28 (Peck [37]) *Let $\mathscr{H} = \mathrm{Forb}(C_h^2)$ where C_h^2 is the square of the cycle on h vertices.*

(a) If $\mathscr{H} = \mathrm{Forb}(C_8^2)$ then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p(1-p)}{2-p}, \frac{1-p}{2}\right\}$ for all $p \in [0,1]$.

(b) If $\mathscr{H} = \mathrm{Forb}(C_9^2)$ then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{2-p}, \frac{p(1-p)}{1+p}\right\}$ for all $p \in [0,1]$.

(c) If $\mathscr{H} = \mathrm{Forb}(C_{10}^2)$ then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{1-p}{3}\right\}$ for all $p \in [0,1]$.

(d) If $\mathscr{H} = \mathrm{Forb}(C_{11}^2)$ then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p(1-p)}{2}\right\}$ for all $p \in [0,1/2]$ and $\mathrm{ed}_{\mathscr{H}}(p) \le \min\left\{\frac{p(1-p)}{2}, \frac{1-p}{3}\right\}$ for all $p \in [1/2,1]$.

(e) If $\mathscr{H} = \mathrm{Forb}(C_{12}^2)$ then $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p(1-p)}{2}$ for all $p \in [0,1/2]$ and $\mathrm{ed}_{\mathscr{H}}(p) \le \min\left\{\frac{p(1-p)}{2}, \frac{1-p}{3}\right\}$ for all $p \in [1/2,1]$.

Theorem 28, together with Theorem 13 are enough to determine the value of $p^*_{\mathrm{Forb}(C_h^2)}$ and of $d^*_{\mathrm{Forb}(C_h^2)}$ for $h \in \{5,\dots,12\}$.

In work in progress, Berikkyzy, Peck and Martin have extended the results from [37] to apply to powers of cycles, provided the number of vertices is large enough.

Theorem 29 (Berikkyzy-Martin-Peck [17]) *Let $\mathscr{H} = \mathrm{Forb}(C_h^t)$ where C_h^t is the t^{th} power of the cycle on h vertices. For $t \ge 1$ and h sufficiently large, let $\ell_0 = \lceil h/(t+1)\rceil$, $\ell_t = \lceil h/(2t+1)\rceil$ and $p_0 = \ell_t^{-1}$.*

(a) If $(t+1) \nmid h$ then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{t+1}, \frac{p(1-p)}{t(1-p)+(\ell_t-1)p}, \frac{1-p}{\ell_0-1}\right\}$ for $p \in [0,1]$.

(b) If $(t+1) \mid h$, then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{t(1-p)+(\ell_t-1)p}, \frac{1-p}{\ell_0-1}\right\}$ for $p \in [p_0,1]$.

If $t = 2$, then $h \ge 13$ suffices for Theorem 29. In general, the bound that is proven to suffice is $h \ge 4t^2 + \Omega(t)$ although this is likely not best possible.

5.5 Complete bipartite graphs

5.5.1 $\mathrm{Forb}(K_{s,s})$

The case of $\mathscr{H} = \mathrm{Forb}(K_{2,2})$ was established by Marchant and Thomason [31] where it was shown that $\mathrm{ed}_{\mathrm{Forb}(K_{2,2})}(p) = p(1-p)$ for all $p \in [0,1]$.

In the case of $\mathscr{H} = \mathrm{Forb}(K_{3,3})$, the values of $p^*_{\mathrm{Forb}(K_{3,3})} = \sqrt{2}-1$ and $d^*_{\mathrm{Forb}(K_{3,3})} = 3-2\sqrt{2}$ were established by Balogh and Martin [16].

For p not too small, the edit distance function for $\mathrm{Forb}(K_{3,3})$ coincides with $\gamma_{\mathrm{Forb}(K_{3,3})}(p)=\frac{p(1-p)}{1+p}$, but for p very small, the edit distance function is strictly smaller.

Theorem 30 (Marchant-Thomason [31]) *Let $\mathscr{H}=\mathrm{Forb}(K_{3,3})$ where $K_{3,3}$ is the complete bipartite graph with 3 vertices in each part.*

(a) $\mathrm{ed}_{\mathscr{H}}(p)=\frac{p(1-p)}{1+p}$, *for* $p\in[1/9,1]$.
(b) $\mathrm{ed}_{\mathscr{H}}(p)<\frac{p(1-p)}{1+p}$, *for* $p\in(0,1/124]$.

The CRGs used to establish Theorem 30(b) are defined by constructions due to Brown [22] to address a related Zarankiewicz problem. Specifically, for a prime power r, the constructions are (r^2-r)-regular bipartite graphs on $2r^3$ vertices. Such graphs have no copy of $K_{3,3}$ and, of course, no copy of K_3. For such a graph G, we construct the CRG K for which the vertices of G are (black) vertices of K, the edges of G are gray edges of K and the nonedges of G are white edges of K. By (5), this construction gives

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

With $r=19$, we obtain strict inequality for $p=1/124$ and the continuity and concavity of the edit distance function gives Theorem 30(b) for all $p\leq 1/124$.

It is also established in [31] that the value of $d^*_{\mathrm{Forb}(K_{s,s})}$ cannot be determined by the clique spectrum. The only extreme point of the clique spectrum is $(1,s-1)$ and the resulting CRG has $g_{K(1,s-1)}=\frac{p(1-p)}{1+(s-2)p}=\gamma_{\mathrm{Forb}(K_{s,s})}(p)$. The construction is a CRG $K^{(s-1)}$ on $2s-2$ black vertices consisting of $s-1$ disjoint white edges. It is easy to show that $K_{s,s}\not\mapsto K^{(s-1)}$ and since the g function of each component is $1/2$, Proposition 8 gives $g_{K^{(2s-2)}}(p)=1/(2s-2)$.

So, $g_{K^{(2s-2)}}(p)$ is less than the maximum value of $\gamma_{\mathrm{Forb}(K_{s,s})}(p)$ for $s\geq 7$. We ask in Problem 2 if $1/(2s-2)$ is, indeed, the maximum value of $\mathrm{ed}_{\mathrm{Forb}(K_{s,s})}(p)$ and if that value is achieved for a positive length interval.

5.5.2 $\mathrm{Forb}(K_{2,t})$

McKay and Martin [35] establish some surprising results for the hereditary property $\mathrm{Forb}(K_{2,t})$.

Theorem 31 (Martin-McKay [35]) *Let $\mathscr{H}=\mathrm{Forb}(K_{2,t})$ where $K_{2,t}$ is the complete bipartite graph with 2 vertices in one part and t vertices in the other part. For all $t\geq 2$, $\gamma_{\mathscr{H}}(p)=\min\{p(1-p),(1-p)/(t-1)\}$.*

(a) If $t=3$ then $\mathrm{ed}_{\mathscr{H}}(p)=\min\left\{p(1-p),\frac{1-p}{2}\right\}$ *for all* $p\in[0,1]$.
(b) If $t=4$ then $\mathrm{ed}_{\mathscr{H}}(p)=\min\left\{p(1-p),\frac{7p+1}{15},\frac{1-p}{3}\right\}$ *for all* $p\in[0,1]$.
(c) If $t\geq 5$ and is odd, then $d^*_{\mathscr{H}}=\frac{1}{t+1}$ *and* $p^*_{\mathscr{H}}\supseteq\left[\frac{2t-1}{t(t+1)},\frac{2}{t+1}\right]$.

(d) If $t \geq 9$, there exists a $p_0(t) < 1/2$ such that $\mathrm{ed}_{\mathscr{H}}(p) < p(1-p)$.

There are a number of interesting consequences arising from the study of these hereditary properties. First, we note that the CRG that gives the portion of the function in Theorem 31(b) corresponding to $(1+7p)/15$ results from a strongly-regular graph construction.

Definition 32 *A* (n,k,λ,μ)-strongly regular graph *or* (n,k,λ,μ)-SRG *is a k-regular graph on n vertices for which each pair of adjacent vertices has exactly λ common neighbors and for which each pair of nonadjacent vertices has exactly μ common neighbors.*

The CRG we use for Theorem 31(b) is constructed from a $(15,6,1,3)$-SRG, commonly called $\mathrm{GQ}(2,2)$. It is a member of the family of so-called generalized quadrangles. Given a $\mathrm{GQ}(2,2)$ G', the CRG K' has 15 white vertices that correspond to the vertices of the graph. An edge of K' is gray if and only if the corresponding pairs of vertices are adjacent in G'. See Figure 5 for the graph $K_{2,4}$ and Figure 6 for the 15-vertex CRG mentioned above. In [35], it is shown that $K_{2,4} \not\mapsto K'$ and $g_{K'}(p) = (1+7p)/15$.

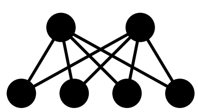

Fig. 5 The complete bipartite graph $K_{2,4}$.

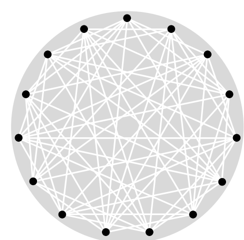

Fig. 6 The 15-vertex CRG that gives $(1+7p)/15$ in Theorem 31(b). The white edges are shown. The remaining edges are gray, and form a graph isomorphic to $\mathrm{GQ}(2,2)$.

Similar constructions from strongly regular graphs are in $\mathscr{K}(\mathrm{Forb}(K_{2,t}))$ and have a smaller g function than $\gamma_{\mathscr{H}}(p) = \min\{p(1-p),(1-p)/(t-1)\}$ for certain values of t and p.

For Theorem 31(c), the corresponding CRG $K^{(t)}$ has $t+1$ black vertices and a perfect matching of $(t+1)/2$ white edges. The remaining edges are gray. It is easy to show that $K_{2,t} \not\mapsto K^{(t+1)}$ and $g_{K^{(t+1)}}(p) = 1/(t+1)$.

For Theorem 31(d), the constructions are due to Füredi [28] to address a related Zarankiewicz problem. If q is a prime power such that $t-1$ divides $q-1$, then there exists a graph on $2(q^2-1)/(t-1)$ vertices that is q-regular with no copy of $K_{2,t}$ and no triangle. This is enough to ensure that $K_{2,t}$ does not map to the corresponding CRG. By (5), this construction gives

$$f_K(p) = \frac{t-1}{2(q^2-1)}\left[1+p\left(\frac{2(q^2-1)}{t-1}-q-2\right)\right]$$

With $t \geq 9$, we can find a sufficiently large prime power q so that $f_k(p) < p(1-p)$.

5.6 Split graphs

A graph H on at least two vertices is a *split graph* if there is a partition of $V(H)$ into one independent set and one clique. For a split graph with independence number $\alpha \geq 2$ and clique number $\omega \geq 2$, either $\alpha+\omega = h$ or $\alpha+\omega = h+1$. We can compute the edit distance function of hereditary properties defined by such graphs.

Theorem 33 (Martin [34]) *Let H be a split graph which has independence number $\alpha = \alpha(H) \geq 2$ and clique number $\omega = \omega(H) \geq 2$. If $\mathscr{H} = \mathrm{Forb}(H)$ then*

$$\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{\omega-1}, \frac{1-p}{\alpha-1}\right\}.$$

Hence $p^*_{\mathscr{H}} = (\omega-1)/(\alpha+\omega-2)$ and $d^*_{\mathscr{H}} = 1/(\alpha+\omega-2)$.

6 Quantities related to the edit distance function

We get the following notation from Balogh, et al. [11]. For a graph property[4] $\mathscr{H}$, the *labeled slice* of $\mathscr{H}$ is the set $\mathscr{H}^n$ of graphs in $\mathscr{H}$ with vertex set $\{1,\ldots,n\}$. The *labeled speed* of $\mathscr{H}$ is the function $n \mapsto |\mathscr{H}^n|$.

Theorem 34 ([13, 14, 15, 19]) *If $\mathscr{H}$ is a hereditary property of graphs then one of the following holds:*

(i) There exist $N, k \in \mathbb{N}$ and polynomials $\{p_i(n)\}_{i=0}^k$ such that, for all $n > N$, $|\mathscr{H}^n| = \sum_{i=0}^k p_i(n) i^n$.
(ii) For some $t \in \mathbb{N}$, $t > 1$, we have $|\mathscr{H}^n| = n^{(1-1/t+o(1))n}$.
(iii) For n sufficiently large, $n^{(1+o(1))n} \leq |\mathscr{H}^n| \leq 2^{o(n^2)}$.
(iv) For some $k \in \mathbb{N}$, $k > 1$, we have $|\mathscr{H}^n| = 2^{(1-1/k+o(1))n^2/2}$.

Here $k = \chi_B(\mathscr{H}) - 1$.

This partition of hereditary properties was first discovered by Scheinerman and Zito [46]. As for the precise results, parts (i) and (ii) were established by Balogh, Bollobás and Weinreich [13], part (iii) was also established by Balogh, Bollobás and Weinreich [14, 15] and part (iv) was established by Bollobás and Thomason [19].

[4] For us, the property will be hereditary, although that is not necessary in order to define the speed.

Part (iv) has been well-studied, including the $o(1)$ error term [12]. It has also been generalized. Bollobás and Thomason [20] define $c_p(\mathscr{H})$ as follows:

$$c_{\mathscr{H}}(p) = \lim_{n\to\infty} -\log_2 \Pr(G(n,p) \in \mathscr{H}) / \binom{n}{2}. \tag{12}$$

They showed that the limit exists, based on work by Alekseev [1] (see also [2]). Thomason [49] compiled these results to show the relationship to the edit distance function:

Theorem 35 (Thomason [49]) *Let $\mathscr{H}$ be a nontrivial hereditary property and let $c_{\mathscr{H}}(p)$ be defined as in (12). Then*

$$c_{\mathscr{H}}(p) = (-\log_2 p(1-p))\,\mathrm{ed}_{\mathscr{H}}\left(\frac{\log_2(1-p)}{\log_2 p(1-p)}\right).$$

As we see, $c_{\mathscr{H}}(p)$ can be derived directly from the edit distance function.

Remark 36 *The function $c_{\mathscr{H}}(p)$ is not necessarily concave down, however $\mathrm{ed}_{\mathscr{H}}(p)$ is. Concavity is a key tool in finding the elusive lower bounds on the edit distance function which can then be used to compute lower bounds for $c_{\mathscr{H}}(p)$.*

Perhaps other functions of hereditary properties can be defined from the edit distance function. We are particularly interested in other metrics. For each positive integer n, let d be a metric on the space of graphs with vertex set $\{1,\ldots,n\}$. For any hereditary property $\mathscr{H}$, define

$$\mathrm{d}(G,\mathscr{H}) = \min\left\{\mathrm{d}(G,G') : V(G') = V(G), G' \in \mathscr{H}\right\},$$

and define the function

$$\phi_{\mathscr{H}}(p) := \limsup_{n\to\infty} \max\left\{\mathrm{d}(G,\mathscr{H}) : |V(G)| = n, |E(G)| = \left\lfloor p\binom{n}{2}\right\rfloor\right\}. \tag{13}$$

In Question 3 from Section 8.2, we ask whether $\phi_{\mathscr{H}}(p)$ can be expressed as a function of p and $\mathrm{ed}_{\mathscr{H}}(\cdot)$ if d satisfies a natural property. Recall that dist represents the edit metric. It is clear that in order for any such result to exist, the metric d should be *continuous with respect to the edit metric*. That is, for every $\varepsilon > 0$, there exists a δ such that $\mathrm{d}(G,G') < \varepsilon$ whenever $\mathrm{dist}(G,G') < \delta$. This is a natural restriction because, for example, the trivial metric where $\mathrm{d}(G,G) = 0$ but $\mathrm{d}(G,G') = 1$ when $G \neq G'$ produces no useful results.

A well-studied metric is the so-called *cut metric* (or *cut norm*). Introduced by Frieze and Kannan [27] and investigated further for graph limits (see, e.g., Borgs et al. [21]) is defined as follows for graphs on the same labeled vertex set $V = \{1,\ldots,n\}$:

$$\mathrm{d}_{\square}(G,G') = \max_{S,T\subset V} \frac{1}{n^2}\left|e_G(S,T) - e_{G'}(S,T)\right|,$$

where $e_G(S,T)$ is the number of ordered pairs (i,j) with $i \in S$ and $j \in T$ and $ij \in E(G)$. If S and T are disjoint, it counts the number of edges between S and T in G.

The cut metric is useful for comparing random graphs. Although two typical graphs selected according to $G(n,p)$ have edit distance close to $2p(1-p)$, their $\mathrm{d}_\square$ distance is $O(1/n)$.

7 Generalizations of edit distance

Axenovich and Martin have investigated natural generalizations of the edit distance problem. The paper [9] addressed editing matrices (Section 7.1 below). The paper [10] addressed both editing the edges of multicolorings of a complete graph (Section 7.2) and editing the edges of a directed graph (Section 7.3).

7.1 Matrices

Let $\mathscr{A} = \{A_1,\ldots,A_r\}$ be a partition of pairs from $[m]\times[n]$ into r nonemtpy classes. An $m\times n$ matrix $A = (a_{ij})$ is said to have a *pattern* $\mathscr{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i,j),(i',j') \in A_t$ for some $t \in \{1,\ldots,r\}$. A pattern is *non-trivial* if $r \geq 2$. For a matrix M, if there is a submatrix M' with pattern $\mathscr{A}$ then we say that M has a *subpattern* $\mathscr{A}$.

For a pattern $\mathscr{A}$ and positive integers m,n,s, we define $\mathrm{Forb}(m,n;s,\mathscr{A})$ to be the set of all $m\times n$ matrices with at most s distinct entries and not containing subpattern $\mathscr{A}$.

For two matrices A and B of the same dimensions, we say that $\mathrm{dist}(A,B)$ is the number of positions in which A and B differ; i.e., it is the matrix Hamming distance. For a class of matrices $\mathscr{F}$ and a matrix A, all of the same dimensions, we denote $\mathrm{dist}(A,\mathscr{F}) = \min\{\mathrm{dist}(A,\mathscr{F}) : F \in \mathscr{F}\}$. Finally,

$$f(m,n;s,\mathscr{A}) := \max\left\{\mathrm{dist}(A,\mathscr{F}) : A \in \mathscr{M}(m,n;s), \mathscr{F} = \mathrm{Forb}(m,n;s,\mathscr{A})\right\}/mn. \tag{14}$$

The function f in (14) counts the maximum proportion of edits required to remove a pattern with r places from an $m\times n$ matrix with s distinct entries.[5]

Theorem 37 (Axenovich-Martin [9]) *Let s,r be positive integers, $s \geq r$. Let b_1,b_2 be positive constants such that $b_1 \leq m/n \leq b_2$. Let $\mathscr{A}$ be a non-trivial pattern with r distinct entries. Then*

$$f(m,n;s,\mathscr{A}) = (1+o(1))\left(\frac{s-r+1}{s}\right).$$

[5] In [9], f counts the number of edits, but we normalize by dividing by mn to make it consistent with the rest of this paper.

Without loss of generality, the case of $s = 2$ corresponds to a $\{0,1\}$-matrix. If the pattern has both zeros and ones, then $r = 2$ and the edit distance is $1/2$; i.e., an asymptotically most efficient editing algorithm is to make all entries zero or all entries one, whichever is most prevalent and the worst case is that there is the same number of each. If the pattern has, say only zeroes, then the edit distance is 1 because the worst case is that the original matrix is all zeros and almost all of them must be changed to one.

The setting for matrices is identical to the case of editing the $m \times n$ complete bipartite graph in which the edges are colored with s distinct colors.

7.2 Multicolor edit distance

We will use slightly different terminology from [10] so as not to confuse it with similar notation for hypergraphs in Section 8.1. For any integer $r \geq 2$, an *r-colored-graph* is pair (V, c) such that V is a finite labeled set and $c : E \to \{1, \ldots, r\}$. For r-colored graph G and $\rho \in \{1, \ldots, r\}$, we denote $E_\rho(G)$ to be the set of edges colored ρ.

If G and G' are two r-colored graphs on the same labeled vertex set, then the edit distance between them, $\mathrm{dist}(G, G')$ is the proportion of edges that receive a different color. For example, if $r = 2$ then graphs correspond to black edges and the complement corresponds to white edges. The following definitions for $r = 2$ are consistent with the graph case.

Further, we may define $\mathrm{dist}(G, \mathscr{H})$ for any hereditary property of r-colored-graphs as in (1). In this setting, a property is still hereditary if it is closed under isomorphism and the deletion of vertices. For any r-colored-graph H, we write $\mathrm{Forb}(H)$ to be the set of all r-colored-graphs that have no copy of H. Note that "induced" is not necessary here because all edges receive a color. For an integer $r \geq 2$, a *density vector* $\mathbf{p} = (p_1, \ldots, p_r)$ is a nonnegative real vector with the property that $\sum_{\rho=1}^{r} p_i = 1$. The domain of r dimensional density vectors is the *(standard)* $(r-1)$*-simplex*.

If $\mathscr{H}$ is a hereditary property of r-colored-graphs then we may define the edit distance function parallel to (3) as follows.

$$\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) := \lim_{n \to \infty} \max \left\{ \mathrm{dist}(G, \mathscr{H}) : |V(G)| = n, |E_\rho(G)| = p_\rho \binom{n}{2}, \rho = 1, \ldots, r \right\}.$$

The limit was proven to exist in [10]. We omit floors and ceilings in defining $|E_\rho(G)|$ because they play no role in the limit.

We can also define the equivalent of CRGs in this setting. In [10], the term *type* is used, though for consistency of this paper, we will just call them r-CRGs.[6]

[6] In Section 8.1, we refer to r-CRHs when discussing the edit distance on r-uniform hypergraphs.

Definition 38 *An r-CRG K is a pair (U,ϕ) where U is a finite set of vertices and $\phi : U \times U \to 2^{\{1,\dots,r\}} - \emptyset$ such that $\phi(x,y) = \phi(y,x)$ and $\phi(x,x) \neq \{1,\dots,r\}$. The* sub-$r$-CRG induced by $W \subseteq U$ *is the r-CRG that results from deleting $U - W$.*

We say that an r-colored-graph $H = (V,c)$ embeds in r-CRG K, *and write $H \mapsto K$, if there is a map $\gamma : V \to U$ such that $c(vv') = c_0$ implies $c_0 \in \phi(\gamma(v)\gamma(v'))$. For any hereditary property $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$, let $\mathscr{K}(\mathscr{H})$ be the set of r-CRGs for which none of $\mathscr{F}(\mathscr{H})$ embeds in that r-CRG.*

The notion of the binary chromatic number is more complicated in the r-colored-graph case when $r > 2$. There are weak and strong colorings.

Definition 39 *Let $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$ be a hereditary property of r-colored-graphs.*

- *An r-tuple $(a_1,\dots,a_r)$ of nonnegative integers is* weakly-good *if for some $H \in \mathscr{F}(\mathscr{H})$ the vertex set $V(H)$ can be partitioned into sets $S_1,\dots,S_r$ such that for each $\rho \in \{1,\dots,r\}$ with $a_\rho \neq 0$, the partition can be further refined $S_\rho = V_{\rho,1} \cup \dots \cup V_{\rho,a_\rho}$ where each edge in $V_{\rho,j}$* ***does not*** *have color ρ.*
- *An r-tuple $(a_1,\dots,a_r)$ of nonnegative integers is* strongly-good *if for some $H \in \mathscr{F}(\mathscr{H})$ the vertex set $V(H)$ can be partitioned into sets $S_1,\dots,S_r$ such that for each $\rho \in \{1,\dots,r\}$ with $a_\rho \neq 0$, the partitioned can be further refined $S_\rho = V_{\rho,1} \cup \dots \cup V_{\rho,a_\rho}$ where each edge in $V_{\rho,j}$* ***must have*** *color ρ.*

We can then define spectra and r-ary chromatic numbers based on weak and strong colorings.

Definition 40
- *The* weak clique spectrum *of $\mathscr{H}$ is the set of all tuples $(a_1,\dots,a_r)$ that are* ***not*** *weakly-good. The* weak r-ary chromatic number *of $\mathscr{H}$, denoted $\chi_r^{\mathrm{wk}}(\mathscr{H})$, is the largest $a_1 + \dots + a_r + 1$ such that $(a_1,\dots,a_r)$ is in the weak clique spectrum of $\mathscr{H}$.*
- *The* strong clique spectrum *of $\mathscr{H}$ is the set of all tuples $(a_1,\dots,a_r)$ that are* ***not*** *strongly-good. The* strong r-ary chromatic number *of $\mathscr{H}$, denoted $\chi_r^{\mathrm{st}}(\mathscr{H})$, is the largest $a_1 + \dots + a_r + 1$ such that $(a_1,\dots,a_r)$ is in the strong clique spectrum of $\mathscr{H}$.*

The f and g functions are defined similar to the graph case.

Definition 41 *Let $K = (\{u_1,\dots,u_k\},\phi)$ be an r-CRG and for $\mathbf{p} = (p_1,\dots,p_r)$, let $\mathbf{M}_K(\mathbf{p})$ denote the matrix with entries defined as follows:*

$$m_K(\mathbf{p})_{ij} = 1 - \sum_{\rho\in\phi(u_i,u_j)} p_\rho .$$

The functions f_K and g_K are defined as follows:

$$f_K(\mathbf{p}) = \frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(\mathbf{p})\mathbf{1} \tag{15}$$

$$g_K(\mathbf{p}) = \min\left\{\mathbf{x}^T\mathbf{M}_K(\mathbf{p})\mathbf{x} : \mathbf{x}^T\mathbf{1} = 1, \mathbf{x} \geq \mathbf{0}\right\}. \tag{16}$$

We say that a CRG K is $\mathbf{p}$-*core* if, for any proper sub-r-CRG K' of K, $g_{K'}(\mathbf{p}) > g_K(\mathbf{p})$.

We summarize the basic properties of this version of the edit distance function that generalize Theorems 13, 3 and 6. For $\mathbf{p} = (p_1, \ldots, p_r)$, the random r-colored-graph $G(n, \mathbf{p})$ is the complete graph on n vertices in which each edge independently receives color ρ with probability p_ρ.

Theorem 42 (Axenovich-Martin [10]) *Let $\mathscr{H}$ be a hereditary property of r-colored-graphs.*

(a) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is continuous over the* $(r-1)$*-simplex.*
(b) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is concave down over the* $(r-1)$*-simplex.*
(c) $\mathrm{ed}_{\mathscr{H}}(r^{-1}\mathbf{1}) \geq 1/(r(\chi_r^{\mathrm{st}}(\mathscr{H})-1))$.
(d) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) \leq 1/(\chi_r^{\mathrm{wk}}(\mathscr{H})-1)$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*
(e) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \lim_{n\to\infty} \mathbb{E}[\mathrm{dist}(G(n,\mathbf{p}),\mathscr{H})]$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*
(f) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \inf\{f_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\} = \inf\{g_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\}$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*

Finally, we give some examples of results in the case $r = 3$.

Theorem 43 (Axenovich-Martin [10]) *Let $r = 3$ and let $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ be a hereditary property of r-colored-graphs. Let $d^*_{\mathscr{H}} := \max\{\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) : \mathbf{p}^T\mathbf{1} = 1, \mathbf{p} \geq \mathbf{0}\}$.*

(a) If $\mathscr{F}$ is a family that consists of a single monochromatic triangle, then $d^*_{\mathscr{H}} = 1/2$.
(b) If $\mathscr{F}$ is a family that consists of a single triangle with two edges colored 1 and the other edge colored 2, then $d^*_{\mathscr{H}} = 1/2$.
(c) If $\mathscr{F}$ is a family that consists of two monochromatic triangles of different colors, then $d^*_{\mathscr{H}} = 1/2$.
(d) If $\mathscr{F}$ is a family that consists of all six bi-chromatic triangles, then $d^*_{\mathscr{H}} = 2/3$.
(e) If $\mathscr{F}$ is a family that consists of a single rainbow triangle, then $d^*_{\mathscr{H}} = 1/3$.

7.3 Directed edit distance

A *simple directed graph* or *digraph* G is a pair (V, c) such that V is a finite labeled set and, if $(V)_2 = V \times V - \{(v,v) : v \in V\}$ then $c : (V)_2 \to \{\bigcirc, -, \leftarrow, \rightarrow\}$ where

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

In the standard representation of digraphs as a pair (V,E) where $E \subseteq (V)_2$, we interpret $c(v,w) = \bigcirc$ to mean that neither (v,w) nor (w,v) is in E, $c(v,w) = -$ to mean that both (v,w) and (w,v) are in E, and $c(v,w) = \rightarrow$ to mean that $(v,w) \in E$ but $(w,v) \notin E$. We also define the following for any digraph G:

- $E_{\bigcirc}(G)$ 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 $c(v,w) = -$.

If $G = (V,c)$ and $G' = (V,c)$ are two digraphs on the same labeled vertex set with the same fixed palette, then the edit distance between them, $\mathrm{dist}(G,G')$ is the proportion of ordered pairs on which G and G' differ. We define $\mathrm{dist}(G,\mathscr{H})$ for any hereditary property of digraphs on any palette as in (1). A property is, of course, hereditary if it is closed under isomorphism and the deletion of vertices. For any digraph, we write $\mathrm{Forb}(H)$ to be the set of all digraphs that have no induced copy of H.

The digraph case encompasses several well-studied subclasses of digraphs. Just as the number of colors must be specified in Section 7.2, the palette must be specified for the digraph case.

Definition 44 *We say that $\mathscr{P} \subseteq \{\bigcirc, -, \leftarrow, \rightarrow\}$ is a* palette *if either none or both of "$\leftarrow$" and "$\rightarrow$" are in $\mathscr{P}$. There are 5 possible nontrivial palettes:*

(0) $\mathscr{P}_0 = \{\bigcirc, -, \leftarrow, \rightarrow\}$ is the general case.
(1) $\mathscr{P}_{\mathrm{compl}} = \{-, \leftarrow, \rightarrow\}$ is the case of simple digraphs such that every pair of vertices has at least one arc between them.
(2) $\mathscr{P}_{\mathrm{orien}} = \{\bigcirc, \leftarrow, \rightarrow\}$ is the case of oriented graphs.
(3) $\mathscr{P}_{\mathrm{undir}} = \{\bigcirc, -\}$ is the usual case of simple, undirected graphs.
(4) $\mathscr{P}_{\mathrm{tourn}} = \{\leftarrow, \rightarrow\}$ is the case of tournaments.

Definition 45 *A* directed density vector (p,q) *is a pair such that $p \geq 0$, $q \geq 0$ and $p+2q \leq 1$. For different palettes, there are further restrictions.*

(0) If $\mathscr{P} = \mathscr{P}_{\mathrm{compl}}$ then $p+2q = 1$.
(1) If $\mathscr{P} = \mathscr{P}_{\mathrm{orien}}$ then $p = 0$ and $q \leq 1/2$.
(2) If $\mathscr{P} = \mathscr{P}_{\mathrm{undir}}$ then $q = 0$ and $p \leq 1$; i.e., the usual graph case.
(3) If $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$ then $p = 0$ and $q = 1/2$.

If $\mathscr{H}$ is a hereditary property of digraphs with palette $\mathscr{P}$, then for all directed density vectors $\mathbf{p} = (p,q)$, we define the edit distance function for hereditary property $\mathscr{H}$ as follows:

$$\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) := \lim_{n\to\infty} \max\left\{\mathrm{dist}(G,\mathscr{H}) : \begin{array}{l} |V(G)| = n, |E_{-}(G)| = \lfloor p\binom{n}{2}\rfloor, \\ |E_{\leftarrow}(G)| = |E_{\rightarrow}(G)| = \lfloor q\binom{n}{2}\rfloor \end{array}\right\}.$$

The limit was proven to exist in [10].

In [10], the equivalent of CRGs (called *dir-types* in [10], but it would be natural to call them $\mathscr{P}$-dir-CRGs for palette $\mathscr{P}$) are defined as well as the notion of $H \mapsto K$ for any digraph H and any K a $\mathscr{P}$-dir-CRG. The matrix $\mathbf{M}_K(\mathbf{p})$ and functions $f_K(\mathbf{p})$ and $g_K(\mathbf{p})$ are defined analogously. In addition, the *strong directed clique spectrum*, *strong directed chromatic number* $\chi_{\mathscr{P}}^{\mathrm{st,dir}}(\mathscr{H})$, *weak directed clique spectrum* and *weak directed chromatic number* $\chi_{\mathscr{P}}^{\mathrm{wk,dir}}(\mathscr{H})$ are defined for each palette, although for $\mathscr{P}_{\mathrm{undir}}$ and $\mathscr{P}_{\mathrm{tourn}}$ "strong" and "weak" are the same, where we use the notation $\chi_{\mathscr{P}}^{\mathrm{dir}}(\mathscr{H})$.

We will not give the detailed definitions of these quantities or of the random digraph $G(n,\mathbf{p})$. The natural notions are defined precisely in [10]. We have similar basic results for the directed case as for the multicolored case in Theorem 42.

Theorem 46 (Axenovich-Martin [10]) *Let $\mathscr{P}$ be a palette and let $\mathscr{H}$ be a hereditary property of digraphs with palette $\mathscr{P}$. Let the domain be defined as in Definition 45.*

(a) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is continuous over the domain.*
(b) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is concave down over the domain.*
(c) $\mathrm{ed}_{\mathscr{H}}(r^{-1}\mathbf{1}) \geq 1/(r(\chi_{\mathscr{P}}^{\mathrm{st,dir}}(\mathscr{H})-1))$.
(d) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) \leq 1/(\chi_{\mathscr{P}}^{\mathrm{wk,dir}}(\mathscr{H})-1)$ *for all* $\mathbf{p}$ *in the domain.*
(e) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \lim_{n\to\infty}\mathbb{E}[\mathrm{dist}(G(n,\mathbf{p}),\mathscr{H})]$ *for all* $\mathbf{p}$ *in the domain.*
(f) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \inf\{f_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\} = \inf\{g_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\}$ *for all* $\mathbf{p}$ *in the domain.*

We give some examples involving triangles.

Theorem 47 (Axenovich-Martin [10]) *Let $\mathscr{H}$ be a hereditary property of digraphs with palette $\mathscr{P}$.*

(a) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{dir}})$ where H_{dir} is a directed triangle, then $d^*_{\mathscr{H}} = 1/2$ regardless of $\mathscr{P}$.
(b) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{tra}})$ where H_{tra} is a transitive triangle, then $\mathscr{H}$ is a trivial hereditary property as long as $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$.
(c) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{tra}})$ where H_{tra} is a transitive triangle, then $d^*_{\mathscr{H}} = 1/2$, as long as $\mathscr{P} \neq \mathscr{P}_{\mathrm{tourn}}$.
(d) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{dir}}) \cap \mathrm{Forb}(H_{\mathrm{tra}})$ where H_{dir} is a directed triangle and H_{tra} is a transitive triangle, then $d^*_{\mathscr{H}} = 1/2$, as long as $\mathscr{P} \neq \mathscr{P}_{\mathrm{tourn}}$.

The case of tournaments turns out to be trivial. Theorem 47(b) is a simple consequence of Ramsey theory, a hereditary property $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})}\mathrm{Forb}(H)$ is nontrivial if and only if no member of $\mathscr{F}(\mathscr{H})$ is transitive. In the case of tournaments, the density vector must be $\mathbf{p} = (0,1/2)$. The edit distance function is, therefore, a constant.

Theorem 48 *Let $\mathscr{H}$ be a nontrivial hereditary property of tournaments and $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$. Then*

$$\mathrm{ed}_{\mathscr{H}}(0,1/2) = \frac{1}{2(\chi_{\mathscr{P}}^{\mathrm{dir}}(\mathscr{H})-1)}.$$

8 Future directions

8.1 Hypergraph edit distance

Berikkyzy and the author have been investigating the extension of the edit distance problem to r-uniform hypergraphs (r-graphs). A *colored regularity hypergraph of order r (r-CRH)* is a triple (V,E,ϕ) in which V is a vertex set, E is the collection of all r-multisets on V and $\phi : E \to \{\mathbf{W},\mathbf{G},\mathbf{B}\}$ with the restrictions that, (a) for any $v \in V$, $\phi(\{v,\ldots,v\}) \in \{\mathbf{W},\mathbf{B}\}$ and (b) for any permutation $\sigma \in \Sigma_r$, $\phi(\{v_1,\ldots,v_r\}) = \phi(\{v_{\sigma(1)},\ldots,v_{\sigma(r)}\})$. Therefore, a 2-CRH is just a CRG.

In parallel to the graph case, we can define colored homomorphisms from r-uniform hypergraphs to r-CRHs so that if r-graph H does not map to a r-CRH K, then an r-graph G which is edited according to the "recipe" defined by K will have no induced copy of H.

We can then define, for each r-CRH K, an r-linear form which we can also call $g_K(p)$. It is easy to prove, for a hereditary property $\mathscr{H}$ of r-graphs, that there is a family $\mathscr{F}(\mathscr{H})$ of r-CRHs such that

$$\mathrm{ed}_{\mathscr{H}}(p) \leq \inf_{K\in\mathscr{K}(\mathscr{H})} g_K(p). \tag{17}$$

The difficulty in extending the theory of the edit distance in graphs to hypergraphs is in proving that (17) is, in fact, an equality.

The above definition of the r-CRH would not seem to be adequate to capture the subtleties of hypergraphs. Consider the common example of a hypergraph whose 3-edges are cyclic triangles in an underlying random tournament. See, e.g., the survey of hypergraph Turán theory by Keevash [30]. This hypergraph has no copy of the tetrahedron K_4^3 but crossing triples would be gray in any 3-CRH that models it.

Strong hypergraph regularity was developed in the 3-uniform case by Frankl and Rödl [26] and then for the general r-uniform case by Gowers [29], Rödl and Skokan [44, 45] and Nagle, Rödl and Schacht [36]. In these formulations, the notion of how overlapping hyperedges interact is captured by structures known as *complexes*. The structure of complexes inherent in strong hypergraph regularity would seem to be necessary in order to define r-CRHs and colored homomorphisms in order for the existence of a particular induced hypergraph to be determined.

The edit distance problem is, asymptotically, a general case of the Turán problem. In the context of Turán-type problems, a hypergraph property is *monotone* if it is closed under the taking of (not necessarily induced) subgraphs. Therefore, a monotone property is also hereditary. For a monotone property $\mathscr{M}$, the Turán density is $\pi(\mathscr{M}) = \limsup_{n\to\infty} \max\{|E(G)|/\binom{n}{2} : |V(G)| = n, G \in \mathscr{M}\}$. It is easy to see that

$$\pi(\mathscr{M}) = 1 - \mathrm{ed}_{\mathscr{M}}(1).$$

The Turán density for most monotone properties is not currently known, even though a great deal of work has been done on the subject.

In the graph case, it is trivial to derive $\mathrm{ed}_{\mathscr{M}}(1)$ using symmetrization. In addition, the classification of p-core CRGs established by Marchant and Thomason [31] allow for a trivial proof of the asymptotic Erdős-Stone-Simonovits result [25, 24].

In Question 4 from Section 8.2, we ask several questions that are related to a general theory of edit distance in r-uniform hypergraphs.

8.2 Open Problems

We first ask about powers of cycles and the questions left open in Section 5.4.

Question 1 *Let $\mathscr{H} = \mathrm{Forb}(C_h^t)$. What is $\mathrm{ed}_{\mathscr{H}}(p)$ for small values of p, where $t+1$ divides h? What is $\mathrm{ed}_{\mathscr{H}}(p)$ for small values of h? In particular:*

- *For $\mathscr{H} = \mathrm{Forb}(C_h)$, what is $\mathrm{ed}_{\mathscr{H}}(p)$ for even values of h and all values of p?*
- *For $\mathscr{H} = \mathrm{Forb}(C_h)$, what is $d^*_{\mathscr{H}}$ for all $t \geq 2$ and $h \geq 2t+1$?*

Next we consider complete bipartite graphs and some interesting questions from Section 5.5

Question 2 *What is $\mathrm{ed}_{\mathscr{H}}(p)$ for $\mathscr{H} = \mathrm{Forb}(K_{s,t})$? In particular:*

- *For $\mathscr{H} = \mathrm{Forb}(K_{s,s})$, is $d^*_{\mathscr{H}} = 1/(2s-2)$ if $s \geq 7$?*
- *For $\mathscr{H} = \mathrm{Forb}(K_{s,s})$, is $p^*_{\mathscr{H}}$ is an interval of positive length if $s \geq 7$?*
- *For $\mathscr{H} = \mathrm{Forb}(K_{s,t})$, which values of s and t give that $p^*_{\mathscr{H}}$ is an interval of positive length?*

Other metrics on the space of graphs are of interest, as we discussed in Section 6.

Question 3 *Let $\mathscr{H}$ be a nontrivial hereditary property of graphs.*

- *For the cut metric $\mathrm{d}_{\square}$, is it the case that the function $\phi_{\mathscr{H}}$, as defined in (13), can be expressed only in terms of p and of $\mathrm{ed}_{\mathscr{H}}(\cdot)$?*
- *For any metric d that is continuous with respect to the edit metric, is it the case that the function $\phi_{\mathscr{H}}$, as defined in (13), can be expressed only in terms of p and of $\mathrm{ed}_{\mathscr{H}}(\cdot)$?*

The question of the edit distance in hypergraphs is wide open, as we discussed in Section 8.1.

Question 4 *Let $\mathscr{H}$ be a nontrivial hereditary property of r-uniform hypergraphs.*

- *Is it the case that $\mathrm{ed}_{\mathscr{H}}(p) = \inf_{K\in\mathscr{K}(\mathscr{H})} g_K(p)$ for all $p \in [0,1]$?*
- *If $\mathscr{H}$ is monotone, is it the case that $\mathrm{ed}_{\mathscr{H}}(1) = \inf_{K\in\mathscr{K}(\mathscr{H})} g_K(1)$?*
- *Is there a useful form of generalizations properties do r-linear forms have?*
- *Can we provide a structural characterization for r-CRHs that are p-core, à la Theorem 9?*

The cases of $\mathscr{H} = \mathrm{Forb}(K_{3,3})$ and $\mathscr{H} = \mathrm{Forb}(K_{2,t})$ for $t \geq 9$ suggest an infinite number of CRGs are necessary to define an edit distance function, but only if one wants to compute it for p arbitrarily close to 0 (or by considering the property $\overline{\mathscr{H}}$, arbitrarily close to 1).

Conjecture 1 *Let $\mathscr{H}$ be a nontrivial hereditary property. For every $\varepsilon > 0$ there exists a $\mathscr{K}' = \mathscr{K}'(\varepsilon, \mathscr{H})$ such that*

$$\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{g_K(p) : K \in \mathscr{K}'\right\} \qquad \textit{for all } p \in (\varepsilon, 1-\varepsilon).$$

We ask if the behavior we seem to observe for $\mathrm{Forb}(K_{3,3})$ and $\mathrm{Forb}(K_{2,t})$ for $t \geq 9$ – that is, that an infinite sequence of CRGs are required to compute the edit distance function for all values of p – does, in fact, occur.

Question 5 *Are there hereditary properties of graphs for which the edit distance function cannot be determined from the G functions of a finite number of CRGs?*

Finally, we conclude with an open problem for the random graph. Recall that $G(n, p)$ denotes the Erdős-Rényi random graph on n vertices with probability p.

Conjecture 2 (Martin [33]) *Fix $p_0 \in (0,1)$ and let $\mathscr{H} = \mathrm{Forb}(G(n_0, p_0))$. Then*

$$\mathrm{ed}_{\mathscr{H}}(p) = (1+o(1))\frac{2\log_2 n_0}{n_0}\min\left\{\frac{p}{-\log_2(1-p_0)}, \frac{1-p}{-\log_2 p_0}\right\}$$

with probability approaching 1 as $n_0 \to \infty$.

The functions that define this bound are of the form $p/(\chi - 1)$ and $(1-p)/(\overline{\chi} - 1)$. The case of $p_0 = 1/2$ was proved to be true by Alon and Stav [6].

If Conjecture 2 is true, then it implies that $p^*_{\mathscr{H}} = \frac{\log_2(1-p_0)}{\log_2 p_0(1-p_0)}$, which is only equal to p_0 itself when $p_0 \in \{0, 1/2, 1\}$. Informally, this implies the counterintuitive notion that it is harder to remove induced copies of $G(n_0, p_0)$ from $G(n, p^*_{\mathscr{H}})$ than it is to remove them from $G(n, p_0)$.

If Conjecture 2 is false, then it implies that the structure of random graphs and the behavior of editing induced graphs is quite complex and very unexpected.

Acknowledgments

The author would like to extend a heartfelt appreciation many people for valuable conversations on the topic of edit distance, including Noga Alon, Ed Marchant, Richard Mycroft, Jozef Skokan and Andrew Thomason. Thanks also to his collaborators József Balogh, Zhanar Berikkyzy, André Kézdy, Tracy McKay and Chelsea Peck. Special appreciation to Maria Axenovich, who suggested the problem in the first place and without whom this work would not be possible.

The author would like to profoundly thank the National Security Agency (NSA) which partially supported his research via grant H98230-13-1-0226 as well as the

Institute for Mathematics and its Applications (IMA). This paper was written while the author was a long-term visitor at the IMA and would not have been possible without their resources and generous support.

References

1. V. E. Alekseev, *Hereditary classes and coding of graphs*, Problemy Kibernet. (1982), no. 39, 151–164. MR 694829 (85a:05071)
2. ———, *Range of values of entropy of hereditary classes of graphs*, Diskret. Mat. **4** (1992), no. 2, 148–157. MR 1181539 (93k:05140)
3. Noga Alon, *Testing subgraphs in large graphs*, Random Structures Algorithms **21** (2002), no. 3-4, 359–370, Random structures and algorithms (Poznan, 2001). MR 1945375 (2003k:05129)
4. Noga Alon, Richard A. Duke, Hanno Lefmann, Vojtěch Rödl, and Raphael Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms **16** (1994), no. 1, 80–109. MR 1251840 (94j:05112)
5. Noga Alon, Eldar Fischer, Michael Krivelevich, and Mario Szegedy, *Efficient testing of large graphs*, Combinatorica **20** (2000), no. 4, 451–476. MR 1804820 (2002f:05143)
6. Noga Alon and Uri Stav, *The maximum edit distance from hereditary graph properties*, J. Combin. Theory Ser. B **98** (2008), no. 4, 672–697. MR 2418765 (2009d:05111)
7. ———, *What is the furthest graph from a hereditary property?*, Random Structures Algorithms **33** (2008), no. 1, 87–104. MR 2428979 (2009c:05219)
8. Maria Axenovich, André Kézdy, and Ryan R. Martin, *On the editing distance of graphs*, J. Graph Theory **58** (2008), no. 2, 123–138. MR 2407000 (2009c:05107)
9. Maria Axenovich and Ryan R. Martin, *Avoiding patterns in matrices via a small number of changes*, SIAM J. Discrete Math. **20** (2006), no. 1, 49–54 (electronic). MR 2257244 (2007j:05015)
10. ———, *Multicolor and directed edit distance*, J. Comb. **2** (2011), no. 4, 525–556. MR 2911190
11. József Balogh, Béla Bollobás, Michael Saks, and Vera T. Sós, *The unlabelled speed of a hereditary graph property*, J. Combin. Theory Ser. B **99** (2009), no. 1, 9–19. MR 2467814 (2009k:05105)
12. József Balogh, Béla Bollobás, and Miklós Simonovits, *The fine structure of octahedron-free graphs*, J. Combin. Theory Ser. B **101** (2011), no. 2, 67–84. MR 2763070 (2012a:05157)
13. József Balogh, Béla Bollobás, and David Weinreich, *The speed of hereditary properties of graphs*, J. Combin. Theory Ser. B **79** (2000), no. 2, 131–156. MR 1769217 (2001d:05092)
14. ———, *The penultimate rate of growth for graph properties*, European J. Combin. **22** (2001), no. 3, 277–289. MR 1822715 (2002b:05079)
15. Jòzsef Balogh, Béla Bollobás, and David Weinreich, *Measures on monotone properties of graphs*, Discrete Appl. Math. **116** (2002), no. 1-2, 17–36. MR 1877113 (2002i:05066)
16. József Balogh and Ryan R. Martin, *Edit distance and its computation*, Electron. J. Combin. **15** (2008), no. 1, Research Paper 20, 27. MR 2383440 (2008j:05175)
17. Zhanar Berikkyzy, Ryan R. Martin, and Chelsea Peck, *On the edit distance of powers of cycles*, In progress.
18. Béla Bollobás and Andrew Thomason, *Projections of bodies and hereditary properties of hypergraphs*, Bull. London Math. Soc. **27** (1995), no. 5, 417–424. MR 1338683 (96e:52006)
19. ———, *Hereditary and monotone properties of graphs*, The mathematics of Paul Erdős, II, Algorithms Combin., vol. 14, Springer, Berlin, 1997, pp. 70–78. MR 1425205 (98a:05085)
20. ———, *The structure of hereditary properties and colourings of random graphs*, Combinatorica **20** (2000), no. 2, 173–202. MR 1767020 (2001e:05114)

21. C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, *Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing*, Adv. Math. **219** (2008), no. 6, 1801–1851. MR 2455626 (2009m:05161)
22. William G. Brown, *On graphs that do not contain a Thomsen graph*, Canad. Math. Bull. **9** (1966), 281–285. MR 0200182 (34 #81)
23. Duhong Chen, Oliver Eulenstein, David Fernández-Baca, and Michael Sanderson, *Supertrees by flipping*, Computing and combinatorics, Lecture Notes in Comput. Sci., vol. 2387, Springer, Berlin, 2002, pp. 391–400. MR 2064534 (2005b:92030)
24. P. Erdős and M. Simonovits, *A limit theorem in graph theory*, Studia Sci. Math. Hungar **1** (1966), 51–57. MR 0205876 (34 #5702)
25. P. Erdös and A. H. Stone, *On the structure of linear graphs*, Bull. Amer. Math. Soc. **52** (1946), 1087–1091. MR 0018807 (8,333b)
26. Peter Frankl and Vojtěch Rödl, *Extremal problems on set systems*, Random Structures Algorithms **20** (2002), no. 2, 131–164. MR 1884430 (2002m:05192)
27. Alan Frieze and Ravi Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), no. 2, 175–220. MR 1723039 (2001i:68066)
28. Zoltán Füredi, *New asymptotics for bipartite Turán numbers*, J. Combin. Theory Ser. A **75** (1996), no. 1, 141–144. MR 1395763 (97f:05096)
29. W. T. Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. (2) **166** (2007), no. 3, 897–946. MR 2373376 (2009d:05250)
30. Peter Keevash, *Hypergraph Turán problems*, Surveys in combinatorics 2011, London Math. Soc. Lecture Note Ser., vol. 392, Cambridge Univ. Press, Cambridge, 2011, pp. 83–139. MR 2866732
31. Edward Marchant and Andrew Thomason, *Extremal graphs and multigraphs with two weighted colours*, Fete of combinatorics and computer science, Bolyai Soc. Math. Stud., vol. 20, János Bolyai Math. Soc., Budapest, 2010, pp. 239–286. MR 2797967 (2012f:05149)
32. Edward James Marchant, *Graphs with weighted colours and hypergraphs*, Ph.D. thesis, University of Cambridge, 2011.
33. Ryan Martin, *The edit distance function and symmetrization*, Electron. J. Combin. **20** (2013), no. 3, Paper 26, 25. MR 3104524
34. Ryan R. Martin, *On the computation of edit distance functions*, Discrete Math. **338** (2015), no. 2, 291–305. MR 3279281
35. Ryan R. Martin and Tracy McKay, *On the edit distance from $K_{2,t}$-free graphs*, J. Graph Theory **77** (2014), no. 2, 117–143. MR 3246171
36. Brendan Nagle, Vojtěch Rödl, and Mathias Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms **28** (2006), no. 2, 113–179. MR 2198495 (2007d:05084)
37. Chelsea Peck, *On the edit distance from a cycle- and squared cycle-free graph*, Master's thesis, Iowa State University, United States, 2013.
38. Oleg Pikhurko, *An exact Turán result for the generalized triangle*, Combinatorica **28** (2008), no. 2, 187–208. MR 2399018 (2009c:05161)
39. Lajos Pósa, *Hamiltonian circuits in random graphs*, Discrete Math. **14** (1976), no. 4, 359–364. MR 0389666 (52 #10497)
40. Hans Jürgen Prömel and Angelika Steger, *Excluding induced subgraphs: quadrilaterals*, Random Structures Algorithms **2** (1991), no. 1, 55–71. MR 1099580 (92j:05157)
41. ______, *Excluding induced subgraphs. III. A general asymptotic*, Random Structures Algorithms **3** (1992), no. 1, 19–31. MR 1139486 (93d:05065)
42. ______, *Excluding induced subgraphs. II. Extremal graphs*, Discrete Appl. Math. **44** (1993), no. 1-3, 283–294. MR 1227710 (94f:05121)
43. Duncan C. Richer, *Graph theory and combinatorial games*, Ph.D. thesis, University of Cambridge, 2000.
44. Vojtěch Rödl and Jozef Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms **25** (2004), no. 1, 1–42. MR 2069663 (2005d:05144)
45. ______, *Applications of the regularity lemma for uniform hypergraphs*, Random Structures Algorithms **28** (2006), no. 2, 180–194. MR 2198496 (2006j:05099)

46. Edward R. Scheinerman and Jennifer Zito, *On the size of hereditary classes of graphs*, J. Combin. Theory Ser. B **61** (1994), no. 1, 16–39. MR 1275261 (95e:05061)
47. Alexander Sidorenko, *Boundedness of optimal matrices in extremal multigraph and digraph problems*, Combinatorica **13** (1993), no. 1, 109–120. MR 1221180 (94c:05043)
48. Endre Szemerédi, *Regular partitions of graphs*, Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS, vol. 260, CNRS, Paris, 1978, pp. 399–401. MR 540024 (81i:05095)
49. Andrew Thomason, *Graphs, colours, weights and hereditary properties*, Surveys in combinatorics 2011, London Math. Soc. Lecture Note Ser., vol. 392, Cambridge Univ. Press, Cambridge, 2011, pp. 333–364. MR 2866736
50. A. A. Zykov, *On some properties of linear complexes*, Amer. Math. Soc. Translation **1952** (1952), no. 79, 33. MR 0051516 (14,493a)