$2 \times n$ grids have unbounded anagram-free chromatic number*

Saman Bazargani
Department of Computer Science and Electrical Engineering
University of Ottawa
Ontario, Canada
saman.bazargani@gmail.com

Paz Carmi
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
carmip@cs.bgu.ac.il

Vida Dujmović
Department of Computer Science and Electrical Engineering
University of Ottawa
Ontario, Canada
vida.dujmovic@uottawa.ca

Pat Morin
School of Computer Science
Carleton University
Ottawa, Canada
morin@scs.carleton.ca

Submitted: May 5, 2021; Accepted: Apr 27, 2022; Published: Aug 26, 2022

Abstract

We show that anagram-free vertex colouring a $2 \times n$ square grid requires a number of colours that increases with n. This answers an open question in Wilson's thesis and shows that there are even graphs of pathwidth 2 that do not have anagram-free colourings with a bounded number of colours.

Mathematics Subject Classifications: 05C15

*This research was partly funded by NSERC.

1 Introduction

Two words s and t are *anagrams* of each other if s is a permutation of t. A single word $w := w_1, \ldots, w_{2r}$ is *anagramish* if its first half $w_1, \ldots, w_r$ and its second half $w_{r+1}, \ldots, w_{2r}$ are anagrams of each other. A word is *anagram-free* if it does not contain an anagramish factor. Anagramish words are also known as *abelian squares* and anagram-free words are also known as *abelian square free* words [7, 8], *strongly nonrepetitive* words [11], or *strongly asymmetric* sequences [4, 5].

In 1961, Erdős [3] asked if there exist arbitrarily long anagram-free words over an alphabet of size 4.[1] In 1968 Evdokimov [4, 5] showed the existence of arbitrarily long anagram-free words over an alphabet of size 25 and in 1971 Pleasants [11] showed that an alphabet of size 5 is sufficient. Erdős's question was not fully resolved until 1992, when Keränen [7] answered it in the affirmative.

A path $v_1, \ldots, v_{2r}$ in a graph G is *anagramish* under a vertex c-colouring $\phi : V(G) \to \{1, \ldots, c\}$ if $\phi(v_1), \ldots, \phi(v_{2r})$ is an anagramish word. The colouring ϕ is an *anagram-free colouring* of G if no path in G is anagramish under ϕ. The minimum integer c for which G has an anagram-free vertex c-colouring is called the *anagram-free chromatic number* of G and is denoted by $\dot{\chi}_\pi(G)$.

For a non-empty graph family $\mathcal{G}$, $\dot{\chi}_\pi(\mathcal{G}) := \max\{\dot{\chi}_\pi(G) : G \in \mathcal{G}\}$ or $\dot{\chi}_\pi(\mathcal{G}) := \infty$ if the maximum is undefined. The results on anagram-free words discussed in the preceding paragraph can be interpreted in terms of $\dot{\chi}_\pi(\mathcal{P})$ where $\mathcal{P}$ is the family of all paths. Indeed, Keränen's result shows that $\dot{\chi}_\pi(\mathcal{P}) \leqslant 4$. Slightly more complicated than paths are trees. Wilson and Wood [14] showed that $\dot{\chi}_\pi(\mathcal{T}) = \infty$ for the family $\mathcal{T}$ of trees and Kamčev, Łuczak, and Sudakov [6] showed that $\dot{\chi}_\pi(\mathcal{T}_2) = \infty$ even for the family $\mathcal{T}_2$ of binary trees.

One positive result in this context is that of Wilson and Wood [14], who showed that every tree T of pathwidth p has $\dot{\chi}_\pi(T) \leqslant 4p + 1$. Since trees are graphs of treewidth 1 it is natural to ask if this result can be extended to show that every graph G of treewidth t and pathwidth p has $\dot{\chi}_\pi(G) \leqslant f(t, p)$ for some function $f : \mathbb{N}^2 \to \mathbb{N}$. Carmi, Dujmović, and Morin [2] showed that such a generalization is not possible for any $t \geqslant 3$ by giving examples of n-vertex graphs of pathwidth 3 (and treewidth 3) with $\dot{\chi}_\pi(G) \in \Omega(\log n)$. The obvious remaining gap left by these two works is graphs of treewidth 2. Our main result is to show that G_n, the $2 \times n$ square grid has $\dot{\chi}_\pi(G_n) \in \omega_n(1)$. Since G_n has pathwidth 2, we have:

Theorem 1. For every $c \in \mathbb{N}$, there exists a graph of pathwidth 2 that has no anagram-free vertex c-colouring.

Wilson [12, Section 7.1] conjectured that $\dot{\chi}_\pi(G_n) \in \omega_n(1)$, so this work confirms this conjecture. Prior to the current work, it was not even known if the family of $n \times n$ square grids had anagram-free colourings using a bounded number of colours.

[1] This was an incredibly prescient question since it is not at all obvious that there exist arbitrarily long anagram-free words over *any* finite alphabet. The only justification for choosing the constant 4 is that a short case analysis rules out the possibility of length-8 anagram-free words over an alphabet of size 3.

In a larger context, this lower bound gives more evidence that, except for a few special cases (paths [4, 7, 11], trees of bounded pathwidth [14], and highly subdivided graphs [13]), the qualitative behaviour of anagram-free chromatic number is not much different than that of treedepth/centered colouring [10]. Very roughly: For most graph classes, every graph in the class has an anagram-free colouring using a bounded number of colors precisely when every graph in the class has a colouring using a bounded number of colours in which every path contains a colour that appears only once in the path.

The remainder of this paper is organized as follows: Section 2 gives some definitions and states a key lemma which shows, under a certain periodicity condition, that every sufficiently long word contains a factor that is ϵ-close to being anagramish. In Section 3 we use this key lemma to prove Theorem 1. In Section 4 we prove the key lemma. Section 5 concludes with some final remarks about the (non-)constructiveness of our proof technique.

2 Periodicity in words

An *alphabet* Σ is a finite non-empty set and each element of Σ is called a *letter*. A *word* over Σ is a (possibly empty) sequence $w := w_1, \ldots, w_n$ with $w_i \in \Sigma$ for each $i \in \{1, \ldots, n\}$. The *length* $|w|$ of w is the length, n, of the sequence. A word of length at least 1 is *non-empty*. The unique word of length 0 is denoted by ε. For a word $w := w_1, \ldots, w_n$, we let $\Sigma_w := \{w_i : i \in \{1, \ldots, n\}\}$. For each $1 \leqslant i \leqslant j \leqslant n+1$, $w_i, w_{i+1}, \ldots, w_{j-1}$ is called a *factor* of w.[2] The set of all factors of w is denoted by $F(w)$.

A *language* L over Σ is a set of words over Σ. A language L is *infinite* if $|L| = \infty$. We let $\Sigma_L := \bigcup_{w \in L} \Sigma_w$ and $F(L) := \bigcup_{w \in L} F(w)$. The language L is *factor-closed* if $F(w) \subseteq L$ for each $w \in L$. For any alphabet Σ and any $k \in \mathbb{N}$, Σ^k (the k-fold cartesian product of Σ with itself) is the language consisting of all length-k words over Σ. The *Kleene closure* $\Sigma^* := \bigcup_{k=0}^{\infty} \Sigma^k$ is the language consisting of all words over Σ.

Let $w := w_1, \ldots, w_n$ be a word over an alphabet Σ and, for each $a \in \Sigma$, define $|w|_a := |\{i \in \{1, \ldots, n\} : w_i = a\}|$. The *Parikh vector* of w is the integer-valued $|\Sigma|$-vector $\mathrm{p}(w) := (|w|_a : a \in \Sigma)$ that is indexed by elements of Σ. Observe that a word $w_1, \ldots, w_{2r}$ is anagramish if and only if $\mathrm{p}(w_1, \ldots, w_r) = \mathrm{p}(w_{r+1}, \ldots, w_{2r})$ or, equivalently, $\mathrm{p}(w_1, \ldots, w_r) - \mathrm{p}(w_{r+1}, \ldots, w_{2r}) = \mathbf{0}$. For each $a \in \Sigma$, let $\delta_a(w) := |w_1, \ldots, w_r|_a - |w_{r+1}, \ldots, w_{2r}|_a$ and let $\tau_a(w) := |\delta_a|$. Then $\tau(w) := \sum_{a \in \Sigma} \tau_a(w)$ is a useful measure of how far a word is from being anagramish and $\tau(w) = 0$ if and only if w is anagramish.

A word $w := w_1, \ldots, w_n$ is *ℓ-periodic* over an alphabet $\Sigma \supseteq \Sigma_w$ if each length-ℓ factor of w contains every letter in Σ. A language L is *ℓ-periodic* if each of its words is ℓ-periodic over Σ_L. A language L is *periodic* if it is ℓ-periodic for some finite ℓ. We make use of the following key lemma, which states that every sufficiently long ℓ-periodic word contains arbitrarily long factors that are ϵ-close to being anagramish.

[2]Note that this definition includes the empty factor ε obtained when $i = j$.

Lemma 2. For each $r_0, \ell \in \mathbb{N}$ and each $\epsilon > 0$, there exists a positive integer n such that every ℓ-periodic word $w_1, \ldots, w_n$ contains a factor $w := w_{i+1}, \ldots, w_{i+2r}$ of length $2r \geqslant 2r_0$ such that $\tau(w) \leqslant \epsilon r$.

The proof of Lemma 2 is deferred to Section 4. We now give some intuition as to how it is used. The process of checking if a word is anagramish can be viewed as finding common letters in the first and second halves and crossing them both out. If this results in a complete cancellation of all letters, then the word is an anagram. Lemma 2 tells us that we can always find a long factor w where, after exhaustive cancellation, only an ϵ-fraction of the original letters remain.

Stated differently, if up to ϵr letters in each half of w were each allowed to cancel two of the same letters in the other half of w, then it would be possible to complete the cancellation process. To achieve this type of one-versus-two cancellation in our setting, we decompose our coloured pathwidth-2 graph into pieces of constant size. The vertices in each piece can be covered with one path or partitioned into two paths (see Figure 4). In this way an occurrence H_z of a particular coloured piece in one half can be matched with two like-coloured pieces H_x and H_y in the other half. We construct a single path P that contains all vertices in H_z and only half the vertices in each of H_x and H_y. In this way, the colours of vertices $P \cap H_z$ can cancel the colours of the vertices in $P \cap (H_x \cup H_y)$.

Since Lemma 2 requires that the word w be ℓ-periodic, the following lemma will be helpful in obtaining words that can be used with Lemma 2. (Although Lemma 3 follows immediately from considerably stronger results in combinatorics on words—for example Lothaire [9, Proposition 1.5.12]—we provide a proof here for the sake of completeness.)

Lemma 3. Let L be an infinite factor-closed language over an alphabet Σ. Then there exists an infinite periodic factor-closed language $M \subseteq L$.

Proof. Let $\Xi \subseteq \Sigma$ be any minimal subset of Σ with the property that $M := \Xi^* \cap L$ is infinite. Observe that, since L is factor closed, M is also factor-closed. Therefore M is infinite and factor-closed. All that remains is to show that M is ℓ-periodic for some integer ℓ.

Since Ξ is minimal, $\Pi^* \cap L$ is finite for any strict subset $\Pi \subsetneq \Xi$. Since L is factor-closed, $\Pi^* \cap L$ is factor-closed. Since $\Pi^* \cap L$ is factor-closed and finite, $|w|$ is finite for each $w \in \Pi^* \cap L$. Therefore, the value

$$\ell := 1 + \max\{|w| : w \in \Pi^* \cap L, \Pi \subsetneq \Xi\} \tag{1}$$

is finite. Now, consider any factor f of any word in M that does not contain every letter of Ξ, i.e., $f \in \Pi^*$ for some $\Pi \subsetneq \Xi$. Since $\Pi^* \cap L$ is factor closed, $f \in \Pi^* \cap L$ so, by Equation (1), $|f| < \ell$. Contrapositively, any factor of length at least ℓ of any word in M contains every element of Ξ, so M is ℓ-periodic. □

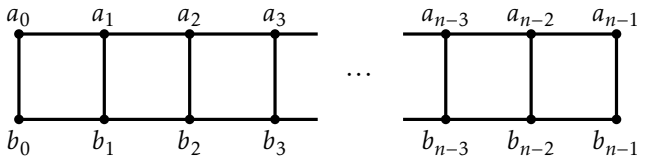

Figure 1: The graph G_n

3 Proof of Theorem 1

For each $n \in \mathbb{N}$, let G_n be the $2 \times n$ square grid with top row $a_0, \ldots, a_{n-1}$ and bottom row $b_0, \ldots, b_{n-1}$ (see Figure 1). Our proof strategy is a proof by contradiction. Ultimately, we will assume (for the sake of contradiction) that there exists an integer c such that $\dot{\chi}_\pi(G_n) \leqslant c$ for each $n \in \mathbb{N}$ and use this to derive a contradiction. In the following, we will define several languages L_1, L_2, and L_3 that would be infinite if the preceding assumption were true. The final contradiction will occur if L_3 contains a sufficiently long string.

As a first step, observe that for any $n \in \mathbb{N}$ there is a bijection between vertex c-colourings of G_n and words of length $2n$ over the alphabet $\Sigma_0 := \{1, \ldots, c\}$ where a word $w := w_0, \ldots, w_{2n-1}$ corresponds to the colouring $\phi_w : V(G_n) \to \{1, \ldots, c\}$ defined by $\phi_w(a_i) := w_{2i}$ and $\phi_w(b_i) := w_{2i+1}$, for each $i \in \{0, \ldots, n-1\}$.

For reasons that will become apparent, it is useful to break G_n into blocks consisting of 4 columns each. Let $\Sigma_1 := \Sigma_0^8$ and define the language L_1 to consist of exactly those words $w_1, \ldots, w_r \in \Sigma_1^*$ for which $\phi_{w_1 \cdots w_r}$ is an anagram-free colouring of G_{4r}.

From the definition of anagram-free colouring, it follows that L_1 is factor-closed. Define the lange $L_2 \subseteq L_1$ as follows: If L_1 is finite, then $L_2 := L_1$. If L_1 is infinite then, by Lemma 3, there exists an infinite periodic factor-closed language contained in L_1 and we let L_2 be any such language. Note that in either case L_2 is periodic, since every finite language is periodic.

Let x be any letter in Σ_{L_2}. Note that the word xx is not in L_2 since the corresponding colouring ϕ_{xx} of G_8 has, for example, the anagramish path $a_0, \ldots, a_7$. Since L_2 is factor-closed, no word in L_2 contains xx as a factor, i.e., $xx \notin F(L_2)$. Let $\Sigma_3 \subset F(L_2)$ contain only those factors of L_2 that do not contain x. Thus, any word in L_2 is of the form $w_0 x w_1 x w_2 \cdots x w_r x w_{r+1}$ where each of $w_1, \ldots, w_r$ is in Σ_3 and each of w_0, w_{r+1} is in $\Sigma_3 \cup \{\varepsilon\}$.

Since L_2 is periodic, each word in Σ_3 has bounded length, so Σ_3 is finite, i.e., Σ_3 is an alphabet. Let L_3 be the language consisting of all words $w_1, \ldots, w_r \in \Sigma_3^*$ such that $x w_1 x w_2 \cdots x w_r \in L_2$. If L_2 is infinite, then so is L_3. Since L_2 is factor closed and periodic, L_3 is also factor-closed and periodic.

See Figure 2 for what follows. Each word $w := w_1, \ldots, w_r \in L_3$ defines a word $w' := x w_1 x w_2 \cdots x w_r \in L_2$. In graph colouring terms, w defines an integer $n(w) := 4 \sum_{i=1}^{r} (1 + |w_i|)$ and an anagram-free colouring $\phi_w := \phi_{w'}$ of $G_{n(w)}$. The word w' partitions

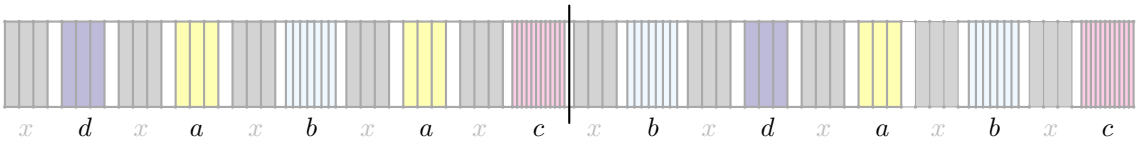

Figure 2: A word $w := dabacbdabc$, its corresponding word $w' := xdxaxbxaxcxbxdxaxbxc$, and its corresponding coloured graph $G_{n(w)}$.

the columns of $G_{n(w)}$ into *blocks*, each of which corresponds to a letter a of w' of length $4|a|$. There are two types of blocks:

- There are *boring* blocks $Q_0, \ldots, Q_{r-1}$, each containing 4 columns of $G_{n(w)}$, and each corresponding to an occurrence of x in w'.
- There are *colourful blocks* $H_1, \ldots, H_r$, where H_i is the subgraph of $G_{n(w)}$ corresponding to w_i and contains $4|w_i|$ columns of $G_{n(w)}$.

The following lemma shows that any sufficiently long string w in L_3 must define a colouring ϕ_w of $G_{n(w)}$ that is not anagram-free.

Lemma 4. There exists $\epsilon > 0$ and $r_0 \in \mathbb{N}$ such that, for any $w \in L_3$ that has even length $2r := |w| \geqslant 2r_0$ and that has $\tau(w) \leqslant \epsilon r$, the graph $G_{n(w)}$ contains a path that is anagramish under ϕ_w.

Proof. Let $w_1, \ldots, w_{2r} := w$ so that the colouring ϕ_w is defined by the word $xw_1xw_2 \cdots xw_{2r}$. For each $a \in \Sigma_3 \setminus \{x\}$, define $X_a^+ := \{i \in \{1, \ldots, r-1\} : w_i = a\}$ and $Y_a^+ := \{i \in \{r+1, \ldots, 2r-1\} : w_i = a\}$. (The sets X_a^+ and Y_a^+ index the occurrences of a in the first and second half of w, respectively.) We will now define two sets $X_a \subseteq X_a^+$ and $Y_a \subseteq Y_a^+$ with the following properties:

1. If $\delta_a(w) = 0$ then $X_a = Y_a = \emptyset$.
2. If $\delta_a(w) > 0$ then $|X_a| = 2\delta_a(w) = 2|Y_a|$.
3. If $\delta_a(w) < 0$ then $|X_a| = \delta_a(w) = \frac{1}{2}|Y_a|$.

Let $X := \bigcup_{a \in \Sigma_3} X_a$, let $Y := \bigcup_{a \in \Sigma_3} Y_a$, and observe that the three preceding properties imply that $|X| = |Y| = 3\tau(w)/2$. The sets X_a and Y_a will be chosen so that $X \cup Y$ satisfies the following *global independence constraint*: There is no pair $i, j \in X \cup Y$ such that $i - j = 1$. We now explain why, with appropriately chosen ϵ and r_0, it is always possible to find sets X_a and Y_a (for each $a \in \Sigma_3 \setminus \{x\}$) that satisfy these three properties.

We choose the elements of $X \cup Y$ using the following greedy strategy. Suppose X_a or Y_a does not yet contain enough elements, for some $a \in \Sigma_3$. Without loss of generality, assume it is X_a. Choose an arbitrary element from $X_a^+ \setminus \bigcup_{i \in X \cup Y} \{i-1, i, i+1\}$ and add it to X_a. To see that this is always possible observe that at the beginning of each step,

$$\begin{aligned} \left| X_a^+ \setminus \bigcup_{i \in X \cup Y} \{i-1, i, i+1\} \right| &\geqslant |X_a^+| - 3|X \cup Y| \\ &\geqslant |X_a^+| - 3\left(3\tau(w) - 1\right) \end{aligned}$$

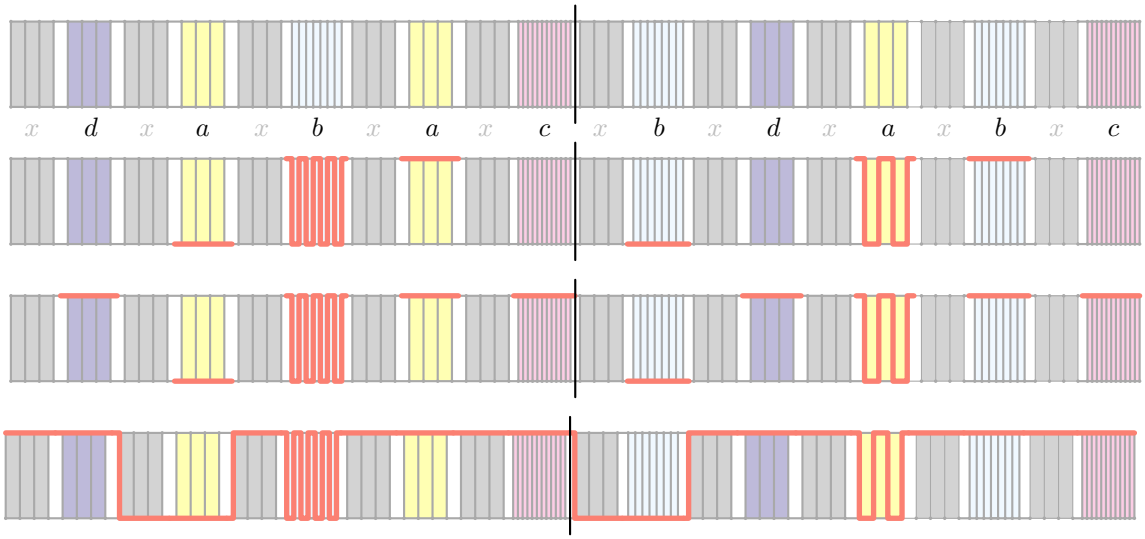

Figure 3: Constructing the anagramish path P: (1) pairs of top and bottom paths are matched with zig-zag paths; (2) all remaining colourful blocks receive top paths; (3) all boring blocks receive top, updown, or downup paths.

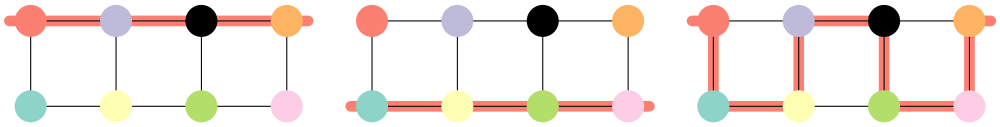

Figure 4: Subpaths of P through colourful blocks: A pair of top and bottom paths contribute the same amount as a single zig-zag path.

$$\geqslant |X_a^+| - 9\epsilon r - 3 \ ,$$

so there will always be an element to choose provided that $|X_a^+| > 9\epsilon r + 3$. Since $w \in L$, w is ℓ-periodic, so $|X_a^+| \geqslant \lfloor (r-1)/\ell \rfloor$. Therefore, this procedure will succeed in finding sets X and Y with the global independence property provided that $9\epsilon r - 3 < \lfloor (r-1)/\ell \rfloor$. In particular, $\epsilon < 1/(9\ell)$ satisfies this requirement.

To simplify notation, let $G := G_{n(w)}$. For each $i \in \{1, \ldots, 2r\}$, let H_i be the subgraph of G induced by the vertices corresponding to w_i; the subgraphs $H_1, \ldots, H_{2r}$ are referred to above as colourful blocks. We now construct the anagramish path P in a piecewise fashion, as illustrated in Figure 3.

1. For each $a \in \Sigma_3$ such that $\delta_a(w) > 0$, group the elements of X_a into pairs. For each pair (i, j), P contains the path through the top row of H_i and the path through the bottom row of H_j. For each element $i \in Y_a$ the path P contains the zig-zag path with both endpoints in the top row of H_i and that contains every vertex of H_i.
2. For each $a \in \Sigma$ such that $\delta_a(w) < 0$ we proceed symmetrically to the previous case, but reversing the roles of X_a and Y_a. Specifically, we group the elements of Y_a into pairs. For each pair (i, j), P contains the path through the top row of H_i

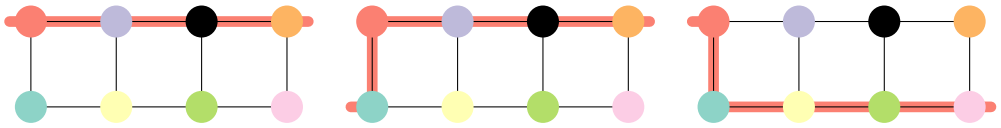

Figure 5: The paths taken by P through boring blocks: A top path, a downup path, and an updown path.

and the path through the bottom row of H_j. For each element $i \in X_a$, P contains the zig-zag path with both endpoints in the top row of H_i and that contains every vertex of H_i.

3. For each $i \in \{1, \ldots, 2r\} \setminus (X \cup Y)$, P contains the top row of H_i.

The rules above define the intersection, P_i, of P with each colourful block H_i of G. If P_i is the path through the bottom row of H_i then we call H_i a *bottom block*. If P_i is the path through the top row of H_i then we call H_i a *top block*. If P_i is the zig-zag path that contains every vertex of H_i then we call H_i a *zig-zag block*. Note that $\sum_{a \in \Sigma_3} \delta_a(w) = 0$ and this implies that the number of bottom blocks among $H_1, \ldots, H_{r-1}$ is the same as the number of bottom blocks among $H_{r+1}, \ldots, H_{2r}$. Indeed, this number is exactly $\frac{1}{2}\tau(w)$.

We now define how P behaves for the boring blocks, denoted by $Q_0, \ldots, Q_{2r-1}$. The first boring block Q_0 comes immediately before H_1. Each boring block Q_j, for $j \in \{1, \ldots, 2r-1\}$ comes immediately after H_j and immediately before H_{j+1}. In almost every case, P uses the path through the top row of Q_j. The only exceptions are when H_j or H_{j+1} are bottom blocks. Note that, because of the global independence constraint, these two cases are mutually exclusive. See Figure 5.

1. When H_j is a bottom block, P uses a path that begins at the bottom row of Q_j but moves immediately to the top row of Q_j and uses the entire path along the top row. We call this a *downup* path.
2. When H_{j+1} is a bottom block, P uses a path that begins at the top row of Q_j and moves immediately to to the bottom row of Q_j and uses the entire path along the bottom row. We call this a *updown* path.

This completely defines the path $P := v_1, \ldots, v_{2m}$. All that remains is to argue that the word $\rho := \phi_w(v_1), \ldots, \phi_w(v_{2m})$ is anagramish.

Observe that the number of downup paths and the number of updown paths that appear in $Q_0, \ldots, Q_{r-1}$ is exactly the same as the number of bottom blocks among $H_1, \ldots, H_{r-1}$ which is exactly $\frac{1}{2}\tau(w)$. Similarly, the number of updown paths and downup paths that appear in $Q_{r+1}, \ldots, Q_{2r-1}$ is exactly $\frac{1}{2}\tau(w)$. Now every path that is neither downup nor updown uses the top row. This implies that the sequence of colours contributed to ρ by the intersection of P with $Q_0, \ldots, Q_{r-1}$ is a permutation of the sequence of colours contributed to ρ by the intersection of P with $Q_{r+1}, \ldots, Q_{2r-1}$. These two sequences cancel perfectly.

Next, by construction, each pair of top and bottom blocks in $H_1, \ldots, H_{r-1}$ contributes exactly the same amount as a single matching zig-zag block in $H_{r+1}, \ldots, H_{2r-1}$. Specif-

ically, if $x, y \in X_a$, $z \in Y_a$, H_x is a top block, H_y is a bottom block and H_z is a zig-zag block, then the contributions of P_x and P_y to ρ cancels out the contribution of P_z. After doing this cancellation exhaustively, all that remains are top blocks, which also cancel each other perfectly. This completes the proof. □

To be explicit, we finish the proof of Theorem 1 by pointing out the contradiction between Lemma 2 and Lemma 4:

Proof of Theorem 1. Assume for the sake of contradiction that there exists some $c \in \mathbb{N}$ such that $\dot{\chi}_\pi(G_n) \leqslant c$ for each $n \in \mathbb{N}$. Then the language L_1 is infinite and, by Lemma 3 so are L_2 and L_3. Therefore L_3 is an infinite periodic language. Therefore, Lemma 2 implies that for every $\epsilon > 0$ and $r_0 \in \mathbb{N}$ there exists a word $w \in L_3$ with $2r := |w| \geqslant 2r_0$, and $\tau(w) \leqslant \epsilon r$. In particular, this is true for the specific values of r_0 and ϵ that appear in Lemma 4. However, Lemma 4 implies that ϕ_w is not an anagram-free colouring of $G_{n(w)}$, which contradicts the fact that $w \in L_3$. □

4 Proof of Lemma 2

All that remains is to prove Lemma 2, which we do now.

Proof of Lemma 2. First observe that, for any ℓ-periodic word w, $|\Sigma_w| \leqslant \ell$. Define an even-length word t to be *a-unbalanced* if $\tau_a(t) > \epsilon|t|/\ell$ and *a-balanced* otherwise. If t is a-balanced for each $a \in \Sigma_t$ then t is *balanced*. Observe that, if t is balanced then $\tau(t) \leqslant |\Sigma_t|\epsilon|t|/\ell \leqslant \epsilon|t|$. A word is *everywhere unbalanced* if it contains no balanced factor of length $r \geqslant r_0$. Our goal therefore is to show that there is an upper bound $n := n(\ell, \epsilon, r_0)$ on the length of any ℓ-periodic everywhere unbalanced word.

Let h be a positive integer whose value will be discussed later and let $n := r_0 2^h$. Let w be an ℓ-periodic everywhere unbalanced word of length n over the alphabet Σ. Assume, without loss of generality, that r_0 is a multiple of ℓ.

Consider the complete binary tree T of height h whose leaves, in order, are length-r_0 words whose concatenation is w and for which each internal node is the factor obtained by concatenating the node's left and right child. Note that for each $v \in V(T)$ and each $a \in \Sigma$, the fact that w is ℓ-periodic and r_0 is multiple of ℓ implies that $|v|_a \geqslant |v|/\ell$.

For each $a \in \Sigma$, let $S_a := \{v \in V(T) : v \text{ is } a\text{-unbalanced}\}$. Since w is everywhere unbalanced, $\bigcup_{a\in\Sigma} S_a = V(T)$. Therefore,

$$(h+1)n = \sum_{v\in V(T)} |v| \leqslant \sum_{a\in\Sigma} \sum_{v\in S_a} |v| \ .$$

Therefore, there exists some $\alpha \in \Sigma$ such that $\sum_{v\in S_\alpha} |v| \geqslant (h+1)n/|\Sigma| \geqslant (h+1)n/\ell$. At this point we are primarily concerned with appearances of α, so let $X := S_\alpha$, and, for each node $v \in V(T)$, let $W(v) := |v|_\alpha$.

For each non-leaf node v of T, let $R(v)$ denote a child of v such that $W(R(v)) \leqslant \frac{1}{2} \cdot W(v)$. It is helpful to think of T as being ordered so that each right child y with sibling x has $W(y) \leqslant W(x)$. For a non-leaf node $v \in X$ the fact that v is α-unbalanced implies that

$$W(R(v)) \leqslant \tfrac{1}{2} \cdot W(v) - \tfrac{\epsilon}{2\ell} \cdot |v| \leqslant (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell})W(v) \ .$$

From this point on we use the following shorthands. For any $S \subseteq V(T)$, $L(S) := \sum_{v \in S} |v|$, $W(S) := \sum_{v \in S} W(v)$, and $R(S) = \{R(v) : v \in S\}$. Summarizing, we have a complete binary tree T of height h and $X \subseteq V(T)$ with the following properties:

1. For each $v \in V(T)$, $W(v) \geqslant |v|/\ell$.
2. $L(X) \geqslant (h+1)n/\ell$.
3. For each non-leaf node $v \in X$, $W(R(v)) \leqslant (\frac{1}{2} - \frac{\epsilon}{2\ell})W(v)$.

For each $i \in \{0, \ldots, h\}$, let $X_i \subseteq X$ denote the set of nodes $v \in X$ for which the path from the root of T to v contains exactly i nodes in X, excluding v. See Figure 6. Observe that, since each node in X_i has an ancestor in X_{i-1},

$$n \geqslant L(X_0) \geqslant L(X_1) \geqslant \cdots \geqslant L(X_h) \ .$$

We will show that there exists an integer $t := t(\epsilon, \ell, r_0)$ such that, for each $i \in \{0, \ldots, h-t\}$,

$$L(X_{i+t}) \leqslant (1 - (1/2)^{t+1})L(X_i) \ . \tag{2}$$

In this way,

$$\begin{aligned}
\frac{(h+1)n}{\ell} \leqslant L(X) &= \sum_{i=0}^{h} L(X_i) \\
&\leqslant \sum_{i=0}^{h} L(X_{t\lfloor i/t \rfloor}) && \text{(since } i \geqslant t\lfloor i/t \rfloor\text{, so } L(X_i) \leqslant L(X_{i\lfloor i/t \rfloor})\text{)} \\
&= t \cdot \sum_{i=0}^{h/t} L(X_{it}) && \text{(for } h \text{ a multiple of } t\text{)} \\
&\leqslant t \cdot \sum_{i=0}^{\infty} (1 - (1/2)^{t+1})^i L(X_0) && \text{(by Equation (2))} \\
&\leqslant tn \cdot \sum_{i=0}^{\infty} (1 - (1/2)^{t+1})^i && \text{(since } |X_0| \leqslant n\text{)} \\
&= tn2^{t+1}
\end{aligned}$$

which is a contradiction for sufficiently large h; in particular, for $h > \ell t 2^{t+1} - 1$.

It remains to establish Equation (2), which we do now. Fix some $i \in \{0, \ldots, h-1\}$, define $A_0 := X_i$ and, for each $j \geqslant 1$, define A_j to be the subset of X_{i+j} that are descendants

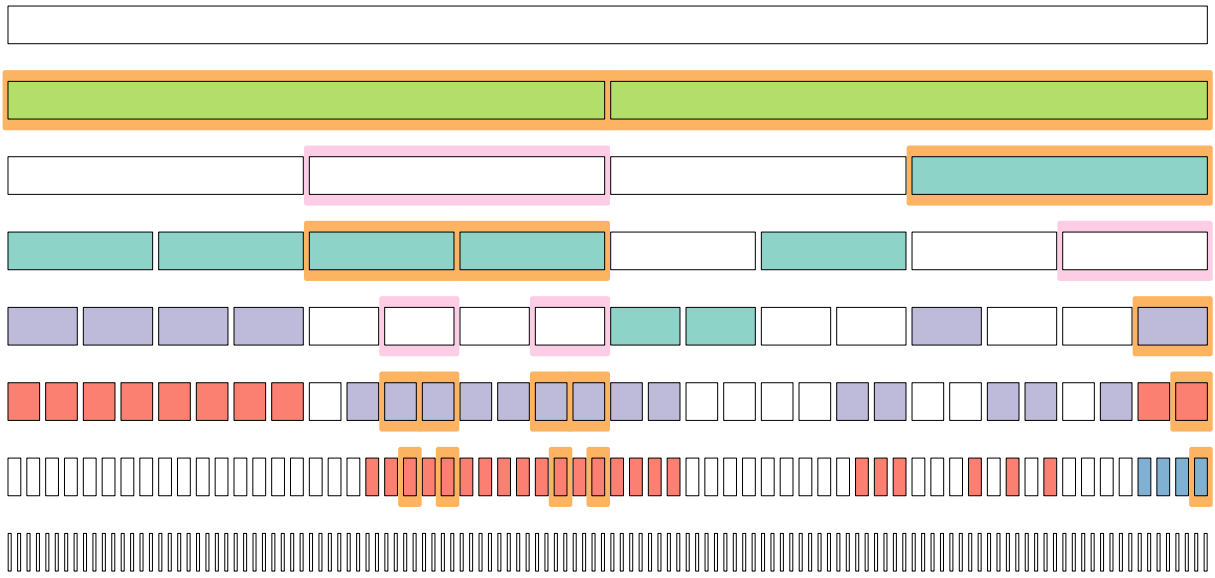

Figure 6: The partitioning of X into $X_0, \ldots, X_h$. Shaded nodes are in X and all nodes in X_i are shaded with the same colour. Starting with $A_0 = X_0$, the elements of $A_0, \ldots, A_h$ are highlighted (in orange). The elements of $R(A_0), \ldots, R(A_h)$ are also highlighted (in pink).

of some node in $R(A_{j-1})$. See Figure 6. To upper bound $L(X_{i+t})$ observe that X_{i+t} can be split into two sets A'_0 and B defined as follows: The nodes A'_0 do not have an ancestor in $R(A_0)$ and therefore $L(A'_0) \leqslant (1/2)L(A_0)$. The nodes in B do have an ancestor in $R(A_0)$ and therefore have an ancestor in A_1. Iterating this argument, we obtain

$$\begin{aligned} L(X_{i+t}) &\leqslant \sum_{j=0}^{t-1} (1/2)L(A_j) + L(A_t) \\ &\leqslant \sum_{j=0}^{t-1} (1/2^{j+1})L(A_0) + L(A_t) \\ &= (1-(1/2)^t)L(A_0) + L(A_t) \\ &= (1-(1/2)^t)L(X_i) + L(A_t) \ . \end{aligned}$$

So all that remains to establish 2 is to prove that $L(A_t) \leqslant (1/2)^{t+1}L(X_i)$. To do this, observe that, for each $j \in \{1, \ldots, t\}$,

$$W(A_j) \leqslant W(R(A_{j-1})) \leqslant (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell}) \cdot W(A_{j-1}) \tag{3}$$

which implies

$$W(A_t) \leqslant (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell})^t W(A_0) \leqslant (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell})^t \cdot L(A_0) = (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell})^t \cdot L(X_i)$$

Since w is ℓ-periodic,

$$L(A_t) \leqslant \ell \cdot W(A_t) \leqslant \ell \cdot (\tfrac{1}{2} - \tfrac{\epsilon}{2\ell})^t \cdot L(X_i) \leqslant L(X_i)/2^{t+1}$$

for $t := \lceil \log(2\ell)/\log((1/(1-\frac{\epsilon}{\ell}))\rceil$. □

5 Reflections

Although an explicit upper bound on $n := n(\epsilon, \ell, r_0)$ could be extracted from the proof of Lemma 2 it would likely be far from tight. We suspect that there is a Fourier analytic proof that would give better quantitative bounds. We have not pursued this, because we have no idea how to explicitly upper bound ℓ, for reasons discussed in the next paragraph.

Lemma 3 and its proof give absolutely no clues to help find a concrete bound on ℓ or to find a minimal set Ξ. Indeed, for some choices of L, doing so can be a difficult problem. Consider the example where $|\Sigma| = 5$ and L is the language of all anagram free words in Σ^*. It is easy to see that this language L is factor-closed and the result of Pleasants [11], published in 1970, shows that this L is infinite. The question of whether $|\Xi| = 4$ or $|\Xi| = 5$ is then the question of determining whether there exist arbitrarily long anagram-free words on an alphabet of size 4. This was the open problem posed by Erdős [3] in 1961 and again by Brown [1] in 1971 and not resolved until 1992 when Keränen [7, 8] showed that the answer, in this case, is that $|\Xi| = 4$. However, if this were not the case, then determining ℓ would be the question of determining the length of the longest anagram-free word over an alphabet of size 4.

Our proof uses Lemma 3 twice and each uses a language L_3 that is considerably more complicated than the language of anagram-free words. It seems unlikely that we will obtain concrete upper bounds on ℓ as a function of c except, possibly, through the use of computer search. The resulting value ℓ is used in the application of Lemma 2 and also within the proof of Lemma 4.

Acknowledgement

We would like to thank an anonymous referee for pointing out existing results and terminology from the field of combinatorics on words that has helped streamline the presentation in Sections 2 and 3.

References

[1] T. C. Brown. Is there a sequence on four symbols in which no two adjacent segments are permutations of one another? *The American Mathematical Monthly*, 78(8):886–888, 1971. doi:10.1080/00029890.1971.11992892.

[2] Paz Carmi, Vida Dujmović, and Pat Morin. Anagram-free chromatic number is not pathwidth-bounded. In Andreas Brandstädt, Ekkehard Köhler, and Klaus Meer, editors, *Graph-Theoretic Concepts in Computer Science - 44th International Workshop, WG 2018, Cottbus, Germany, June 27-29, 2018, Proceedings*, volume 11159 of *Lecture Notes in Computer Science*, pages 91–99. Springer, 2018. doi:10.1007/978-3-030-00256-5_8.

[3] P. Erdős. Some unsolved problems. *Magyar Tud Akad Mat Kutato Int Kozl*, 6:221–254, 1961.

[4] A. A. Evdokimov. Strongly asymmetric sequences generated by finite number of symbols. *Doklady Akademii Nauk SSSR*, 179:1268–1271, 1968.

[5] A. A. Evdokimov. Strongly asymmetric sequences generated by finite number of symbols. *Soviet Mathematics Doklady*, 9:536–539, 1968.

[6] Nina Kamčev, Tomasz Łuczak, and Benny Sudakov. Anagram-free colourings of graphs. *Comb. Probab. Comput.*, 27(4):623–642, 2018. doi:10.1017/S096354831700027X.

[7] Veikko Keränen. Abelian squares are avoidable on 4 letters. In Werner Kuich, editor, *Automata, Languages and Programming, 19th International Colloquium, ICALP92, Vienna, Austria, July 13-17, 1992, Proceedings*, volume 623 of *Lecture Notes in Computer Science*, pages 41–52. Springer, 1992. doi:10.1007/3-540-55719-9_62.

[8] Veikko Keränen. A powerful abelian square-free substitution over 4 letters. *Theor. Comput. Sci.*, 410(38-40):3893–3900, 2009. doi:10.1016/j.tcs.2009.05.027.

[9] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, U.K., 2002.

[10] Jaroslav Nešetřil and Patrice Ossona de Mendez. Tree-depth, subgraph coloring and homomorphism bounds. *Eur. J. Comb.*, 27(6):1022–1041, 2006. doi:10.1016/j.ejc.2005.01.010.

[11] P. A. B. Pleasants. Non-repetitive sequences. *Proceedings of the Cambridge Philosophical Society*, 68:267–274, 1970. doi:10.1017/S0305004100046077.

[12] Tim E. Wilson. *Anagram-free Graph Colouring and Colour Schemes*. Ph.D. thesis, Monash University, 2019. doi:10.26180/5c72eca26d5c7.

[13] Tim E. Wilson and David R. Wood. Anagram-free colorings of graph subdivisions. *SIAM J. Discret. Math.*, 32(3):2346–2360, 2018. doi:10.1137/17M1145574.

[14] Tim E. Wilson and David R. Wood. Anagram-free graph colouring. *Electron. J. Comb.*, 25(2):P2.20, 2018. doi:10.37236/6267.