POSET SATURATION AND OTHER COMBINATORIAL RESULTS

Maria-Romina Ivan

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
St John's College

ST JOHN'S COLLEGE
UNIVERSITY OF CAMBRIDGE

This dissertation is submitted for the degree of
Doctor of Philosophy
May 2023

Poset saturation and other combinatorial results

Maria-Romina Ivan

Abstract

In this dissertation we discuss a number of combinatorial results. These results fall into four broad areas: poset saturation, Ramsey theory, pursuit and evasion, and union-closed families.

Chapter 2 is dedicated to the area of poset saturation. Given a finite poset $\mathcal{P}$, we call a family $\mathcal{F}$ of subsets of $[n]$ $\mathcal{P}$-saturated if $\mathcal{F}$ does not contain an induced copy of $\mathcal{P}$, but adding any other set to $\mathcal{F}$ creates an induced copy of $\mathcal{P}$. The size of the smallest $\mathcal{P}$-saturated family with ground set $[n]$ is called the induced saturated number of $\mathcal{P}$, which is denoted by $\text{sat}^*(n, \mathcal{P})$.

In this chapter we look at four posets: the butterfly, the diamond, the antichain and the poset $\mathcal{N}$. We establish a linear lower bound for the butterfly, a lower bound of $(2\sqrt{2} - o(1))\sqrt{n}$ for the diamond, a lower bound of $\sqrt{n}$ for the poset $\mathcal{N}$, and the exact saturation number for the 5-antichain and the 6-antichain.

Chapter 3 is dedicated to two different Ramsey theory questions. In Section 3.1 we establish a Ramsey characterisation of eventually periodic words. More precisely, for a finite colouring of X^* (the set of finite words on alphabet X) we say that a factorisation $x = u_1 u_2 \cdots$ of an infinite word x is 'super-monochromatic' if each word $u_{k_1} u_{k_2} \cdots u_{k_n}$, where $k_1 < \cdots < k_n$, is the same colour. We show that a word x is eventually periodic if and only if for every finite colouring of X^* there is a suffix of x having a super-monochromatic factorisation. This has been a conjecture for quite some time.

In Section 3.2 we investigate the question of whether or not, given a finite colouring of the rationals or the reals, we can find an infinite subset with the property that the set of all its finite sums and products is monochromatic. The main result of this section is the existence of a finite colouring of the rationals with the property that no infinite set whose denominators contain only finitely many primes has the set of all of its finite sums and products monochromatic.

In Chapter 4 we explore the game of cops and robbers on infinite graphs. The main question is: for which graphs can one guarantee that the cop has a winning strategy? In the finite case these graphs are precisely the 'constructible' graphs, but the infinite case is not well understood. For example, we exhibit a graph that is cop-win but not constructible. This is the first known such example.

On the other hand, every constructible graph is a weak cop win (meaning that the cop can eventually force the robber out of any finite set). We also investigate how this notion relates to the notion of 'locally constructible' (every finite graph is contained in a finite constructible subgraph). The main result of this chapter is the construction of a locally constructible graph that is not a weak cop win. Surprisingly, this graph may even be chosen to be locally finite.

Finally, in Chapter 5 we discuss the union-closed conjecture which asserts that for any union-closed family of sets, there exists an element of the ground set contained in at least half of the sets of the family. Our attention is on the small sets of union-closed families. More precisely, we construct a class of union-closed families of sets such that the frequency of the elements of the minimal sets is $o(1)$ – so that these elements are not generally in half of the sets of union-closed families.

Declaration

This dissertation is the result of my own work. Chapter 2 includes nothing which is the outcome of work done in collaboration except for Section 2.4, which was done in collaboration with Irina Đanković. For Chapter 3, Section 3.1 represents work done in collaboration with Imre Leader and Luca Q. Zamboni, and Section 3.2 represents work done in collaboration with Neil Hindman and Imre Leader. Chapter 4 is the outcome of work done in collaboration with Imre Leader and Mark Walters, and Chapter 5 is the outcome of work done in collaboration with David Ellis and Imre Leader. This dissertation bears no resemblance to any dissertation that I have submitted, or is being submitted, for any other degree or qualification at the University of Cambridge or at any other university, or similar institution.

Acknowledgements

I would like to take this opportunity and thank my supervisor, Professor Imre Leader, for everything he has done for me. Not only has he mentored me mathematically and shared with me his wealth of knowledge and inspiring dedication for academia, but he has very quickly and simultaneously became my second father and my best friend. I would like to thank him for all the coffees, all the late night phone calls, all the last minute meetings, and the countless jokes that brought us closer. My life would be completely different without him, and I feel so extremely lucky to have been taken under his wing. I will forever be in his debt, although he is ridiculously modest to acknowledge this.

I also want to thank my family for their life long support and for all the sacrifices that they have made just because they believed in me (or because sometimes they did not have a choice). Firstly, I would like to thank my grandparents, Elena and Niculae Sorescu, who raised me in a little mountain village and showed me that life can be both hard and extremely beautiful. They have planted in me the strength that I needed to power through what the future has thrown at me. Secondly, I would not be the person that I am today without my parents, Nicoleta and Cristian-Mihai Ivan. They have been and still are my role models. They have taught the importance of integrity, honesty and hard work, and they have done so by being an impeccable example themselves. I watched them build their life from zero with no short-cuts and no despair. This was my second proof that life is both hard and beautiful.

There is one member of my family that needs special acknowledgement: my partner, Paul Minter. I want to thank him from all my heart for how much he has helped me during some of the hardest moments over the past four years. He has been my rock, my home away from home, my closest confidant, and my biggest fan. He showed me what true love really means, and how much we can achieve if we help each other.

I have also been blessed with one of the most loyal and kind friends, Anđela Šarković, who has been by my side since October 2015, our first day in Cambridge. She helped me surpass many hard moments over the last eight years thanks to her intoxicating positivity and relatable sense of humour.

It is impossible not to thank my dear friend and dancing partner, Ivan Scales, for all the dances and the laughter we have shared over the last six years. I want to thank him for our occasional coffees that detached me from work and recharged my batteries, for his help and advice regarding academia and life in general, and for always being there when ‘the maths nerd’ needed a ‘normal’ friend.

The list can go on. I am extremely grateful for all the people in my life that have turned it into a beautiful adventure. Here I am talking about all my friends, old and new, fellow mathematicians I met and exchanged ideas with, staff members such as my dear St John’s porters who made me feel at home, and each and every other person who influenced my journey up to this point – I have never taken any of you for granted.

Art is what created me,
Art is what I seek.
Art is what set me free
To create what is unique.

I leave below a humble mark,
A faint brush stroke on life's painting,
A new light in the infinite dark,
A piece of art worth framing.

Written in Crêpeaffaire, Cambridge UK

Contents

1 Introduction **1**

2 Poset Saturation **5**
2.1 Introduction . . . 5
2.2 The butterfly . . . 9
2.2.1 A linear lower bound on $\mathrm{sat}^*(n, \mathcal{B})$ . . . 9
2.2.2 Further analysis of singletons . . . 12
2.3 The diamond . . . 17
2.3.1 The main result . . . 17
2.3.2 Could $\mathrm{sat}^*(n, \mathcal{D}_2)$ be $O(\sqrt{n})$? . . . 23
2.4 Small antichains . . . 25
2.4.1 5-antichain saturation . . . 25
2.4.2 6-antichain saturation . . . 29
2.4.3 Recent developments . . . 31
2.5 An improved bound on $\mathrm{sat}^*(n, \mathcal{N})$ . . . 33
2.6 Extensions and further work . . . 35

3 Two Ramsey Theory Questions **39**
3.1 A Ramsey characterisation of eventually periodic words . . . 39
3.1.1 Introduction . . . 39
3.1.2 The link between the two main theorems . . . 41
3.1.3 Constructing the colouring Λ . . . 45
3.1.4 Conclusions and open problems . . . 57
3.2 Monochromatic sums and products over the rationals . . . 59
3.2.1 Introduction . . . 59
3.2.2 Some useful lemmas . . . 60
3.2.3 Colouring the naturals . . . 62
3.2.4 Colouring the reals . . . 64
3.2.5 Combining an extension of θ over the rationals with ν . . . 68
3.2.6 Exploring μ further . . . 69
3.2.7 Unbounded sequences in the rationals . . . 71
3.2.8 Concluding remarks . . . 75
3.2.9 Appendix . . . 76

4 Constructible Graphs and Pursuit **79**
4.1 Introduction . . . 79
4.2 The graph K and and a finite application . . . 82
4.3 A non-constructible cop-win graph . . . 84
4.4 Locally constructible graphs . . . 85
4.5 Locally constructible graphs may not be weak cop-win . . . 88
4.6 The construction time of constructible graphs . . . 96

4.7 Open problems 99

5 Small Sets in Union-Closed Families **101**

5.1 Introduction 101

5.2 Small sets in union-closed families 103

5.3 An open problem 105

6 Bibliography **107**

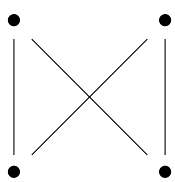
1 Introduction

This dissertation is divided into four chapters about poset saturation, Ramsey theory, pursuit and evasion and the union-closed conjecture. Every chapter is split into smaller sections, each dedicated to a separate result in the area. Below we give an overview of the results presented in each chapter.

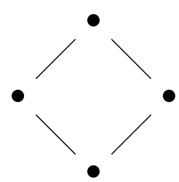
2. Poset Saturation

Given a finite poset $\mathcal{P}$, we call a family $\mathcal{F}$ of subsets of $[n]$ $\mathcal{P}$-saturated if $\mathcal{F}$ does not contain an induced copy of $\mathcal{P}$, but adding any other set to $\mathcal{F}$ creates an induced copy of $\mathcal{P}$. We want to find the induced saturated number of $\mathcal{P}$, denoted by $\mathrm{sat}^*(n, \mathcal{P})$, which is the size of the smallest $\mathcal{P}$-saturated family with ground set $[n]$. It is worth mentioning that this question is very different from the Turán-type questions that ask for the maximal size of such families or graphs. Surprisingly, the saturation question is not at all trivial even for small posets.

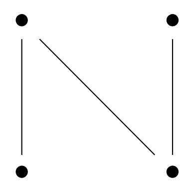
In this chapter we analyse four posets: the butterfly, the diamond, the poset $\mathcal{N}$ and the k-antichain (collection of k pairwise incomparable elements). The Hasse diagrams of the first three posets are displayed below:



The butterfly, or $\mathcal{B}$



The diamond, or $\mathcal{D}_2$



The poset $\mathcal{N}$

Ferrara, Kay, Krammer, Martin, Reiniger, Smith and Sullivan [19] proved that for a large class of posets, including the butterfly and $\mathcal{N}$, the induced saturated number is at least the biclique cover number of the complete graph on n vertices, namely $\log_2 n$. We improve on this result by establishing a linear lower bound for the butterfly, denoted by $\mathcal{B}$, and a lower bound of $\sqrt{n}$ for the poset $\mathcal{N}$. The linear lower bound for the butterfly was later proved to be sharp [38].

For the diamond poset, denoted by $\mathcal{D}_2$, (the two-dimensional Boolean lattice), Martin, Smith and Walker [41] proved that $\sqrt{n} \leq \mathrm{sat}^*(n, \mathcal{D}_2) \leq n + 1$. We prove that $\mathrm{sat}^*(n, \mathcal{D}_2) \geq (2\sqrt{2} - o(1))\sqrt{n}$. We also explore the properties that a diamond-saturated family of size $c\sqrt{n}$, for a constant c, would have to have.

For the antichain with k elements, denoted by $\mathcal{A}_k$, Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] conjectured that $\mathrm{sat}^*(n, \mathcal{A}_k) = (k-1)n(1+o(1))$, and proved this for $k \leq 4$. We prove this conjecture for $k = 5$ and $k = 6$. Moreover, we give the exact value for $\mathrm{sat}^*(n, \mathcal{A}_5)$ and $\mathrm{sat}^*(n, \mathcal{A}_6)$. Since then, Bastide, Groenland, Jacob and Johnston [7] have proved the conjecture for general k.

The work presented in this chapter has been published in [31], [32] and [3].

3. Two Ramsey Theory Questions

Ramsey theory has at its core the question 'Can we find some order in enough disorder?'. The majority of Ramsey-type questions involve a *finite* colouring of a certain 'chaotic' object and asks for a *monochromatic* more 'ordered' substructure. In this chapter we discuss two very different such questions.

In the first section we provide a Ramsey characterisation of eventually periodic words. To start with, a factorisation $x = u_1u_2\cdots$ of an infinite word x on alphabet X is called 'monochromatic', for a given colouring of the finite words X^* on alphabet X, if each u_i is the same colour. Trivially, every periodic word has a monochromatic factorisation for any finite colouring of X^*. Wojcik and Zamboni [57] proved that this is in fact a necessary condition. In other words, an infinite word x is periodic if and only if for every finite colouring of X^* there is a monochromatic factorisation of x. This provides a Ramsey characterisation of periodic words.

A much stronger notion for a factorisation than being monochromatic is being 'super-monochromatic'. More precisely, say that a factorisation $x = u_1u_2\cdots$ is super-monochromatic if each word $u_{k_1}u_{k_2}\cdots u_{k_n}$, where $k_1 < \cdots < k_n$, is the same colour. A direct application of Hindman's theorem shows that, given a finite colouring of X^*, every eventually periodic word has a super-monochromatic factorisation. It has been a conjecture in the community for quite some time that a word x is eventually periodic if and only if for every finite colouring of X^* there is a suffix of x having a super-monochromatic factorisation. In this section we prove this conjecture. This result is published in [35].

The second section is concerned with finite colourings of the naturals, rationals, or reals numbers and the question of whether we can find an infinite set whose finite sums and products all have the same colour. Hindman [27] showed that one cannot ask for sums and products, even just pairwise: there is a finite colouring of the naturals for which no (injective) sequence has the set of all of its pairwise sums and products monochromatic. Our work focuses on the question of what happens over the rationals, for which not much was known.

Our main result is that, for any k, there is a finite colouring of the set of rationals whose denominators contain only the first k primes such that no infinite set has all of its finite sums and products monochromatic. We actually prove a 'uniform' form of this: there is a finite colouring of the rationals with the property that no infinite set whose denominators contain only finitely many primes has all of its finite sums and products monochromatic. We also give various other related results, including a new short proof of the result of Hindman mentioned above, that there is a finite colouring of the naturals such that no infinite set has the set of all of its pairwise sums and products monochromatic.

All results presented in this chapter are published in [33].

4. Constructible Graphs and Pursuit

The area of pursuit and evasion deals with problems where, in a certain fixed set-up with predetermined rules, two players track each other, one trying to pursue and the other trying to evade.

In this chapter we study a problem where the set-up is a graph and the two players are the 'cop' and the 'robber'. They choose their initial vertices and then they take turns and move along edges to new vertices. The question is: for what graphs can one guarantee that the cop always has a strategy to catch the robber (that is when the cop lands on the robber's vertex)? For example, on a triangle, no matter where the players are, the cop catches the robber on his first move, while on a square, if the players are at opposite corners, the robber always moves away from the cop, thus always being diagonally opposite. The finite cop-win graphs were characterised by Nowakowski and Winkler [44], and they are precisely the 'constructible' graphs: a graph is called constructible if it may be obtained recursively from the one-point graph by repeatedly adding dominated vertices.

What about infinite graphs? It turns out that this question is wildly different from the finite version. One of the main results in this chapter is the construction of a graph that is cop-win but not constructible. This is the first known such example. We also show that every countable ordinal arises as the rank of some constructible graph, answering a question of Evron, Solomon and Stahl [18]. As an unexpected spin-off of our methods, we are able to exhibit a finite constructible graph for which there is no construction order whose associated domination map is a homomorphism, answering a question of Chastand, Laviolette and Polat [13].

Lehner [40] showed that every constructible graph is a weak cop win (meaning that the cop can eventually force the robber out of any finite set). We also investigate how this notion relates to the notion of 'locally constructible' (every finite graph is contained in a finite constructible subgraph). We show that, under mild extra conditions, every locally constructible graph is a weak cop win. But we also give an example to show that, in general, a locally constructible graph need not be a weak cop win. Surprisingly, this graph may even be chosen to be locally finite. All results are published in [34].

5. Small Sets in Union-Closed Families

If X is a set, a family $\mathcal{F}$ of subsets of X is said to be *union-closed* if the union of any two sets in $\mathcal{F}$ is also in $\mathcal{F}$. The union-closed conjecture (a conjecture of Frankl [20]) states that if X is a finite set and $\mathcal{F}$ is a union-closed family of subsets of X (with $\mathcal{F} \neq \{\emptyset\}$), then there exists an element $x \in X$ such that x is contained in at least half of the sets in $\mathcal{F}$. Despite the efforts of many researchers over the last forty-five years, and a recent Polymath project [1] aimed at resolving it, this conjecture remains open. Recently Gilmer [24] showed that there exists $c > 0$ such that for any union-closed family there exists an element of the ground set contained in a proportion of at least c of the sets.

In order to solve the conjecture, one might be tempted to look at the elements of the minimal-size sets of a union-closed family as possible candidates for the elements contained in at least half of the sets. In all previous examples of union-closed families, there was at least one element of a minimal-size set that was contained in at least a third of the sets. However, in this chapter we show that, for any $\epsilon > 0$, there exists a union-closed family $\mathcal{F}$ with (unique) smallest set S such that no element of S belongs to more than a fraction ϵ of the sets in $\mathcal{F}$. More precisely, we give an example of a union-closed family with smallest set of size k such that no element of this set belongs to more than a fraction $(1+o(1))\frac{\log_2 k}{2k}$ of the sets in $\mathcal{F}$. This work has been published in [17].

2 Poset Saturation

2.1 Introduction

We say that a poset $(\mathcal{P}, \preceq)$ contains an *induced copy* of a poset $(\mathcal{Q}, \preceq')$ if there exists an injective order-preserving function $f : \mathcal{Q} \to \mathcal{P}$ such that $(f(\mathcal{Q}), \preceq)$ is isomorphic to $(\mathcal{Q}, \preceq')$. If elements a and b of a poset are not related, or incomparable, we write $a \parallel b$.

In this chapter we consider the power-set of $[n] = \{1, 2, \cdots, n\}$ with the partial order induced by inclusion. If $\mathcal{Q}$ is a finite poset and $\mathcal{F}$ is a family of subsets of $[n]$, we say that $\mathcal{F}$ is $\mathcal{Q}$ *saturated* if $\mathcal{F}$ does not contain an induced copy of $\mathcal{Q}$, and for any $S \notin \mathcal{F}$, $\mathcal{F} \cup S$ contains an induced copy of $\mathcal{Q}$. The smallest size of a $\mathcal{Q}$-saturated family of subsets of $[n]$ is called the *induced saturated number*, denoted by $\mathrm{sat}^*(n, \mathcal{Q})$.

We mention that induced and non-induced poset saturation is a growing area in combinatorics. Saturation for posets was introduced by Gerbner, Keszegh, Lemons, Palmer, Pálvölgyi and Patkós [22], although this was not for *induced* saturation. We refer the reader to the textbook of Gerbner and Patkós [23] for a nice introduction to the area.

As remarked in the introduction, determining the *exact* saturation number proves to be a difficult question. However, there have been a couple of global results which reveal that $\mathrm{sat}^*(n, \mathcal{P})$ has a dichotomy of behaviour.

Keszegh, Lemons, Martin, Pálvölgyi and Patkós [38] proved that for any poset, the induced saturated number is either constant, or at least the biclique cover number of the complete graph on n vertices, namely $\log_2 n$. Recently, Freschi, Piga, Sharifzadeh and Treglown [21] improved their result by replacing $\log_2 n$ with $2\sqrt{n-2}$.

Keszegh, Lemons, Martin, Pálvölgyi and Patkós [38] conjectured that for any poset the saturation number is either constant, or at least $n + 1$. Finally, since there is no known poset $\mathcal{P}$ for which $\mathrm{sat}^*(n, \mathcal{P}) = \omega(n)$, it is in fact believed that for any poset, the saturation number is either constant, or linear.

The posets for which we will analyse the induced saturated number are the butterfly (Figure 1) which we denote by $\mathcal{B}$, the diamond (Figure 2) which we denote by $\mathcal{D}_2$, the poset $\mathcal{N}$ (Figure 3), and the k-antichain (collection of k pairwise incomparable elements) which we denote by $\mathcal{A}_k$.

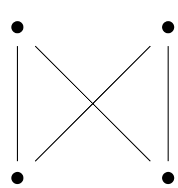


Figure 1

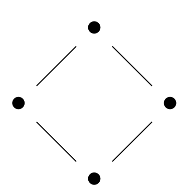


Figure 2

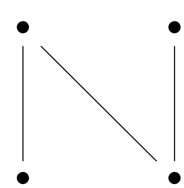


Figure 3

Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] showed that the saturation number for both the butterfly and the poset $\mathcal{N}$ is at least $\log_2 n$. They also provided an upper bound of $\binom{n}{2}+2n-1$ for the butterfly (the family $\{\emptyset, \{i\}, \{i,j\}, \{1,2,\cdots,i\}$:

$1 \leq i \leq n, 1 \leq j \leq n\}$ is butterfly-saturated), and an upper bound of $2n$ for $\text{sat}^*(n, \mathcal{N})$ (the family $\{\emptyset, \{i\}, \{1, 2, \cdots, i\} : 1 \leq i \leq n\}$ is $\mathcal{N}$-saturated).

In Section 2.2 we prove the following result:

Theorem. *([31]) sat*$^*(n, \mathcal{B}) \geq n + 1$.

Shortly after, Keszegh, Lemons, Martin, Pálvölgyi and Patkós [38] showed that $\text{sat}^*(n, \mathcal{B}) \leq 6n$, thus $\text{sat}^*(n, \mathcal{B}) = \Theta(n)$. It is worth mentioning that the butterfly poset is one of the few non-trivilal posets for which the saturation number is known, up to constants.

In Section 2.5 we improve on the lower bound for $\mathcal{N}$:

Theorem. *([31]) sat*$^*(n, \mathcal{N}) \geq \sqrt{n}$.

For the diamond poset, despite its simplicity, the question of its induced saturation number is still open. Martin, Smith and Walker [41] proved that $\sqrt{n} \leq \text{sat}^*(n, \mathcal{D}_2) \leq n+1$, and Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] conjectured that $\text{sat}^*(n, \mathcal{D}_2) = \Theta(n)$.

In Section 2.3 we improve on the constant factor:

Theorem. *([32]) sat*$^*(n, \mathcal{D}_2) \geq (2\sqrt{2} - o(1))\sqrt{n}$.

. We remark that the bound of $\sqrt{n}$, proved in [41], is the result of an argument about the 'local structure' of a diamond-saturated family, and in fact this type of argument cannot get beyond $\sqrt{n}$. To get beyond the $\sqrt{n}$ barrier and achieve $2\sqrt{2n}$, we develop a more 'global' kind of argument which makes full use of the properties of minimal/maximal sets in a diamond-saturated family.

Most importantly, our proof explores in depth what it means for a diamond-saturated family to be of size $c\sqrt{n}$ for a constant c. Surprisingly, such a structure is very rich in properties and yet, as far as we can see, there is no indication that such a family cannot exist. This suggests that perhaps the induced saturation number for the diamond in not of linear growth.

Finally, the question for the antichain poset is perhaps the most natural of them since it can be rephrased in a purely set theoretical way: for a positive integer k we say that a family $\mathcal{F}$ of subsets of $[n] = \{1, \ldots, n\}$ is *k-antichain saturated* if $\mathcal{F}$ does not contain k pairwise incomparable sets, but for every set $X \notin \mathcal{F}$, the family $\mathcal{F} \cup \{X\}$ does contain k incomparable sets. Observe that, by Dilworth's theorem, $\text{sat}^*(n, \mathcal{A}_k)$ is the size of the smallest family that is maximal subject to being the union of $k - 1$ chains.

We call a chain of subsets of $[n]$ *full* if it has size $n + 1$. It is easy to see that a collection of $k - 1$ full chains that intersect only at $\emptyset$ and $[n]$ is a k-antichain saturated family. Thus, for n large enough, we certainly have $\text{sat}^*(n, \mathcal{A}_k) \leq (k - 1)(n - 1) + 2$.

Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] improved this upper bound slightly, showing that for $n \geq k \geq 4$, we have $\text{sat}^*(n, \mathcal{A}_k) \leq (n - 1)(k - 1) - (\frac{1}{2}\log_2 k + \frac{1}{2}\log_2 \log_2 k + c)$, for some absolute constant c.

In the other direction, they also showed that $\mathrm{sat}^*(n, \mathcal{A}_k) \geq 3n-1$ for $n \geq k \geq 4$. This immediately implies that for $n \geq 4$ we have $\mathrm{sat}^*(n, \mathcal{A}_4) = 3n-1$. They also showed that $\mathrm{sat}^*(n, \mathcal{A}_2) = n+1$ and $\mathrm{sat}^*(n, \mathcal{A}_3) = 2n$, and conjectured that $\mathrm{sat}^*(n, \mathcal{A}_k) = (k-1)n(1+o(1))$. Here $o(1)$ denotes a function that tends to 0 as n tends to infinity for each fixed k, in other words we are thinking of k as fixed and n growing. Later on, Martin, Smith and Walker [41] improved the lower bound by showing that for $k \geq 4$ and n large enough $\mathrm{sat}^*(n, \mathcal{A}_k) \geq (1 - \frac{1}{\log_2(k-1)})\frac{(k-1)n}{\log_2(k-1)}$.

In Section 2.4 we determine the exact value for $k=5$ and $k=6$.

Theorem. *([3]) Let $n \geq 5$, then sat*$^*(n, \mathcal{A}_5) = 4n-2$.

Theorem. *([3]) Let $n \geq 6$, then sat*$^*(n, \mathcal{A}_6) = 5n-5$.

2.2 The butterfly

2.2.1 A linear lower bound on $\mathbf{sat}^*(n, \mathcal{B})$

In this subsection we prove that any butterfly-saturated family has size at least $n+1$. The strategy is to look at singletons that are not in the family and associate in an injective manner to each one of them a butterfly, satisfying some maximality conditions, that is formed when the singleton is added to the family. This injective association will allow us to construct an explicit injection from the ground set $[n]$ to $\mathcal{F}$. Together with the observation that the empty set has to be in $\mathcal{F}$, we establish:

Theorem 2.1. *$sat^*(n, \mathcal{B}) \geq n+1$.*

Before we start the proof of this result, we make the following observation:

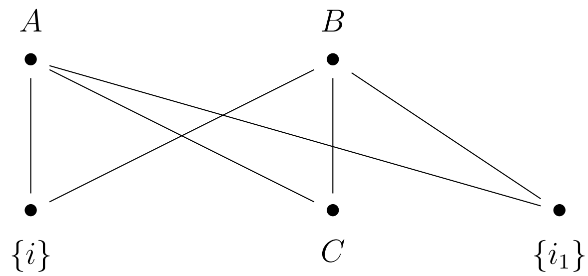
Lemma 2.2. *Let $\mathcal{F}$ be a $\mathcal{B}$-saturated family. If $\{i\}$ and $\{j\} \in \mathcal{F}$, then the pair $\{i, j\}$ is an element of $\mathcal{F}$.*

Proof. Assume $\{i, j\}$ is not an element of $\mathcal{F}$. Since $\mathcal{F}$ is $\mathcal{B}$-saturated, this implies that $\mathcal{F} \cup \{i, j\}$ contains a butterfly. That butterfly must involve the pair $\{i, j\}$, or else the initial family will contain a butterfly.
If $\{i, j\}$ is one of the maximal elements of the butterfly, then the two incomparable elements below it must be the singletons $\{i\}$ and $\{j\}$. But then they will be included in the other maximal element of the butterfly, call it M, and thus $\{i, j\} \subset M$, contradicting the incomparability of the maximal elements.
If $\{i, j\}$ is one of the minimal elements, then, by replacing $\{i, j\}$ in the newly formed butterfly with $\{i\}$ or $\{j\}$, we form a butterfly in $\mathcal{F}$, unless $\{i\}$ and $\{j\}$ are comparable to the other minimal element, call it N. Thus $\{i, j\} \subset N$, contradicting $N \parallel \{i, j\}$. $\square$

Note that if a butterfly-saturated family $\mathcal{F}$ contains $\Theta(n)$ singletons, then $\mathcal{F}$ contains $\Theta(n^2)$ pairs, so that $|\mathcal{F}| \geq \Theta(n^2)$.

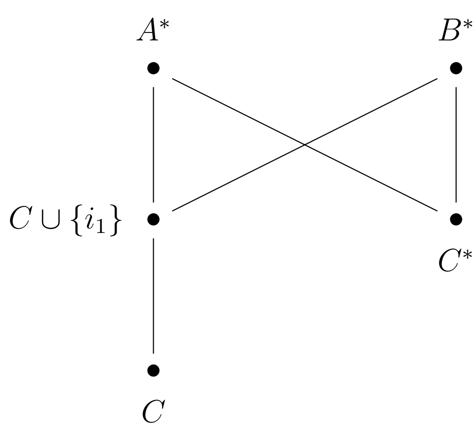
We define a *chevron* to be a triplet (A, B, C) of subsets of $[n]$ with the property that $C \subset A$, $C \subset B$ and $A \parallel B$.

Proof of Theorem 2.1. We will first assign to every $\{i\} \notin F$ a chevron with elements from $\mathcal{F}$ in such a way that no two singletons are assigned the same chevron.
If $\{i\} \notin F$, then $\mathcal{F} \cup \{i\}$ contains a butterfly and that butterfly has to involve the singleton, otherwise $\mathcal{F}$ would not be butterfly-free. Moreover, that singleton has to be one of the minimal elements of the butterfly since it does not have two incomparable elements below it. Therefore we have the structure shown below.



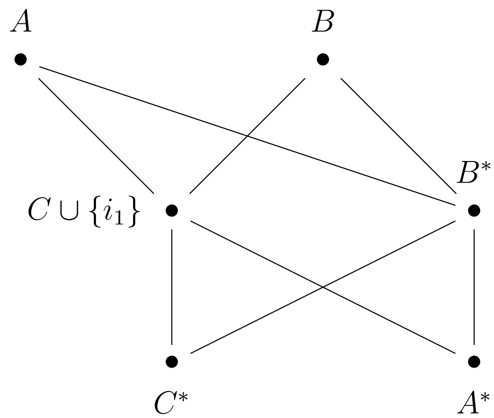
It is obvious that $i \notin C$. Among all these constructions, we pick the one having $|C|$ maximal and assign the chevron (A, B, C) to $\{i\}$. We now need to show that under this construction, a chevron is not assigned to two different singletons. Assume that $\{i_1\}$ has also been assigned the same (A, B, C) chevron, as shown above. Consider the set $C \cup \{i_1\}$. It is clearly incomparable to $\{i\}$ since $i \neq i_1$, it is contained in both A and B, but not equal to either of them since they contain i, thus $C \cup \{i_1\} \notin F$ by maximality of $|C|$. Therefore $C \cup \{i_1\}$ has to form a butterfly with three elements of $\mathcal{F}$.

1. Case 1: $C \cup \{i_1\}$ is one of the minimal elements of the butterfly as shown below, where $A^*, B^*, C^* \in \mathcal{F}$.



To stop A^*, B^*, C^*, C from forming a butterfly in $\mathcal{F}$, we need C and C^* to be comparable, and since $C \cup \{i_1\} \parallel C^*$, the only option is $C \subset C^*$ and $i_1 \notin C^*$. Now the chevron (A^*, B^*, C^*) has the property that $\{i_1\} \parallel C^*$, $i_1 \in A^*, B^*$ and the size of C^* is strictly greater than the size of C. By construction, this contradicts that (A, B, C) was assigned $\{i_1\}$.

2. Case 2: $C \cup \{i_1\}$ is one of the maximal elements of the butterfly as shown in the diagram below.



We obviously have $C \cup \{i_1\} \subset A, B$. To stop A (or B), C^*, A^* and B^* from forming a butterfly in $\mathcal{F}$, we need both A and B to be comparable to B^* and the only option is $B^* \subset A, B$. We notice that we are now in the previous case where $C \cup \{i_1\}$ is the minimal element of a butterfly, namely the one formed with A, B and B^*.

We therefore conclude that we can associate every singleton (not in our family) with a chevron, and no two singletons are associated with the same chevron.
The next step is to show that $C \cup \{i\} \in \mathcal{F}$ where C is the maximal element of the chevron assigned to the singleton $\{i\} \notin \mathcal{F}$. Assume that $C \cup \{i\} \notin \mathcal{F}$ and as before, it will have to form a butterfly with three elements of $\mathcal{F}$.

1. Case 1: $C \cup \{i\}$ is one of the minimal elements of the butterfly. Assume C^* is the other minimal element and A^*, B^* are the two maximal incomparable elements. The same argument we used above for $C \cup \{i_1\}$ will tell us that $C \subset C^*$, $i \notin C^*$ and $i \in A^* \cap B^*$, contradicting the maximality of $|C|$.

2. Case 2: $C \cup \{i\}$ is one of the maximal elements of the butterfly, B^* is the other maximal element and A^* and C^* are the two incomparable minimal elements. Since $C \cup \{i\} \subset A, B$, but is not equal to either of them (if for example $A = C \cup \{i\}$, then $A \subset B$ which cannot happen), the same arguments as above will tell us that $B^* \subset A, B$, which leads us back to the first case.

Therefore we indeed have $C \cup \{i\} \in \mathcal{F}$. Let C_i be the minimal element of the chevron assigned to the singleton $\{i\}$, for every $\{i\} \notin \mathcal{F}$.
Let us now define the following function from $[n]$ to elements of $\mathcal{F}$:

$$i \longmapsto \begin{cases} \{i\} & \text{if } \{i\} \in \mathcal{F} \\ C_i \cup \{i\} & \text{if } \{i\} \notin \mathcal{F}, \end{cases}$$

By what we just proved above, this is a well-defined function from $[n]$ to $\mathcal{F}$.
We claim that this function is an injection. Since $C_i \cup \{i\}$ cannot be a singleton and

the function is obviously injective on singletons, we only need to show that $C_{i_1} \cup \{i_1\} \neq C_i \cup \{i\}$ if $i \neq i_1$.
If $C_{i_1} \cup \{i_1\} = C_i \cup \{i\}$, then $C_{i_1} \neq C_i$ since $i \neq i_1$, but they have the same cardinality which tells us that they are incomparable. Let (A, B, C_i) be the chevron C_i is originating from. By construction we have that $C_i \cup \{i\} \subset A, B$. It would then follow that $C_{i_1} \subset A, B$ which immediately implies that A, B, C_i, C_{i_1} would form a butterfly in $\mathcal{F}$, contradiction.
Hence, the above function is indeed an injection from $[n]$ to non-empty elements of $\mathcal{F}$. It is easy to see that if we add the empty set to $\mathcal{F}$, it cannot form a butterfly as it is comparable to everything, thus by saturation $\emptyset \in \mathcal{F}$.
We therefor conclude that $|\mathcal{F}| \geq n + 1$ for every butterfly-saturated family, implying $\text{sat}^*(n, \mathcal{B}) \geq n + 1$, as claimed. □

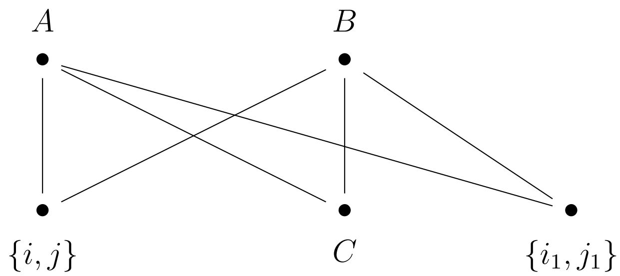
2.2.2 Further analysis of singletons

In the previous subsection we looked at singletons that are not in our family and constructed an injection from them to the set of chevrons with elements in $\mathcal{F}$. Because of the crucial role singletons played in the above proof, it is of interest to see what more can be said about the number of singletons in the family. In this section we use the same techniques and look at pairs that are not in our family. This provides us with better bounds in the case when we have few singletons in the family.

Theorem 2.3. *Let $\mathcal{F}$ be a $\mathcal{B}$-saturated family containing $k \geq 1$ singletons. Then $|\mathcal{F}| \geq \binom{k}{2} + k(n-k) + k + 1$.*

Proof. From Lemma 2.2 we already know that $\mathcal{F}$ contains the $\binom{k}{2}$ pairs made out of singletons of $\mathcal{F}$ only. We look at pairs $\{i, j\} \notin \mathcal{F}$. Also from Lemma 2.2 we get that at least one of the singletons in this pair is not in our family. We now restrict our attention to pairs $\{i, j\}$ that are not in $\mathcal{F}$ and for which exactly one of the singletons $\{i\}$ and $\{j\}$ is in $\mathcal{F}$. As in the previous subsection, we will show that we can uniquely assign a chevron to these pairs and construct an injection from the set of pairs containing at least one singleton of $\mathcal{F}$, to $\mathcal{F}$.

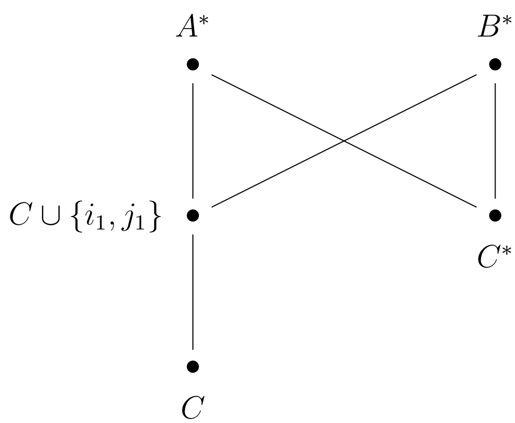
As before, if $\{i, j\} \notin \mathcal{F}$, then $\mathcal{F} \cup \{i, j\}$ contains a butterfly involving the pair $\{i, j\}$, and that this pair has to be one of the minimal elements. This is because if it is one of the maximal elements, the only candidates for the two incomparable minimal elements are the singletons made out of its elements. But we know that one must not be in our family and so this cannot happen. We then have a butterfly formed as shown below.



Among all these configurations, we choose the one having $|C|$ maximal and assign the chevron (A, B, C) to the pair $\{i, j\}$. We have to show that this assignment is an injection for the pairs we are considering.
Assume that under this construction, the same chevron has been assigned to a different pair $\{i_1, j_1\}$. We have the following two cases to analyse:

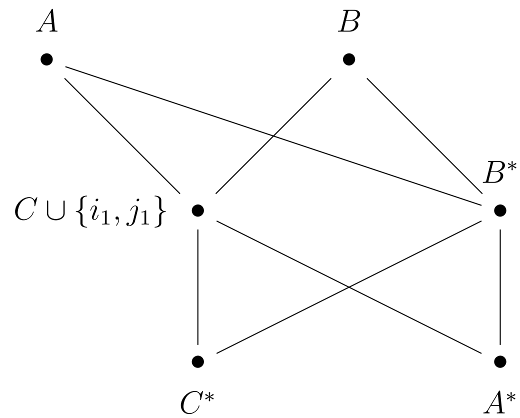
1. $\{i, j\} \parallel C \cup \{i_1, j_1\}$, then $C \cup \{i_1, j_1\} \notin \mathcal{F}$ by the maximality of the chevron assigned to $\{i, j\}$. This means that $C \cup \{i_1, j_1\}$ will form a butterfly in $\mathcal{F}$.

   (a) Case 1: $C \cup \{i_1, j_1\}$ is one of the minimal elements of the butterfly as shown in the diagram below, where $A^*, B^*, C^* \in \mathcal{F}$.

   But A^*, B^*, C^*, C have to not form a butterfly, therefore C and C^* are comparable and thus $C \subseteq C^*$. Because $C^* \parallel C \cup \{i_1, j_1\}$, $C \neq C^*$ and $\{i_1, j_1\} \parallel C^*$. Therefore the chevron (A^*, B^*, C^*) can be assigned to $\{i_1, j_1\}$ and the size of C^* is strictly greater than the size of C, which is a contradiction.

   (b) Case 2: $C \cup \{i_1, j_1\}$ is one of the maximal elements as shown below.

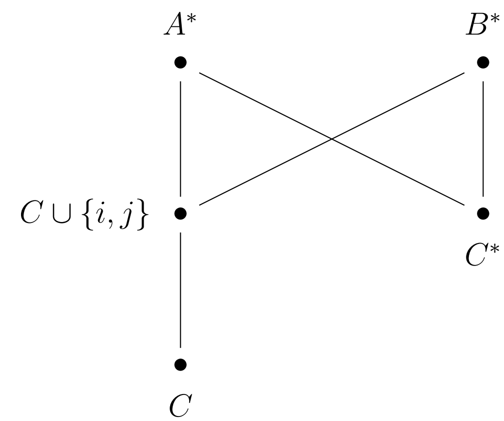


We obviously have $C \cup \{i_1, j_1\} \subset A, B$. To stop A (or B), C^*, A^* and B^* from forming a butterfly in $\mathcal{F}$, we need both A and B to be comparable to B^* and the only option is $B^* \subset A, B$. We notice that we are now in the previous case where $C \cup \{i_1, j_1\}$ is the minimal element of a butterfly, namely the one formed with A, B and B^*.

2. $\{i, j\}$ is comparable to $C \cup \{i_1, j_1\}$ and $\{i_1, j_1\}$ is comparable to $C \cup \{i, j\}$. By cardinality, the only options are $\{i, j\} \subset C \cup \{i_1, j_1\}$ and $\{i_1, j_1\} \subset C \cup \{i, j\}$. Because $\{i, j\} \parallel C$, we need $\{i, j\} \cap \{i_1, j_1\} \neq \emptyset$ and wlog $i = i_1$ and $j \neq j_1$. It then follows that $j, j_1 \in C$ and thus $i \notin C$. Because $i \notin C$, we cannot have $\{i\} \in \mathcal{F}$, otherwise $A, B, C, \{i\}$ will form a butterfly in $\mathcal{F}$. Thus, since the pairs we are analysing consist of exactly one singleton in $\mathcal{F}$, we obtain that $\{j\}$ and $\{j_1\}$ are elements of $\mathcal{F}$. But then we obtain $A, B, \{j\}, \{j_1\}$ a butterfly in $\mathcal{F}$, which is a contradiction.

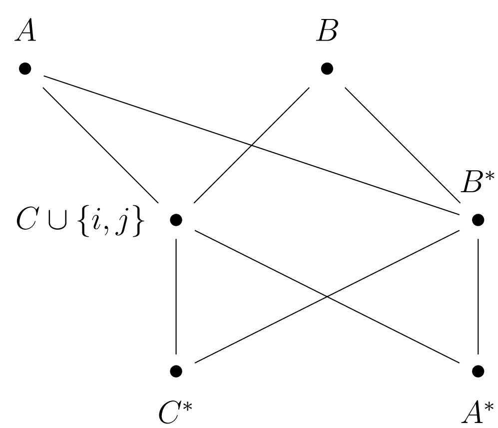
We will construct an injection from the set of pairs $\{i, j\}$ with at least one singleton in $\mathcal{F}$ to elements of $\mathcal{F}$. We know that we can assign a unique chevron to every such pair that is not in our butterfly-saturated family.
Let $\{i, j\}$ be such a pair and (A, B, C) its chevron. If $C \cup \{i, j\} \notin \mathcal{F}$, then it has to form a butterfly when added to the family.

1. Case 1: $C \cup \{i, j\}$ is one of the minimal elements of the butterfly.



As before, it then follows that $C \subset C^*$ and $\{i,j\} \parallel C^*$, and thus the chevron (A^*, B^*, C^*) could have been assigned to the pair, contradicting the maximality of the minimal element of the chevron.

2. Case 2: $C \cup \{i,j\}$ is one of the maximal elements of the butterfly. It then follows by the same arguments we used before, that $C \cup \{i,j\}$ will be one of the minimal elements in a butterfly containing A, B and B^* (as we can see on the diagram below), thus returning to the previous case.



Therefore we must have $C \cup \{i,j\} \in \mathcal{F}$. Let C_{ij} be the minimal element of the chevron assigned to the pair $\{i,j\}$ which contains exactly one singleton in the family. We define the following function from the set of pairs containing at least one singleton in the family to $\mathcal{F}$:

$$\{i,j\} \longmapsto \begin{cases} \{i,j\} & \text{if } \{i,j\} \in \mathcal{F} \\ C_{ij} \cup \{i,j\} & \text{if } \{i,j\} \notin \mathcal{F}. \end{cases}$$

By what we just proved above, this is a well-defined function from this set of pairs to $\mathcal{F}$.

Because in the second case $C_{ij} \parallel \{i,j\}$, $|C_{ij} \cup \{i,j\}| \geq 3$, and thus, in order to prove

injectivity, we need to show that $C_{i_1j_1} \cup \{i_1, j_1\} \neq C_{ij} \cup \{i, j\}$ for any two pairs $\{i, j\} \neq \{i_i, j_1\}$ with the desired property that are not in our family $\mathcal{F}$.
Assume that we do have $C_{i_1j_1} \cup \{i_1, j_1\} = C_{ij} \cup \{i, j\}$ for two different pairs.

1. Case 1: $|C_{ij}| = |C_{i_1j_1}|$. If $C_{ij} = C_{i_1j_1}$ and i is the element not in C_{ij}, then i is an element of the other pair. Now we have the equality $C_{ij} \cup \{i, j\} = C_{ij} \cup \{i, j_1\}$ with $j \neq j_1$. This immediately implies that $j, j_1 \in C_{ij}$, and if (A, B, C_{ij}) is the chevron assigned to $\{i, j\}$, then the same chevron can be assigned to $\{i, j_1\}$, contradicting the uniqueness of the chevrons for this type of pairs. This is because $\{i, j_1\} \subseteq C_{ij} \cup \{i, j\} \subset A, B$ and $C_{ij} \parallel \{i, j_1\}$ as $i \notin C_{ij}$. Therefore $C_{ij} \neq C_{i_1j_1}$ and since they have the same cardinality, $C_{ij} \parallel C_{i_1j_1}$. But if (A, B, C_{ij}) is the chevron corresponding to $\{i, j\}$, then $C_{i_1j_1} \subset C_{ij} \cup \{i, j\} \subseteq A, B$, thus $A, B, C_{ij}, C_{i_1j_1}$ forms a butterfly in $\mathcal{F}$, which is a contradiction.

2. Case 2: $|C_{ij}| \neq |C_{i_1j_1}|$ and wlog, $|C_{ij}| < |C_{i_1j_1}|$. Because adding the pair increases the size by at least one and at most two, we find that $|C_{i_1j_1}| = |C_{ij}| + 1$. This also means that $C_{ij} \cap \{i, j\} = \emptyset$ and $|C_{i_1j_1} \cap \{i_1, j_1\}| = 1$. If $C_{ij} \parallel C_{i_1j_1}$, then we form a butterfly in $\mathcal{F}$ with the chevron where $C_{i_1j_1}$ is coming from. If they are comparable, then $C_{ij} \subset C_{i_1j_1}$ and consequently $\{i, j\} \parallel C_{i_1j_1}$. Moreover, if $(\bar{A}, \bar{B}, C_{i_1j_1})$ is the chevron for $\{i_1, j_1\}$, then $\{i, j\} \subset C_{i_1j_1} \cup \{i_1, j_1\} \subseteq \bar{A}, \bar{B}$, and thus $\{i, j\}$ can be assigned the chevron $(\bar{A}, \bar{B}, C_{i_1j_1})$ and the size of $C_{i_1j_1}$ is strictly greater than the size of C_{ij}, which contradict the choice of the chevron.

Therefore our function is an injection from the set of pairs containing at least one singleton from the family, to elements of $\mathcal{F}$. We have exactly $\binom{k}{2} + k(n-k)$ such pairs, giving $|\mathcal{F}| \geq \binom{k}{2} + k(n-k)$. To finish the proof, we observe that all these elements of $\mathcal{F}$ have size at least 2, thus together with the k singletons and the empty set we obtain $|\mathcal{F}| \geq \binom{k}{2} + k(n-k) + k + 1$, as claimed □

Note that if the number of singletons in $\mathcal{F}$ is $\Theta(n^\alpha)$ for some $\alpha \in (0, 1)$, Theorem 2.3 gives us a better bound than both Lemma 2.2 and Theorem 2.1. Lemma 2.2 gives us $\Theta(n^{2\alpha})$ elements in $\mathcal{F}$ and Theorem 2.1 gives $n + 1$. On the other hand, Theorem 2.3 gives us $\Theta(n^{1+\alpha})$ elements in $\mathcal{F}$, which beats both of the previous bounds.

2.3 The diamond

2.3.1 The main result

The aim of this subsection is to prove the following:

Theorem 2.4. *For every $c < 2\sqrt{2}$ there exists an n_0 such that $sat^*(n, \mathcal{D}_2) \geq c\sqrt{n}$ for any $n \geq n_0$.*

Before we do that, we prove the following lemma, which is a special case of Lemma 9 in paper [41]. What is crucial about this lemma is that it shows the importance of minimal elements in a $\mathcal{D}_2$-saturated family $\mathcal{F}$. Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] proved that if $\mathcal{F}$ contains the empty set (or the full set $[n]$), then $|\mathcal{F}| > n$. This fact will be used repeatedly throughout the proof.

Lemma 2.5. *Let $\mathcal{F}$ be a $\mathcal{D}_2$-saturated family. Let S be a minimal element of $\mathcal{F}$. Then $|\mathcal{F}| \geq |S|$.*

Proof. If S is the empty set, then the statement is trivially true.
Now we assume $S \neq \emptyset$, and for each element i of S we will find an element of $\mathcal{F}$ that contains all elements of S except i. This will give us $|S|$ elements of $\mathcal{F}$, as desired.

More precisely, by the minimality of S we have that $S - \{i\} \notin \mathcal{F}$. Therefore, since $\mathcal{F}$ is diamond-saturated, $S - \{i\}$ will have to form a diamond when added to the family. We obtain three sets A, B and C of $\mathcal{F}$ such that they form a diamond together with $S - \{i\}$. By the minimality again we can only have $S - \{i\}$ the minimal element of the diamond. Let A be the maximal element of the diamond.

Suppose $B \neq S$ and $C \neq S$. If $i \in B$ and $i \in C$, then we observe that A, B, C and S form a diamond in $\mathcal{F}$, contradiction. Thus we can assume, without loss of generality, that $i \notin B$. So we have $S - \{i\} \subset B$ and $i \notin B$, as claimed.

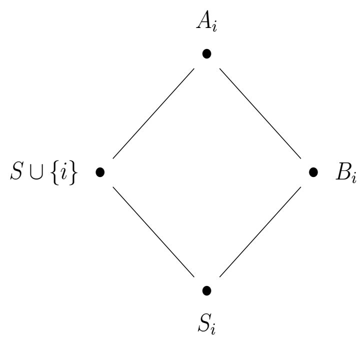
If on the other hand $C = S$, then B and S are incomparable and $S - \{i\} \subset B$. So we again obtain that $i \notin B$ and $S - \{i\} \subset B$, which finishes the proof. □

Proof of Theorem 2.4. Suppose for a contradiction that for some $c < 2\sqrt{2}$ we have $\text{sat}^*(n, \mathcal{D}_2) \leq c\sqrt{n}$ for some arbitrarily large n.

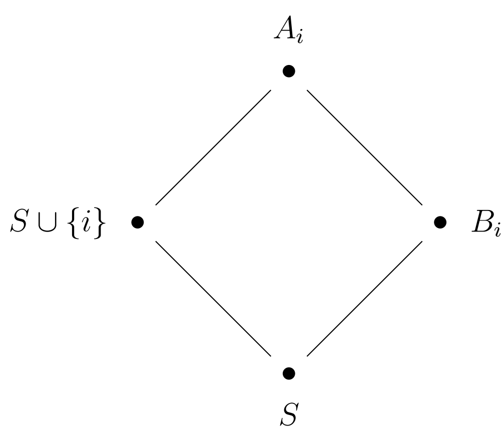
Fix $\mathcal{F}$ an arbitrary diamond-saturated family with cardinality at most $c\sqrt{n}$. This immediately implies that $\emptyset, [n] \notin \mathcal{F}$. Fix $S \in \mathcal{F}$ a minimal set with respect to inclusion. From Lemma 2.5 we know that $|S| \leq c\sqrt{n}$. Thus there exist $n - c\sqrt{n}$ singletons such that $i \notin S$. For those singletons, at least $n - 2c\sqrt{n}$ of the sets $S \cup \{i\}$ are not in $\mathcal{F}$, by our initial assumption on the size of the family.
A set $S \cup \{i\}$ that is not in our family must form a diamond with 3 elements of $\mathcal{F}$ by the saturation of $\mathcal{F}$. Assume $S \cup \{i\}, A, B, C$ form a diamond where $A, B, C \in \mathcal{F}$.

We observe that $S \cup \{i\}$ cannot be the maximal element of such a diamond for more than $c\sqrt{n}$ singletons. Indeed, for each singleton i for which $S \cup \{i\}$ is the maximal element of a diamond, let V_i be minimal among the minimal elements of such diamonds. We observe that each V_i is a minimal element of $\mathcal{F}$ and that $i \in V_i$ by the minimality of S ($V_i \neq S$ as they have different sizes). Moreover, $V_i = (S - K_i) \cup \{i\}$ for some $K_i \subseteq S$. This implies that $S \cup \{i\}$ is the maximal element of the diamond for at most $c\sqrt{n}$ singletons.

Also, $S \cup \{i\}$ cannot be the minimal element of the diamond because then A, B, C, S would form a diamond in $\mathcal{F}$, contradicting the fact that $\mathcal{F}$ is diamond free. Thus, $S \cup \{i\}$ has to be one of the two incomparable elements of the induced diamond for at least $n - 3c\sqrt{n}$ singletons. Therefore, for these singletons i we have the structure below, where $A_i, B_i, S_i \in \mathcal{F}$.



Moreover, we observe that by minimality either $S = S_i$ or $S_i - S = \{i\}$. If $S_i \neq S$ then there exists $K_i \subseteq S$ such that $S_i = S \cup \{i\} - K_i$, and all such S_i are pairwise different because S_i is the only one containing i. Therefore there are at least $n - 4c\sqrt{n}$ singletons i such that $S_i = S$. We will now focus on these singletons for which we have the following diamond, where B_i is of maximal cardinality with respect to this construction, and A_i is of minimal cardinality, after choosing B_i.



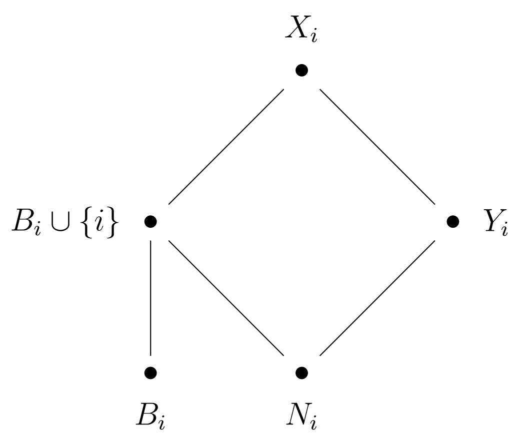
We observe that since B_i and $S \cup \{i\}$ are incomparable, but $S \subset B_i$, then we must have $i \notin B_i$.

Claim 1. *If $i \neq j$, then $B_i \cup \{i\} \neq B_j \cup \{j\}$.*

Proof. Suppose $B_i \cup \{i\} = B_j \cup \{j\}$. Since $i \notin B_i$ and $j \notin B_j$, we must have $i \in B_j$ and $j \in B_i$, which implies that B_i and B_j are incomparable. We now observe that

S, B_i, B_j, A_i form a diamond in $\mathcal{F}$, which is a contradiction. We can choose A_i to be the maximal element because $B_j \subset B_j \cup \{j\} = B_i \cup \{i\} \subseteq A_i$. □

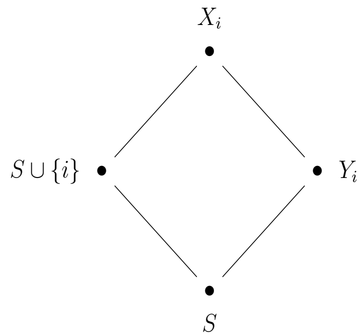
We deduce from Claim 1 and the assumption on the size of $\mathcal{F}$ that for at least $n - 5c\sqrt{n}$ singletons i, $B_i \cup \{i\}$ is not in the family. Therefore, by saturation, each element $B_i \cup \{i\}$ has to form a diamond with 3 different elements of $\mathcal{F}$. Let X_i, Y_i, N_i be three such elements. We notice that $B_i \cup \{i\}$ cannot be the minimal element because then B_i, X_i, Y_i, N_i would form a diamond in $\mathcal{F}$. It also cannot be the maximal element of the diamond because $B_i \cup \{i\} \subset A_i$, thus A_i, X_i, Y_i, N_i will again form a diamond in $\mathcal{F}$. We conclude that $B_i \cup \{i\}$ has to be one of the two incomparable elements as shown in the picture below. We choose N_i of minimal cardinality with respect to this configuration.



Claim 2. *B_i and N_i are incomparable.*

Proof. Suppose they are comparable. There are three cases:

1. $N_i \subset B_i$.
   In this case, if B_i and Y_i were incomparable, then we would have the diamond X_i, B_i, N_i, Y_i inside $\mathcal{F}$. Therefore B_i and Y_i are comparable and the only option is $B_i \subset Y_i$ as $B_i \cup \{i\} \parallel Y_i$. This also implies that $i \notin Y_i$. Finally we have $S \subset B_i \subset Y_i$, $S \cup \{i\} \subset B_i \cup \{i\} \subset X_i$, and $S \cup \{i\} \parallel Y_i$ since $i \notin Y_i$ and clearly $|S \cup \{i\}| \leq |B_i| < |Y_i|$. This means that we have the diamond below, which contradicts the maximality of B_i.



2. $N_i = B_i$.
   This case reduces to the previous case since now we already know $B_i \subset Y_i$.

3. $B_i \subset N_i$.
   This case is impossible by cardinality since $N_i \subset B_i \cup \{i\}$, thus $|B_i| < |N_i| < |B_i \cup \{i\}| = |B_i| + 1$, a contradiction.

□

An immediate consequence of Claim 2 is that S and N_i are incomparable. If they were not, then $S \subset N_i$ would form the diamond X_i, B_i, N_i, S in $\mathcal{F}$, contradiction. Lastly, S cannot be equal to N_i since $S \subset B_i$, but B_i and N_i are incomparable.
We also remark that by the minimality of N_i we have that any two distinct N_i are incomparable, and that $i \in N_i$. The second remark follows from the fact that $N_i \parallel B_i$, but $N_i \subset B_i \cup \{i\}$.

The following claim will be very useful for the construction and consequent modification of a certain bipartite graph at the end of the section.

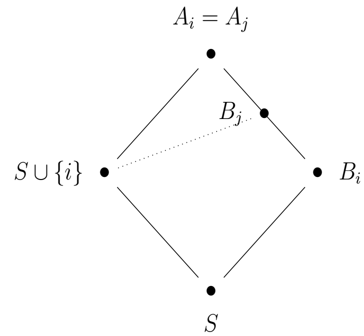
Claim 3. *If $i \neq j$ then we cannot have both $N_i = N_j$ and $B_i = B_j$.*

Proof. Suppose $N_i = N_j$. Then by previous remark we have that $i \in N_i$ and consequently $i \in N_j$. Also $N_j \subset B_j \cup \{j\}$, thus $i \in B_j$. On the other hand $i \notin B_i$, hence $B_i \neq B_j$ which finishes the claim. □

The next claim is not explicitly used in the proof, however it could be of potential interest towards proving more than the result in this paper, and it further illustrates how constraining and structurally rich is the property of being a minimal diamond-saturated family.

Claim 4. *If $B_i \neq B_j$, then $A_i \neq A_j$.*

Proof. Assume for a contradiction that $B_i \neq B_j$ and $A_i = A_j$. We cannot have $B_i \parallel B_j$ because otherwise $A_i = A_j, B_i, B_j, S$ would form a diamond in $\mathcal{F}$. Therefore, without loss of generality, we can assume $B_i \subset B_j$ and we have the following diagram.



If $S \cup \{i\}$ is not comparable to B_j, then the diamond formed by these two, $A_i = A_j$ and S would contradict the maximality of B_i. Hence we need to have $S \cup \{i\}$ and B_j comparable, and by cardinality (there is no set strictly between S and $S \cup \{i\}$), the only possibility is $S \cup \{i\} \subset B_j$.
But now we have the diamond formed by $S, B_i, S \cup \{i\}, B_j$ which contradicts the minimality of A_i with respect to B_i. $\square$

We have already seen previously that $B_i \cup \{i\} \neq B_j \cup \{j\}$ for $i \neq j$. Thus we have at least $n - 5c\sqrt{n}$ sets, $B_i \cup \{i\} \notin \mathcal{F}$. Each of them has a corresponding set N_i, although the N_i's can sometimes coincide for different i's. We build the following bipartite graph: the vertex set is $\mathcal{B} \sqcup \mathcal{N}$, where $\mathcal{B}$ consists of the sets $B_i \cup \{i\}$ and $\mathcal{N}$ consists of the corresponding sets N_i. The only edges are the ones joining $B_i \cup \{i\}$ to the corresponding N_i for each i. We observe that each vertex in $\mathcal{B}$ has degree 1, thus we have at least $n - 5c\sqrt{n}$ edges.

We now modify the graph by replacing $B_i \cup \{i\}$ with B_i for all i, and identifying the same repeating set with a single vertex – in other words, if $B_i = B_j$ for two $i \neq j$, the vertex $B_i \cup \{i\}$ and the vertex $B_j \cup \{j\}$ will both be identified with the vertex $B_i = B_j$. This new graph, which we call G, is bipartite and has vertex set $\mathcal{B}' \sqcup \mathcal{N}$, where $\mathcal{B}'$ consists of the sets B_i. Notice that no vertex in $\mathcal{B}'$ appears in $\mathcal{N}$ since all the B_i contain S, while all the N_i are incomparable to S.

If an edge were to contract, that would mean that for two different i and j we have $N_i = N_j$ and $B_i = B_j$, which contradicts Claim 3. Hence, the modified graph still has at least $n - 5c\sqrt{n}$ edges.

Assume that $|\mathcal{N}| = k$ and that d is the biggest degree in $\mathcal{N}$. Since G is bipartite we have that the number of edges is the sum of degrees in $\mathcal{N}$ which is less or equal to kd. Thus we have that $n - 5c\sqrt{n} \leq kd \Rightarrow d \geq \dfrac{n - 5c\sqrt{n}}{k}$. This also tells us that the size of $\mathcal{B}'$ is at least $d \geq \dfrac{n - 5c\sqrt{n}}{k}$. Moreover, we already have that $|\mathcal{F}| \geq |\mathcal{B}'| + |\mathcal{N}| \geq k + \dfrac{n - 5c\sqrt{n}}{k} \geq 2\sqrt{n - 5c\sqrt{n}}$.

Similarly, by looking at $\mathcal{G} = \{A : \bar{A} \in \mathcal{F}\}$, where $\bar{A}$ is the complement of A in $[n]$, we observe that this is also a diamond-saturated family of the same size as $\mathcal{F}$, where the minimal elements are the complements of the the maximal elements of $\mathcal{F}$. We can do the same analysis as above by fixing T a minimal element of $\mathcal{G}$, and construct a bipartite graph H analogously to the above graph G.

The bipartite graph H has vertex set $\mathcal{C}' \sqcup \mathcal{M}$, where $\mathcal{M}$ consists of the minimal elements, denoted by M_i (equivalent to the N_j), and $\mathcal{C}'$ consists of the elements that contain T, denoted by C_k (equivalent to the B_l). Therefore $\bar{M}_i \in \mathcal{F}$ are maximal elements in $\mathcal{F}$. On the other hand, any N_j is not a maximal element as it is contained in Y_j by construction. Similarly, any B_l is not a maximal element as it is contained in A_l. We conclude that no $\bar{M}_i$ can be equal to any N_j or to any B_l. Moreover, B_l is neither a maximal nor a minimal element in $\mathcal{F}$ since it is between S and A_i, thus no $\bar{C}_k$ is a minimal or a maximal element either, which implies that no $\bar{C}_k$ is equal to any N_j.

Let $|\mathcal{M}| = t$. By the same argument as above, now applied to the graph H, we have that $|\mathcal{C}'| \geq \dfrac{n-5c\sqrt{n}}{t}$. Let $\mathcal{M}^c = \{\bar{A} : A \in \mathcal{M}\}$ and $\mathcal{C}'^c = \{\bar{A} : A \in \mathcal{C}'\}$, thus $\mathcal{M}^c$ and $\mathcal{C}'^c$ are subsets of $\mathcal{F}$. As observed above $\mathcal{N} \cap \mathcal{M}^c = \emptyset$, $\mathcal{N} \cap \mathcal{C}'^c = \emptyset$ and $\mathcal{M}^c \cap \mathcal{B}' = \emptyset$. Therefore we have that $|\mathcal{F}| \geq |\mathcal{N}| + |\mathcal{M}| + |\mathcal{B}' \cup \mathcal{C}'^c|$. Assume without loss of generality that $t \geq k$. We then have that $|\mathcal{F}| \geq k + t + |\mathcal{B}'| \geq 2k + \dfrac{n-5c\sqrt{n}}{k} \geq 2\sqrt{2(n-5c\sqrt{n})}$, which is greater than $c\sqrt{n}$ for n large enough, a contradiction. □

The above proof leads to a natural question, namely how disjoint can $\mathcal{B}'$ and $\mathcal{C}'^c$ be? Suppose the family $\mathcal{F}$ contains a minimal element P and a maximal element R such that P and R are not comparable. It turns out that in that case $\mathcal{B}'$ and $\mathcal{C}'^c$ can be disjoint, thus giving $|\mathcal{F}| \geq |\mathcal{N}| + |\mathcal{B}'| + |\mathcal{M}^c| + |\mathcal{C}'^c| \geq k + \dfrac{n-5c\sqrt{n}}{k} + t + \dfrac{n-5c\sqrt{n}}{t} \geq 4\sqrt{n-5c\sqrt{n}}$. Indeed, we run the above argument once for $\mathcal{F}$ with fixed minimal element P and once for $\mathcal{G}$ with fixed minimal element $\bar{R}$. Now we notice that if $B_l = \bar{C}_k$ for some l and k, then $S \subset B_l = \bar{C}_k \subset \bar{\bar{R}} = R$, a contradiction.

However, whether such minimal and maximal sets exist in any diamond-saturated family (without $\emptyset$ and $[n]$) seems to be a non-trivial question. Using the above notation, the following proposition guarantees the existence of P and R under some mild assumptions.

Proposition 2.6. *Suppose there is no $i \notin S$ such that $S \cup \{i\} \notin \mathcal{F}$, $S = S_i$ and $B_i \cup \{i\} \in \mathcal{F}$. Then there exists a minimal element P and a maximal element R in $\mathcal{F}$ such that P and R are not comparable.*

Proof. We begin by noticing that all the N_i and S are minimal elements in $\mathcal{F}$ and $i \in N_i$ for every i such that $i \notin S$ and $S - K_i \cup \{i\} \notin \mathcal{F}$ for any $K_i \subseteq S$. For the singletons $i \notin S$ such that $S - K_i \cup \{i\} \in \mathcal{F}$ for some $K_i \subseteq S$, we consider $L_i \in \mathcal{F}$ to be the minimal element such that $L_i \subseteq S - K_i \cup \{i\}$. Because S is itself a minimal element, we need to have $i \in L_i$.

If every maximal element is comparable to every minimal element, then any maximal element must contain the union of all N_i, L_i and S which is $[n] - W$, where $W = \{i :$

$i \notin S, S \cup \{i\} \in \mathcal{F}\}$.
Suppose that $|W| \geq 2$. If $S \cup \{i\}$ and $S \cup \{j\}$ are in $\mathcal{F}$ for $i \neq j$ and $i, j \notin S$, then we cannot have a maximal element $T \in \mathcal{F}$ that contains the pair $\{i, j\}$ since by assumption $S \subset T$ and $S, S \cup \{i\}, S \cup \{j\}$ and T will form a diamond in $\mathcal{F}$. This implies that no element of $\mathcal{F}$ contains both i and j (as every element is included in a maximal element), in particular $\{i, j\} \notin \mathcal{F}$. Thus $\{i, j\}$ will have to form a diamond with 3 elements of $\mathcal{F}$ by saturation. However, since the empty set is not in the family, $\{i, j\}$ cannot be the maximal element of the diamond, therefore it will be a subset of one of the three sets it form a diamond with, contradiction.
We conclude that $|W| \leq 1$, thus the union of all the minimal sets is either $[n]$ or $[n] - \{i\}$ for some $i \notin S$ and $S \cup \{i\} \in \mathcal{F}$. If the latter is true, then take D to be the maximal element of $\mathcal{F}$ that contains $S \cup \{i\}$. D will have to contain $[n] - \{i\}$ too by our assumption, thus $D = [n]$. We see that in both cases we have to have $[n] \in \mathcal{F}$, which is a contradiction. □

2.3.2 Could $\mathbf{sat}^*(n, \mathcal{D}_2)$ be $O(\sqrt{n})$?

The above proof explores the extremal behaviour of a diamond-saturated family of size $O(\sqrt{n})$. The square root bound appears to push the minimal and maximal elements closer together, yet spread through most of the layers of the hypercube – note that this is quite unlike the two canonical examples of diamond-saturated families, namely a chain of size $n + 1$, and the family of all singletons and the empty set. Indeed, consider the graphs constructed towards the end of the proof. They show that, under the condition that the diamond-saturated family $\mathcal{F}$ is of square root order, $\mathcal{F}$ must be in a way invariant under taking complements – for example, the antichains formed by the minimal and maximal elements have to roughly look the same and be of $\sqrt{n}$ order.

It is clear from the proof that if the induced saturation number for the diamond is $\Theta(\sqrt{n})$, then the size of the biggest antichain of a family of this size has to be of $\sqrt{n}$ order. This can be seen by looking at the bipartite graph considered in the proof and the fact that one side, namely the N_i's, is an antichain of size k. Indeed, we have that the number of edges is equal to the sum of the degrees of the N_i, thus $n - 5c\sqrt{n} \leq k \times$ max degree. Because for each i, $i \in N_i$, we have $\deg(N_i) \leq |N_i| \leq |\mathcal{F}| \leq c\sqrt{n}$, where the middle inequality comes from the fact that the N_i are minimal elements. Therefore the maximum degree is at most $c\sqrt{n}$, hence $k \geq \dfrac{\sqrt{n}}{c} - 5$.

By applying Dilworth's theorem, we get that $\mathcal{F}$ can be decomposed into roughly $\sqrt{n}$ chains. Since $|\mathcal{F}| = O(\sqrt{n})$, this suggests a family of $c'\sqrt{n}$ disjoint chains, each of constant size and, more importantly, positioned in such a way that there are no common interior gaps for all of them. We believe that this is possible and that the above proof has some of the key clues to construct such a diamond-saturated family. More precisely, we conjecture the following.

Conjecture 2.7. *$sat^*(n, \mathcal{D}_2) = \Theta(\sqrt{n})$. Moreover, there exists a constant c such that* *for n large enough there exists a diamond-saturated family $\mathcal{F}$ consisting of $\sqrt{n}$ chains each of size c, with the property that if $C_i \subseteq A \subseteq B_i$, where C_i and B_i are elements of the i^{th} chain for every i, then $A \in \mathcal{F}$.*

Led on by the above analysis, we can also ask the following question.

Question 2.8. *Let $\mathcal{F}$ be a diamond-saturated family that does not contain $\emptyset$ or $[n]$. Can all the minimal elements of $\mathcal{F}$ be subsets of all the maximal elements of $\mathcal{F}$?*

2.4 Small antichains

In this section we establish the exct saturation number for the 5-antichain and the 6-antichain. We show that $\mathrm{sat}^*(n, \mathcal{A}_5) = 4n - 2$ and that $\mathrm{sat}^*(n, \mathcal{A}_6) = 5n - 5$.

We start by recording two immediate observations that will be used several times. The first is that any k-antichain saturated family must contain $\emptyset$ and $[n]$. The second is the following.

Lemma 2.9. *If $\mathcal{F}$ is an induced k-antichain saturated family, then $\mathcal{F}$ is the union of $k-1$ full chains. In particular, $\mathcal{F}$ must contain at least one element from each layer (a set of size i for every $0 \leq i \leq n$).*

Proof. By Dilworth's theorem, we may partition $\mathcal{F}$ into $k-1$ chains, and so $\mathcal{F}$ is certainly contained in the union of some $k-1$ full chains, say $\mathcal{D}_1, \ldots, \mathcal{D}_{k-1}$. But $\mathcal{D}_1 \cup \ldots \cup \mathcal{D}_{k-1}$ is a k-antichain saturated family, so by maximality of $\mathcal{F}$ we must have that $\mathcal{F} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_{k-1}$. □

2.4.1 5-antichain saturation

Theorem 2.10. *For any positive integer $n \geq 5$ we have sat$^*(n, \mathcal{A}_5) = 4n - 2$.*

Proof. Let $\mathcal{F}$ be an induced 5-antichain saturated family. By Lemma 2.9 we can cover $\mathcal{F}$ with 4 full chains $\mathcal{D}_1, \ldots, \mathcal{D}_4$. For each $i \in \{1, \ldots, n-1\}$ let $\mathcal{F}_i$ be the collection of sets in $\mathcal{F}$ of size i, and $x_i = |\mathcal{F}_i|$. We will now examine the following 4 cases:

Case 1. There exists $i \in \{1, \ldots, n-1\}$ such that $x_i = 1$.
Let A be the unique set in $\mathcal{F}$ of size i. Since each of the chains $\mathcal{D}_1, \ldots \mathcal{D}_4$ is a full chain, it follows that all of them must contain A. Consider the sets of size $i-1$ and $i+1$ in $\mathcal{D}_1$. They must be of the form $A \setminus \{x\}$ and $A \cup \{y\}$ respectively, for some $x \in A$ and $y \in [n] \setminus A$. Let $A' = A \setminus \{x\} \cup \{y\}$. Since $A' \neq A$ and $|A'| = i$, $A' \notin \mathcal{F}$. On the other hand, by setting $\mathcal{D}_1' = \mathcal{D}_1 \setminus \{A\} \cup \{A'\}$, we observe that the chains $\mathcal{D}_1', \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$ cover $\mathcal{F} \cup \{A'\}$ (note that A is still covered by $\mathcal{D}_2$). This implies that $\mathcal{F} \cup \{A'\}$ is 5-antichain free, contradicting the fact that $\mathcal{F}$ is 5-antichain saturated.

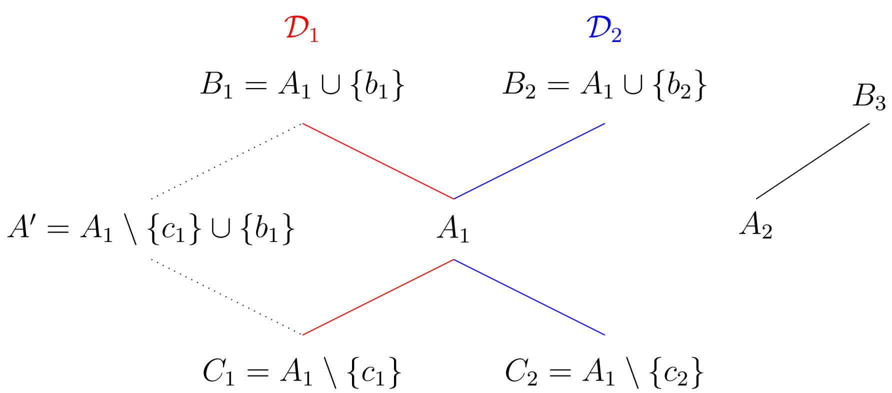
Case 2. There is no j such that $x_j = 1$, but there exists i such that $x_i = 2$.
Since $\mathrm{sat}^*(n, \mathcal{A}_5) \geq 3n - 1$ we get that $|\mathcal{F}| \geq 3n - 1$, thus there must be some $l \in \{1, \ldots, n-1\}$ for which $x_l \geq 3$. Combining this with the fact that there exist i such that $x_i = 2$ and $x_m \neq 1$ for all $1 \leq m \leq n-1$, we deduce that there exists some index $1 \leq j \leq n-1$ such that $x_j = 2$ and $x_{j+1} \geq 3$, or $x_j = 2$ and $x_{j-1} \geq 3$. Since a family is antichain-saturated if and only if the family of the complements of its sets is antichain-saturated, we can assume without loss of generality that there exists j such that $x_j = 2$ and $x_{j+1} \geq 3$. Let A_1 and A_2 be the two sets of size j. Since the 4 chains $\mathcal{D}_1, \ldots, \mathcal{D}_4$ that cover $\mathcal{F}$ are full, they have to go through A_1 and A_2 as well as cover the sets of size $j+1$. This implies that at least two chains with different sets of size

$j+1$ have the same element of size j. Thus we can assume without loss of generality that these chains are $\mathcal{D}_1$ and $\mathcal{D}_2$, and $A_1 \in \mathcal{D}_1, \mathcal{D}_2$. Let also B_1 and B_2 be the two (distinct) sets of size $j+1$ in these two chains respectively. Let B_3 be another set of size $j+1$ and assume without loss of generality that it is part of $\mathcal{D}_3$. We either have $A_2 \in \mathcal{D}_3$, or $A_1 \in \mathcal{D}_3$ which implies $A_2 \in \mathcal{D}_4$. As $\mathcal{D}_4$ must contain an element of size $j+1$, we can assume, after relabelling if necessary, that $A_1 \subset B_1, B_2$, and $A_2 \subset B_3$, and $A_1, B_1 \in \mathcal{D}_1$, and $A_1, B_2 \in \mathcal{D}_2$, and $A_2, B_3 \in \mathcal{D}_3$. Moreover, since $j \neq 0$, there exist sets $C_1, C_2 \subseteq A_1$ of size $j-1$ that are part of the chains $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Note that C_1 may be equal to C_2. Hence we can write

$$C_1 \cup \{c_1\} = A_1 = B_1 \setminus \{b_1\} \text{ and } C_2 \cup \{c_2\} = A_1 = B_2 \setminus \{b_2\},$$

where $b_1 \neq b_2 \in [n] \setminus A_1$ and $c_1, c_2 \in A_1$. Let $A' = A_1 \setminus \{c_1\} \cup \{b_1\}$ and $A'' = A_1 \setminus \{c_2\} \cup \{b_2\}$. If $A' \notin \mathcal{F}$, then by modifying $\mathcal{D}_1$ by replacing A_1 with A' we obtain a cover of $\mathcal{F} \cup \{A'\}$ with 4 chains, contradicting the fact that $\mathcal{F}$ is 5-antichain saturated. Thus $A' \in \mathcal{F}$, and similarly, $A'' \in \mathcal{F}$ too. Moreover, by construction, $|A'| = |A''| = j$ and $A' \neq A_1 \neq A''$. Because $\mathcal{F}$ contains exactly 2 sets of size j, we must have that $A' = A_2 = A''$. However A' contains b_1, while A'' does not, a contradiction.

The picture below summarises the above analysis.



Case 3. For all $i \in \{1, \dots, n-1\}$, $x_i = 3$.
We will show that this implies that $\mathcal{F}$ can be covered by 3 chains, contradicting the 5-saturation property of $\mathcal{F}$.

We start with the 4 full chains $\mathcal{D}_1, \dots, \mathcal{D}_4$ that cover $\mathcal{F}$. By modifying them if necessary, we can choose them in such a way that two of them coincide. Equivalently, we prove that for each $i \in \{0, \dots, n\}$, two of these chains can be chosen to coincide on sets of size less than or equal to i. We proceed by induction on i.

Clearly for $i = 0$ all of $\mathcal{D}_j$ start with the empty set, so they all coincide on sets of size at most 0. For $i = 1$ we have three different options for the sets of size 1 and 4 chains, so two chains must coincide on sets of size at most 1.

Let now $i > 1$ and assume that we can cover $\mathcal{F}$ by 4 full chains, $\mathcal{D}_1^i, \mathcal{D}_2^i, \mathcal{D}_3^i, \mathcal{D}_4^i$, two of which coincide on sets of size less than i. Without loss of generality, $\mathcal{D}_1^i$ and $\mathcal{D}_2^i$

coincide on sets of size less than i. If they coincide on sets of size i, we are done. Thus we now assume that they do not, and let A_1 be the set of size i in $\mathcal{D}_1^i$ and A_2 the set of size i in $\mathcal{D}_2^i$. Let also A_3 be the third set of size i.

If $\mathcal{D}_3^i$ contains A_1, then by replacing the sets of size not more than i in the chain $\mathcal{D}_1^i$ with the sets of size not more than i in $\mathcal{D}_3^i$, we obtain a cover of $\mathcal{F}$ by 4 chains, two of which coincide on all sets of size less than or equal to i, so we are done. Similarly we are done if any of $A_2 \in \mathcal{D}_3^i$, $A_1 \in \mathcal{D}_4^i$ or $A_2 \in \mathcal{D}_4^i$ holds. Therefore we may assume that $A_3 \in \mathcal{D}_3^i, \mathcal{D}_4^i$.

Let B be the set of size $i-1$ in chains $\mathcal{D}_1^i$ and $\mathcal{D}_2^i$. Then A_1 must be of the form $B \cup \{x\}$ for some $x \in [n] \setminus B$. Similarly, $A_2 = B \cup \{y\}$ for some $y \in [n] \setminus B$. We observe that $x \neq y$ as $A_1 \neq A_2$. For any $b \in B$, let $X_b = B \cup \{x\} \setminus \{b\}$ and $Y_b = B \cup \{y\} \setminus \{b\}$. We observe that the family $\mathcal{S} = \{X_b, Y_b : b \in B\}$ has size $2|B| = 2(i-1)$ since the X's are pairwise distinct, the Y's are pairwise distinct, and $X_b \neq Y_{b'}$ for any $b, b' \in B$ (as one set contains x, but the other does not). Moreover, all sets in $\mathcal{S}$ have size $i-1$ and $B \notin \mathcal{S}$.

If $i \geq 3$, then $2(i-1) \geq 4 > 2$, and since there are exactly 2 sets of size $i-1$ in $\mathcal{F}$ that are not equal to B, at least one of the sets in $\mathcal{S}$ is not in $\mathcal{F}$. Without loss of generality, assume $X_b \notin \mathcal{F}$ for some $b \in B$. However, by removing all sets of size less than i from $\mathcal{D}_1^i$ and adding X_b to it, we obtain a 4-chain cover of $\mathcal{F} \cup \{X_b\}$, which contradicts the fact that $\mathcal{F}$ is 5-antichain saturated.

If $i = 2$, then $B = \{b\}$ for some $b \in [n]$, and so $A_1 = \{b, x\}$, $A_2 = \{b, y\}$, $X_b = \{x\}$ and $Y_b = \{y\}$. As in the above case, if $\{x\} \notin \mathcal{F}$ or $\{y\} \notin \mathcal{F}$ we obtain a contradiction. Thus we must have $\{x\}, \{y\} \in \mathcal{F}$. Without loss of generality we can assume that $\{x\} \in \mathcal{D}_3$ and $\{y\} \in \mathcal{D}_4$. As argued previously, we must have $A_3 \in \mathcal{D}_3^2, \mathcal{D}_4^2$, which immediately implies that $A_3 = \{x, y\}$. Now we modify the chains as follows: we set $\mathcal{D}_1^3 = \mathcal{D}_1^2$, $\mathcal{D}_3^3 = \mathcal{D}_3^2$, $\mathcal{D}_2^3 = \mathcal{D}_2^2 \setminus \{\{b\}\} \cup \{\{y\}\}$ and $\mathcal{D}_4^3 = \mathcal{D}_4^2 \setminus \{\{y\}\} \cup \{\{x\}\}$. This forms a cover of $\mathcal{F}$ by 4 full chains such that D_3^3 and D_4^3 coincide on all sets of size not greater than 2. Thus the induction step is complete.

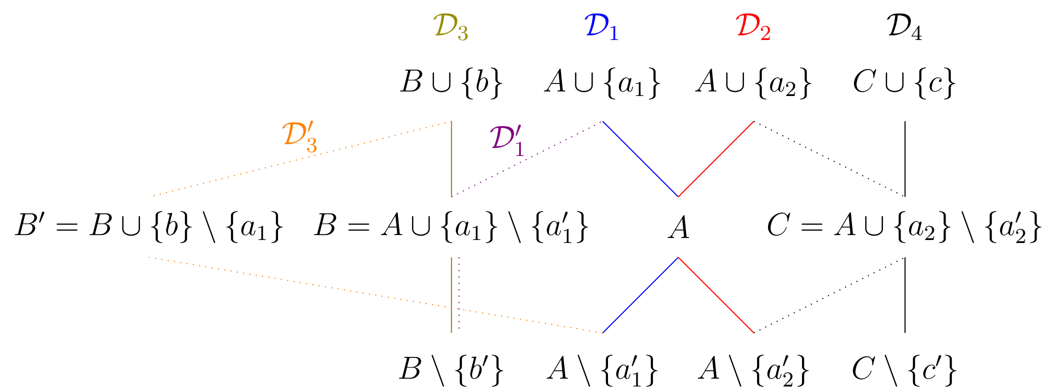
Case 4. There exist $j, t \in \{1, \ldots, n-1\}$ such that $x_j = 3$ and $x_t = 4$.
We know that no x_i is equal to 1 or 2 for $i \in \{1, \ldots n-1\}$, thus there must exist an index l such that $x_l = 3$ and $x_{l+1} = 4$, or $x_l = 3$ and $x_{l-1} = 4$. As in previous cases, we can assume without loss of generality that there exists l such that $x_l = 3$ and $x_{l+1} = 4$. Let A, B and C be the sets of size l in $\mathcal{F}$. Since $\mathcal{F}$ is covered by the 4 full chains $\mathcal{D}_1, \ldots, \mathcal{D}_4$, these 4 chains have to go through the 4 distinct sets of size $l+1$ in $\mathcal{F}$. Moreover, since there are exactly 3 sets of size l, we must have that two chains go through the same set of size l, while the other two chains go through the remaining sets of size l. Putting this together, we can assume without loss of generality that $A \in \mathcal{D}_1, \mathcal{D}_2$, $B \in \mathcal{D}_3$ and $C \in \mathcal{D}_4$. Furthermore, the sets of size $l+1$ are of the form $A \cup \{a_1\} \in \mathcal{D}_1$, $A \cup \{a_2\} \in \mathcal{D}_2$, $B \cup \{b\} \in \mathcal{D}_3$ and $C \cup \{c\} \in \mathcal{D}_4$, where $a_1, a_2 \in [n] \setminus A$ and $a_1 \neq a_2$, $b \in [n] \setminus B$ and $c \in [n] \setminus C$.

We now consider the sets of size $l-1$ corresponding to these chains. They must be of the form $A \setminus \{a_1'\} \in \mathcal{D}_1$, $A \setminus \{a_2'\} \in \mathcal{D}_2$, $B \setminus \{b'\} \in \mathcal{D}_3$ and $C \setminus \{c'\} \in \mathcal{D}_4$, where $a_1', a_2' \in A$, $b' \in B$ and $c' \in C$. We note that these sets need not be distinct.

Let $A' = A \setminus \{a_1'\} \cup \{a_1\}$ and $A'' = A \setminus \{a_2'\} \cup \{a_2\}$. It is clear that $A \neq A'$, $A \neq A''$ and $A' \neq A''$, thus A, A', A'' are 3 distinct sets of size l. If $A' \notin \mathcal{F}$, then by replacing A with A' in the chain $\mathcal{D}_1$ we obtain a cover of $\mathcal{F} \cup \{A'\}$ by 4 chains, which contradicts the fact that $\mathcal{F}$ is 5-antichain saturated. Thus we must have $A' \in \mathcal{F}$, and since it has size l, $A' = B$ or $A' = C$. Similarly we get that $A'' \in \mathcal{F}$. Therefore, the 3 sets of size l in our family are A, A' and A'', and we assume without loss of generality that $B = A'$ and $C = A''$.

Let $B' = B \setminus \{a_1\} \cup \{b\}$. It is clear that $B \neq B'$. If $B' \notin \mathcal{F}$, then by leaving the chains $\mathcal{D}_2$ and $\mathcal{D}_4$ unchanged, swapping the sets of size less than l between the chains $\mathcal{D}_1$ and $\mathcal{D}_3$, then replacing A with B' in chain $\mathcal{D}_3$, and A with B in chain $\mathcal{D}_1$, we obtain a cover of $\mathcal{F} \cup \{B'\}$ with 4 full chains. This implies that $\mathcal{F} \cup \{B'\}$ is still 5-antichain free, a contradiction. Hence $B' \in \mathcal{F}$ and thus it has to be equal to either A or A''.

The picture below illustrates the cover of $\mathcal{F}$ by the modified 4 chains: $\mathcal{D}_1', \mathcal{D}_2, \mathcal{D}_3', \mathcal{D}_4$.



We now examine the two cases:

(a) If $B' = A$, then $A = (A \setminus \{a_1'\} \cup \{a_1\}) \setminus \{a_1\} \cup \{b\} = A \setminus \{a_1'\} \cup \{b\}$, which implies that $a_1' = b$. It then follows that $B \cup \{b\} = (A \setminus \{a_1'\} \cup \{a_1\}) \cup \{a_1'\} = A \cup \{a_1\}$. This contradicts the original assumption that these 4 sets of size $l+1$ are distinct.

(b) If $B' = C$, let $C' = C \setminus \{a_2\} \cup \{c\}$. By the same reasoning as above $C' \in \mathcal{F}$ and $C' \neq A$, thus we must have $C' = B$.
From $B' = C$ we get that $(A \setminus \{a_1'\} \cup \{a_1\}) \setminus \{a_1\} \cup \{b\} = A \setminus \{a_2'\} \cup \{a_2\}$, which implies that $b = a_2$ and $a_1' = a_2'$. Similarly, from $C' = B$ we get that $c = a_1$. This implies that $B \cup \{b\} = (A \setminus \{a_1'\} \cup \{a_1\}) \cup \{a_2\} = (A \setminus \{a_1'\} \cup \{a_2\}) \cup \{a_1\} = C \cup \{c\}$, which contradicts the assumption that there are 4 sets of size $l+1$.

We conclude that none of the 4 cases analysed above is possible, thus we deduce that $x_i = 4$ for all $i \in \{1, \dots, n-1\}$. We already know that $x_0 = x_n = 1$, thus $|\mathcal{F}| \geq 4n-2$. This implies that $\text{sat}^*(n, \mathcal{A}_5) \geq 4n-2$ for $n \geq 5$. On the other hand, a family of 4 full chains that only intersect at $\emptyset$ and $[n]$ is 5-antichain saturated and has size $4n-2$, thus $\text{sat}^*(n, \mathcal{A}_5) \leq 4n-2$, which finishes the proof. $\square$

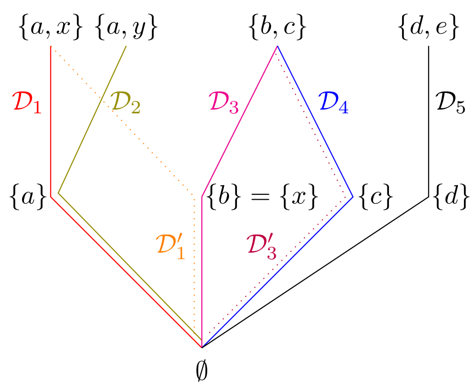
2.4.2 6-antichain saturation

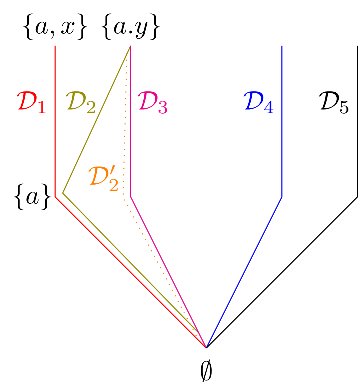
The proof presented in this section is very similar to the proof of Theorem 2.10. We therefore focus only on the parts that are specific to the 6-antichain and, where necessary, direct the reader to the analogous parts in the previous proof.

Theorem 2.11. *For every positive integer $n \geq 6$ we have sat*$(n, \mathcal{A}_6) = 5n - 5$.*

Proof. Let $\mathcal{F}$ be an induced 6-antichain saturated family of subsets of $[n]$. By Lemma 2.9, we can cover $\mathcal{F}$ with 5 full chains $\mathcal{D}_1, \ldots, \mathcal{D}_5$. Let $x_0, \ldots, x_n$ be the numbers of sets of sizes $0, \ldots, n$ respectively in $\mathcal{F}$. In the same way as in the proof of Theorem 2.10, we deduce that we cannot have $x_i \in \{1, 2, 3\}$ for any $i \in \{1, \ldots, n-1\}$.

The case when $x_i = 4$ for all $i \in \{1, \ldots, n-1\}$ is completely analogous to Case 3 in the proof of Theorem 2.10, except for the base case $i = 2$ of the induction. More precisely, we need to show that if the 5 full chains cover $\mathcal{F}$ and two of them agree on sets of size at most 1, then we can modify them in such a way that they still cover $\mathcal{F}$ (and are full chains) and two of them coincide on sets of size at most 2. The figures below are the two situations where we need to modify the chains. The colour coded figures are enough to show that this is possible. For the left figure we note that it is easy to show, and the same argument has been done in the previous section, that $\{x\}$ and $\{y\}$ are in $\mathcal{F}$, thus one of them is in $\mathcal{D}_3$ or $\mathcal{D}_4$. Without loss of generality we assume $\{x\} \in \mathcal{D}_3$.





Finally, suppose that there exist an index i such that $x_i = 4$ and an index j such that $x_j = 5$. Since all x_k are either 4 or 5 for $0 < k < n$, there exists some $l \in \{1, \ldots, n-1\}$ such that $x_l = 4$ and $x_{l+1} = 5$, or $x_l = 4$ and $x_{l-1} = 5$. As before, we can assume without loss of generality that there exists l such that $x_l = 4$ and $x_{l+1} = 5$. Let A, B, C and D be the sets of size l in $\mathcal{F}$. Since there are 4 sets of size l and all 5 chains must go through them and also cover them, it follows that exactly two chains have the same element of size l. On the other hand there are 5 elements of size $l+1$, thus each of them belongs to exactly one of the 5 full chains. Putting this together we can assume without loss of generality that $A \in \mathcal{D}_1, \mathcal{D}_2$, and B, C and D are part of the chains $\mathcal{D}_3$,

$\mathcal{D}_4$ and $\mathcal{D}_5$ respectively. Let $A\cup\{a_1\}$, $A\cup\{a_2\}$, $B\cup\{b\}$, $C\cup\{c\}$ and $D\cup\{d\}$ be the 5 elements of size $l+1$ in the chains $\mathcal{D}_1,\dots,\mathcal{D}_5$ respectively, where $a_1\neq a_2$.

We define the sets A' and A'' as in Case 4 of Theorem 2.10 and deduce by the same exact argument that they both belong to $\mathcal{F}$. Thus, we may assume without loss of generality that $B=A'$ and $C=A''$. We also define B' and C' as in the previous section and deduce in the same way that both B' and C' belong to $\mathcal{F}$. The sets of size l are A, A', A'' and D, two of which have to be B' and C'. By the analogue of the subcases (a) and (b) of Case 4 in the previous section, we have that $B'\neq A$, $B'\neq B=A'$, $C'\neq A$, $C'\neq C$, and $B'=C$ and $C'=B$ cannot both hold. Thus we deduce that either $B'=D$ or $C'=D$. Without loss of generality assume $C'=D$. Moreover, we either have $B'=C$ or $B'=D=C'$. It is an easy exercise to see that both cases imply that $a_1'=a_2'$, and either $b=a_2$ or $b=c$.

Let $W=A\setminus\{a_1'\}\in\mathcal{F}$. We observe that the 4 sets of size l are $W\cup\{w_1\}$, $W\cup\{w_2\}$, $W\cup\{w_3\}$ and $W\cup\{w_4\}$, where $w_1,\dots,w_4$ are a_1,a_2,a_1' and c in some order. We note that each of these sets has at least two supersets of size $l+1$ in $\mathcal{F}$ – for example $W\cup\{c\}=C'$ is comparable to both $C\cup\{c\}$ and $D\cup\{d\}$. This immediately tells us that for every i we can can easily construct full chains $\mathcal{C}_1,\dots,\mathcal{C}_5$ that cover $\mathcal{F}$ such that two of these chains go through the set $W\cup\{w_i\}$. On the other hand, we have that $a_1'=a_2'$, which tells us that the two chains that coincide on level l must also coincide on level $l-1$ and, more importantly, their common set of size $l-1$ has to be a subset of all 4 sets of size l. Combining everything we see that this implies that we must have only one set of size $l-1$ in our family, thus $l=1$. In the analogue case where $x_l=4$ and $x_{l-1}=5$, we get $l=n-1$. To summarise, $x_i=5$ for all $i\in\{2,\dots,n-2\}$, $x_1\geq 4$, $x_{n-1}\geq 4$ and $x_0=x_n=1$. Therefore we have that $|\mathcal{F}|\geq 5n-5$.

We are left to show that this bound is achieved for every $n\geq 6$. Let $\mathcal{F}$ be the following family:

$$\mathcal{F}=\{\emptyset,\{1\},\{2\},\{3\},\{4\},[n]\setminus\{1\},[n]\setminus\{2\},[n]\setminus\{3\},[n]\setminus\{4\},$$

$$\{1,2\},\{1,2,5\},\{1,2,5,6\},\dots,[n]\setminus\{3,4\},$$

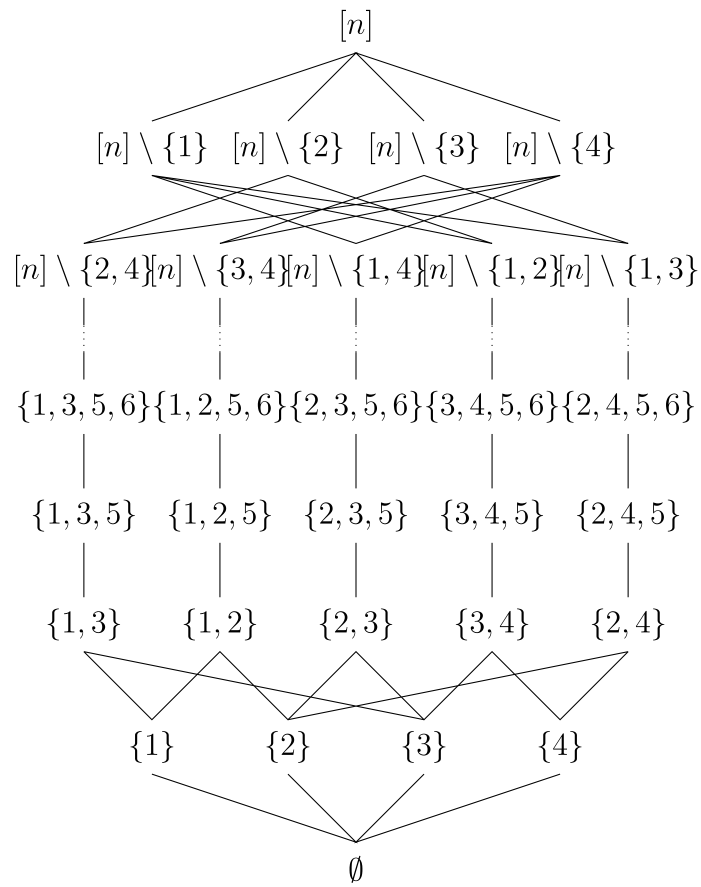
$$\{1,3\},\{1,3,5\},\{1,3,5,6\},\dots,[n]\setminus\{2,4\},$$

$$\{2,3\},\{2,3,5\},\{2,3,5,6\},\dots,[n]\setminus\{1,4\},$$

$$\{4,3\},\{4,3,5\},\{4,3,5,6\},\dots,[n]\setminus\{1,2\},$$

$$\{4,2\},\{4,2,5\},\{4,2,5,6\},\dots,[n]\setminus\{1,3\}\}.$$

This family is pictured below.



It is easy to see that $\mathcal{F}$ is 6-antichain free as it is covered by 5 full chains, and that it has size $1+4+1+4+5(n-3) = 5n-5$. We now prove that whenever we add a set to $\mathcal{F}$ we create a 6-antichain.

Let $X \notin \mathcal{F}$. If $|X| \in \{2, \dots, n-2\}$, then X will form a 6-antichain with the 5 sets in $\mathcal{F}$ that have the same size as X. If $X = \{k\}$ for $k \notin \{1, 2, 3, 4\}$, then X will form a 6-antichain with the sets of size 2 in $\mathcal{F}$. Similarly, if X is the complement of a singleton, it will form a 6-antichain with the sets of size $n-2$ in $\mathcal{F}$.

This proves that $\mathcal{F}$ is 6-antichain saturated. Thus $\text{sat}^*(n, \mathcal{A}_6) = 5n - 5$ for all $n \geq 6$. $\square$

2.4.3 Recent developments

As mentioned at the beginning of this section, prior to our work on the antichain problem, the saturation number for the k-antichain was known to be roughly between $(k-1)n$ and $((k-1)/\log_2(k-1))n$. However, the exact coefficient of n was not known for general k. Based on the work exposed in this subsection, we believed that the following conjecture was true, which strengthened the conjecture in [19] that $\text{sat}^*(n, \mathcal{A}_k) = (k-1)n(1+o(1))$.

Conjecture 2.12. *For each fixed positive integers k we have $sat^*(n, \mathcal{A}_k) = n(k-1) - O(1)$.*

The results in this subsection prove the conjecture for $k = 5$ and $k = 6$, but in addition, the proofs hint at a more general behaviour of antichain-saturated families. In both cases we have seen that almost all levels of the antichain-saturated family have to have the maximal size possible, namely $k-1$, and based on this we made the following conjecture.

Conjecture 2.13. *For each fixed $k > 1$ there exists l with the following property. For n sufficiently large, any k-antichain saturated family $\mathcal{F}$ of subsets of $[n]$ has exactly $k-1$ sets of size i for all $l \leq i \leq n-l$.*

Using the techniques in this subsection, the main obstacle in proving the above conjecture for $k > 6$ came from the increased number of choices the chains we are analysing have when traversing between 2 or 3 consecutive levels of the family. We thought that a first step in proving this conjecture would be to answer the following question.

Conjecture 2.14. *Let $\mathcal{F}$ be a k-antichain saturated family and let x_i be the number of sets of size i in $\mathcal{F}$ for $0 \leq i \leq n$. Then there exist an i such that $x_i = k-1$.*

Recently Bastide, Groenland, Jacob and Johnston [7] showed that all our conjectures were true, thus solving the general saturation problem for the antichain.

2.5 An improved bound on $\mathrm{sat}^*(n, \mathcal{N})$

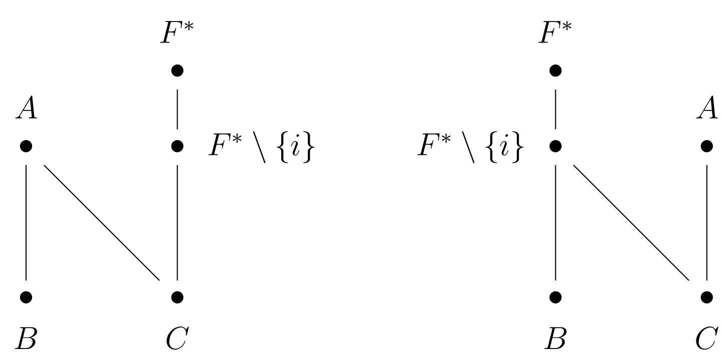
We will show that the saturation number for the poset $\mathcal{N}$ is at least $\sqrt{n}$. The key point is that in every $\mathcal{N}$-saturated family we can find an ordered pair (F, G) such that $F \setminus G = \{i\}$ for every $i \in [n]$. This approach was also used by Martin, Smith and Walker [41] in their analysis of the saturation number of the diamond.

Proposition 2.15. *Let $\mathcal{F}$ be a $\mathcal{N}$-saturated family. Then $|\mathcal{F}| \geq \sqrt{n}$.*

Proof. We will show that for any F in the family and for every $i \in F$, there exist sets A and B in $\mathcal{F}$ such that $A \subseteq F$ and $A \setminus B = \{i\}$. Since the poset $\mathcal{N}$ is invariant under taking complements, this will also tell us that for every $j \notin F$, there exists two sets C and D in $\mathcal{F}$ such that $F \subseteq C$ and $D \setminus C = \{j\}$. From this it would immediately follow that $|\mathcal{F}| \geq \sqrt{n}$, since by fixing an $F \in \mathcal{F}$ we can assign an ordered pair (A, B) to every $i \in [n]$ with the property that $A \setminus B = \{i\}$, and $A, B \in \mathcal{F}$.

Let $F \in \mathcal{F}$ and $i \in F$. If there exist a set $A \in \mathcal{F}$ with $A \subseteq F$, $i \in A$ and $A \setminus \{i\} \in \mathcal{F}$, then we are done. Now suppose that no such A exists and consider an element of the set $\{A \in \mathcal{F} : i \in A, A \subseteq F\}$ of minimal size, which we call F^*. We have that $F^* \setminus \{i\} \notin \mathcal{F}$ and thus it has to form a copy of $\mathcal{N}$ with three other elements of $\mathcal{F}$.

1. Case 1: $F^* \setminus \{i\}$ is one of the maximal elements, then we are in one of the two cases shown below.



   Because we cannot have A, B, C, F^* forming a copy of $\mathcal{N}$ in $\mathcal{F}$, it follows that in both cases A and F^* are comparable. We cannot have $F^* \subseteq A$ as $F^* \setminus \{i\} \parallel A$, therefore we must have $A \subseteq F^*$. Because $F^* \setminus \{i\} \parallel A$, $i \in A$ and $A \neq F^*$. This implies that A is a proper subset of F^* and thus of F, and element of the family containing i. This contradicts minimality of F^*.

2. Case 2: $F^* \setminus \{i\}$ is one of the minimal elements, then we are in the following two cases shown below.

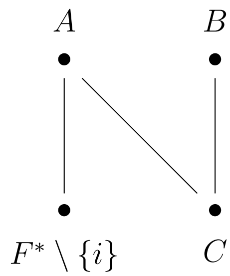


Figure 1

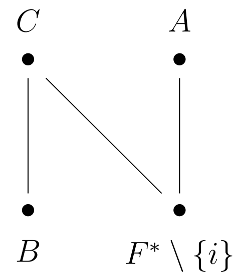


Figure 2

Similarly as before, A, F^*, B, C does not form a copy of $\mathcal{N}$.

In the first case (Figure 1), if $i \notin A$, then $F^* \setminus A = \{i\}$, which gives the pair (F^*, A).
If $i \in A$, then $F^* \subseteq A$, and either $A = F^*$ or $F^* \subset A$.
If $A = F^*$, then C and $F^* \setminus \{i\}$ are incomparable, while $C \subset F^*$. Thus $i \in C$, contradicting the minimality of F^*.
If $F^* \subset A$ then, in order not to create a copy of $\mathcal{N}$ in $\mathcal{F}$, F^* must be comparable to at least one of B or C (which are different from F^* since they are incomparable to $F^* \setminus \{i\}$). Thus $B \subset F^*$ or $C \subset F^*$. Since $C \subset B$, we can assume wlog that $C \subset F^*$, which implies that $i \in C$, contradicting the minimality of F^* again.

In the second case (Figure 2), if $i \notin A$ or $i \notin C$, then similarly as above, we would find the pair (F^*, A) or (F^*, C). So we can assume that $i \in A \cap C$.
If F^* is different from both A and C, then to avoid a copy of $\mathcal{N}$, F^* and B have to be comparable, and since $B \parallel F^* \setminus \{i\}$, we have to have $B \subset F^*$. This gives again $i \in B$ and thus contradicting minimality of F^*.
If $C = F^*$, then we find that $B \subset F^*$ and $i \in B$, leading to the same contradiction. If $A = F^*$, then $F^* \parallel C$ and $F^* \setminus \{i\} \subset C$, so $i \notin C$ and $F^* \setminus C = \{i\}$, which gives the pair (F^*, C) and completes the proof.

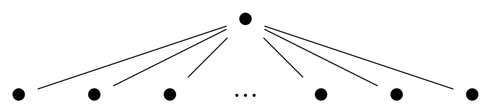
$\square$

2.6 Extensions and further work

In this final section we look at three more general families of posets that include the butterfly. We study their induced saturated number where the ground set is $[n]$.

We call the poset having t maximal incomparable elements and k minimal incomparable elements, all of which are less than both maximal elements, a $K_{t,k}$. First observe that a family $\mathcal{F}$ is $K_{t,k}$-saturated if and only if the family obtained by taking the complements of the sets in $\mathcal{F}$ is $K_{k,t}$-saturated. Therefore $\text{sat}^*(n, K_{t,k}) = \text{sat}^*(n, K_{k,t})$, thus we will only consider the case when $t \leq k$.

We start with the poset $K_{1,k}$, for $k \geq 2$, which is pictured below.



Proposition 2.16. $\sqrt{n} \leq sat^*(n, K_{1,k}) \leq (k-1)(n-1) + 2$.

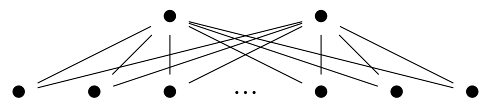
Proof. For the upper bound, consider $\mathcal{F}$ to be the union of $k-1$ full chains that meet at $\emptyset$ and $[n]$ only. Since, by construction, this family does not contain a k-antichain, it also does not contain a copy of $K_{1,k}$. However, if A is a set outside of this family, then the $k-1$ sets of the same size as A from each of the $k-1$ chains, together with $[n]$, will form a copy of $K_{1,k}$. This shows that $\text{sat}^*(n, K_{1,k}) \leq (k-1)(n-1) + 2$ for n large enough.

For the lower bound, we first observe that if $\mathcal{F}$ is a $K_{1,k}$-saturated family and $[n] \in \mathcal{F}$, then $\mathcal{F}$ is in fact an $\mathcal{A}_k$-saturated family, thus $|\mathcal{F}| \geq n(k-1) - O(1) \geq \sqrt{n}$, for n large enough.

If $[n]$ is not in the family, the same argument as the one used to prove the $\sqrt{n}$ lower bound for the poset $\mathcal{N}$ can be used to prove the same lower bound for the poset $K_{1,k}$. using the same tools as the ones presented in the section on the $\mathcal{N}$ poset, we show that if $i \notin A$ for some set A in the family, then we can find two sets F and G in the family such that $F \setminus G = \{i\}$. To finish the proof, we need to know that every element of the ground set is missed by some set of the saturated family. To see that, consider $[n] \setminus \{i\}$. If it is in the family, then we are done. If not, then it must form a $K_{1,k}$ copy when added. Since $[n]$ is not in the family, $[n] \setminus \{i\}$ must be the maximal element, thus the other k elements must miss i. □

If $t > 1$, we observe that no element of $K_{t,k}$ is uniquely *covered* by another element (x covers y if x is greater than y, and there is no w such that $x > w > y$). Therefore $K_{t,k}$ has the *unique twin cover property*: for any element that is uniquely covered, by say x, there exists a different element that is covered by x. Ferrara, Kay, Kramer, Martin, Reiniger, Smith and Sullivan [19] showed that is a poset has the unique twin cover property, then the saturation number is at least $\log_2 n$, thus $\text{sat}^*(n, K_{t,n}) \geq \log_2 n$ for $t > 1$.

We now move on and look at the poset $K_{2,k}$, pictured below.



We remark that the $K_{2,k}$-saturated family of size $O(n^k)$ constructed below has a special structure that will be used later.

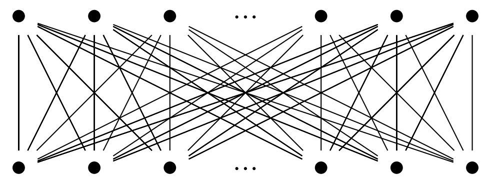
Proposition 2.17. $\log_2 n \leq$ *sat*$^*(n, K_{2,k}) \leq c_k n^k$, *where* c_k *is a constant depending on* k.

Proof. The lower bound has been explained above, thus we only need to provide a $K_{2,k}$-saturated family of size $O(n^k)$. We start with $\mathcal{F}_0$ consisting of all singletons and the chain $\emptyset \cup \{\{1, 2, ..., t\} : 1 \leq t \leq n\}$. It is easy to check that it is $K_{2,k}$-free.
We first prove that If M is a set of size greater than k and $M \notin \mathcal{F}_0$, then $\mathcal{F}_0 \cup M$ contains a $K_{2,k}$.
Let t be the smallest element not in M and $t_1, t_2, ...t_k$ be elements of M, none of which is the maximum of M. Furthermore, assume that t_k is the maximum of the k elements listed above.
If $t = 1$, then M, $\{1, 2, ...t_k\}$ and all singletons $\{t_i\}$ $1 \leq i \leq k$ form a $K_{2,k}$. This is obvious since the singletons form an antichain of size k, they are all contained in both M and $\{1, 2, ...t_k\}$, and M and $\{1, 2, ..., t_k\}$ are incomparable since M does not contain 1 and $\{1, 2, ..., t_k\}$ does not contain the maximum of M.
If $t \neq 1$, then $t < \max(M)$ as M is not in $\mathcal{F}_0$. Let $m = \max\{t, t_k\} < \max(M)$.
We then have the following $K_{2,k}$: M, $\{1, 2, ...m\}$ and $\{t_i\}$, $1 \leq i \leq k$. By a similar argument as above, M and $\{1, 2, ...m\}$ contain all singletons and are incomparable since $t \notin M$ and $\max(M) \notin \{1, 2, ...m\}$.
Now let us list the sets of size at most k that are not currently in F_0: M_1, M_2,..., M_N. We build a $K_{2,k}$-saturated family as follows. We start with $\mathcal{F}_0$. If $\mathcal{F}_0 \cup M_1$ contains a $K_{2,k}$, then we do not add the set to our family, but if it does not, then we add it. We continue this procedure until we reach the end of the list. Let $\mathcal{F}$ be the family obtained in the end. It is $K_{2,k}$ free by construction and also saturated as we showed that if $|M| > k$, then $M \cup \mathcal{F}_0$ contains a $K_{2,k}$, and also the last step ensured that if we add $|M| < k$ to $\mathcal{F}$, then we form a $K_{2,k}$. Observe that all elements of $\mathcal{F}$, apart from the original chain appearing in $\mathcal{F}_0$, have cardinality at most k, thus

$$\mathrm{sat}^*(n, K_{2,k}) \leq |\mathcal{F}| \leq \sum_{i=0}^{k} \binom{n}{i} + n - k = O(n^k).$$

□

We can take a step forward and look at the symmetric poset $K_{k,k}$, which seems to be an even more natural generalisation of the butterfly.



Proposition 2.18. $\log_2 n \leq sat^*(n, K_{k,k}) \leq c_k n^{2k-2}$, *where* c_k *is a constant depending on* k.

Proof. As before, we only need to construct a $K_{k,k}$-saturated family of size $O(n^{2k-2}$. We start with $\mathcal{F}_0$ consisting of all the singletons and the following $k-1$ chains

$$\begin{aligned} \mathcal{C}_1 &: 2, 3, 4, \dots, n, 1 \\ \mathcal{C}_2 &: 1, 3, 4, \dots, n, 2 \\ &\vdots \\ \mathcal{C}_{k-1} &: 1, 2, 3, \dots, n, k-1, \end{aligned}$$

where by the chain $a_1, a_2, \dots, a_n$ we mean the chain $\emptyset, \{a_1\}, \{a_1, a_2\}, \dots, \{a_1, a_2, \dots, a_n\}$. It is clear that $\mathcal{F}_0$ is a $K_{k,k}$-free family since the maximal elements cannot be singletons, but then by the pigeonhole principle at least two will have to be in the same chain, which is impossible since they have to form an antichain.
We have seen in the construction of a $K_{2,k}$-saturated family that if we have a chain and a set M of size greater than k, we can construct a $K_{2,k}$ with M, k arbitrary singletons of M and one element of the chain, as long as those singletons of do not contain the maximum element of M with respect to the order induced by the chain.
Assume now that M is a set of size at least $2k-1$ not in $\mathcal{F}_0$. Let S be the set of maximal elements of M, with respect to the $k-1$ orders induced by the above chains. We have $|M \setminus S| \geq k$, thus we can select $t_1, t_2, \dots, t_k$ elements of M, none of which is the maximum with respect to any of the $k-1$ orders. Therefore, our previous construction gives us $k-1$ sets $A_1, A_2, \dots, A_{k-1}$ with the property that $\{t_1, t_2, \dots, t_k\} \subset A_i$, $A_i \parallel M$ and $A_i \in \mathcal{C}_i$ for all $1 \leq i \leq k-1$.
Observe also that since t_i is not among the maximums, $t_i \geq k$ for every i, and consequently, $t_1, t_2, \dots t_k$ will have the same order in all of the above chains, which is just the usual order. Let t_k be the biggest of them.
If M contains $1, 2, \dots, k-1$, then it contains all maximal elements of the $k-1$ chains, which by our previous construction means that $i \notin A_i$ and we can replace A_i with $[n] - \{i\} \in \mathcal{C}_i$. We then have M, $[n] - \{i\}, \{t_i\}$, $1 \leq i \leq k-1$ forming a copy of $K_{k,k}$.
If M does not contain 1, then the sets that our previous construction gives us for the chains $\mathcal{C}_i$, $i \geq 2$ are $\{1, 3, 4, \dots t_k\}$, $\{1, 2, 4, \dots t_k\}, \dots, \{1, 2, 3, \dots t_k\}$. For $\mathcal{C}_1$ our construction gives $\{2, 3, \dots \max(t_k, t)\}$, where t is the smallest one not in M, with respect to the first order. The first $k-2$ sets have the same size, $t_k - 1$, but are different, thus incomparable. The set obtained from the first chain has size at least $t_k - 1$, but it does not contain 1, so it cannot be comparable to any of the above. Therefore M together with these $k-1$ sets and the k singletons forms a $K_{k,k}$.

If M contains 1, but it does not contain 2, then $t_k \geq 2$ and our construction gives $\{2, 3, \dots t_k\}$, $\{1, 2, 4, \dots t_k\}, \dots, \{1, 2, 3, \dots t_k\}$ for chains $\mathcal{C}_i$, $i \neq 2$.
For $\mathcal{C}_2$ we get $\{1, 3, 4, \dots, \max(t, t_k)\}$. Once again, the first $k-2$ sets are different and have the same size, while the last set has size greater or equal to the size of the others, but it does not contain 2, while all the other set all contain it. Hence, together with M, they form an antichain of size k, and together with the k singletons form a copy of $K_{k,k}$.
It can be shown inductively (if M contains $1, 2, \dots, t$, but not $t+1$) that we always form a copy of $K_{k,k}$.
We therefore have a $K_{k,k}$-free family which forms a $K_{k,k}$ with any set $M \notin \mathcal{F}_0$ with cardinality at least $2k-1$. Thus, the same argument as for for the $K_{2,k}$, gives that

$$\mathrm{sat}^*(n, K_{k,k}) \leq \sum_{i=0}^{2k-2} \binom{n}{i} + (k-1)(n-2k+1) = O(n^{2k-2}).$$

□

These examples illustrate in fact a more general result: together with Bastide, Groenland and Johnston [6], we showed that for every finite poset $\mathcal{P}$ there exist a constant d such that $\mathrm{sat}^*(n, \mathcal{P}) = O(n^d)$.

3 Two Ramsey Theory Questions

3.1 A Ramsey characterisation of eventually periodic words

3.1.1 Introduction

Let X be a non-empty finite or infinite set (called the alphabet) and let X^* denote the set of all finite words $x_1x_2\cdots x_n$ with $n \geq 1$ and $x_i \in X$ for all i. Let $x = x_1x_2x_3\cdots$ be an infinite word on X. Given a finite colouring of X^*, we say that a factorisation $x = u_1u_2u_3\cdots$ (with $u_i \in X^*$) is *monochromatic* if all the u_i have the same colour. When is it the case that x always has a monochromatic factorisation, for any finite colouring of X^*?

This is certainly the case if x is *periodic*. Indeed, if $x = uuu\cdots$ then for any colouring of X^* that very factorisation is trivially monochromatic. In the other direction, Wojcik and Zamboni [57] proved that if x is not periodic then there exists a finite colouring of X^* for which x does not have a monochromatic factorisation. Thus the above Ramsey condition actually characterises the periodic words.

We remark that if we are allowed to pass to a suffix of x then this characterisation breaks down completely. Indeed, *every* word x has the property that for every finite colouring of X^* there is a suffix of x having a monochromatic factorisation. This result is due to Schützenberger [51], and it follows from Ramsey's theorem. To see this, let x be the word $x_1x_2\cdots$. Given a finite colouring ϕ of X^*, we define a colouring of $\mathbb{N}^{(2)}$, the edge set of the complete graph on the natural numbers, by giving the pair (i,j), where $i<j$, the colour $\phi(x_ix_{i+1}\cdots x_{j-1})$. By Ramsey's theorem, there is a monochromatic infinite set for this colouring, say $m_1 < m_2 < \cdots$. But now we note that the finite words $x_{m_i}x_{m_i+1}\cdots x_{m_{i+1}-1}$, for each i, are all assigned the same colour by ϕ, and they form a factorisation of the suffix of x starting at position m_1.

Actually, the above argument shows that more: it shows that, for any colouring, there is a suffix of x having a factorisation $u_1u_2\cdots$ in which every word $u_iu_{i+1}\cdots u_j$, for $i \leq j$, has the same colour. It was shown by de Luca and Zamboni [14] that this strengthened form is actually equivalent to Ramsey's theorem.

In light of these results, it is natural to ask if there is a Ramsey characterisation of the *eventually periodic* words over X, i.e., infinite words of the form $uvvv\cdots$ with $u,v \in X^*$. We say that a factorisation $x = u_1u_2\cdots$ is *super-monochromatic* if each word $u_{k_1}u_{k_2}\cdots u_{k_n}$, where $k_1 < \cdots < k_n$, is the same colour. Our motivation for considering this notion comes from the following observation: if x is eventually periodic then for every finite colouring of X^* there is a suffix of x having a super-monochromatic factorisation.

Indeed, given a finite colouring ϕ of X^*, it suffices to take a suffix of x that is periodic: say $y = uuu\cdots$. We induce a colouring of $\mathbb{N}$ by giving the number n the colour $\phi(u^n)$. By Hindman's theorem [26], there exists an infinite set $M \subset \mathbb{N}$, say $M = \{a_1, a_2, \cdots\}$, where $a_1 < a_2 < \cdots$, such that every (non-empty) finite sum of distinct elements of M has the same colour. But now the factorisation $y = u^{a_1}u^{a_2}\cdots$

is super-monochromatic.

We show in this section that this condition actually characterises the eventually periodic words. In other words, we will show that if the word x has the property that for every finite colouring of X^* there is a suffix of x having a super-monochromatic factorisation, then x is eventually periodic.

Theorem 3.1. *Let x be an infinite word on alphabet X. Then x is eventually periodic if and only if for every finite colouring of X^* there is a suffix of x having a super-monochromatic factorisation.*

(Note that if we 'swap the quantifiers' we would have the statement that x is eventually periodic if and only if there is a suffix of x such that every finite colouring of this suffix has a super-monochromatic factorisation – which is true by the remarks above.)

This result has actually been around in the community as a folklore conjecture for some time (see e.g. [55]). There have been some partial results, of which the strongest is perhaps the result of Wojcik [55], who showed that Theorem 3.1 holds for words x that have at most finitely many distinct square factors, where by a *square factor* we mean a non-empty block of the form uu which occurs in x. But the result was not even known for Sturmian words, which are regarded as the 'simplest' *aperiodic* words, i.e., words that are not eventually periodic.

Let us also remark that if one is allowed to pass to the shift orbit closure then the situation is completely different. Recall that the *shift orbit closure* of an infinite word x is the closure of the set of suffices of x in the product topology: equivalently, it consists of all infinite words y such that every factor of y is a factor of x. Van Thé and Zamboni (see [56]) showed that, for *any* infinite word x over a finite alphabet X, whenever X^* is finitely coloured there is a word y in the shift orbit closure of x having a super-monochromatic factorisation.

Our proof is in two separate parts. In the first part, we reduce the problem to a problem that concerns not colourings of words, but colourings of $\mathbb{N}^{(2)}$. It will turn out from this reduction that Theorem 3.1 is implied by the following result, which concerns alternating sums.

Theorem 3.2. *There exists a finite colouring of $\mathbb{N}^{(2)}$ such that there do not exist $x_1 < x_2 < \cdots$ for which the set of all pairs $(x_{k_1} - x_{k_2} + x_{k_3} - \cdots + x_{k_t}, x_{k_{t+1}})$, where t is odd and $k_1 < k_2 < \cdots < k_{t+1}$, is monochromatic.*

The second part of the proof thus consists of a proof of Theorem 3.2. What is interesting is the role played by the alternation. Indeed, if all the signs were plus-signs then the Ramsey statement would be in the affirmative. In other words, for any finite colouring of $\mathbb{N}^{(2)}$ there exist $x_1 < x_2 < \cdots$ for which the set of all pairs $(x_{k_1} + x_{k_2} + x_{k_3} + \cdots + x_{k_t}, x_{k_{t+1}})$, where $k_1 < k_2 < \cdots < k_{t+1}$, is monochromatic. This follows for example from the Milliken-Taylor theorem ([42], [53]). To see this, recall that the Milliken-Taylor theorem asserts that whenever the set of all pairs (A, B), where A and B are (non-empty) finite subsets of $\mathbb{N}$ with $\max A < \min B$, is finitely coloured there exists a sequence $A_1, A_2, \cdots$ of finite subsets of $\mathbb{N}$, with $\max A_n < \min A_{n+1}$ for all n, such that all of the pairs (S, T), where S and T are finite unions of the A_n with $\max S < \min T$, are the same colour. So we just need to 'transfer' the colouring from numbers to finite sets: given a finite colouring Λ of $\mathbb{N}^{(2)}$, we colour each pair (A, B) as above with the colour $\Lambda(\sum_{i\in A} 2^i, \sum_{i\in B} 2^i)$. Given the sequence $A_1, A_2, \cdots$ as guaranteed by the Milliken-Taylor theorem, we set $x_n = \sum_{i\in A_n} 2^i$, and now we get that every pair $(x_{k_1} + x_{k_2} + x_{k_3} + \cdots + x_{k_t}, x_{k_{t+1}})$, and in fact even every pair $(x_{k_1} + x_{k_2} + x_{k_3} + \cdots + x_{k_t}, x_{k_{t+1}} + x_{k_{t+2}} + \cdots + x_{k_s})$, has the same colour. The interested reader is referred to [29] for a general discussion of the Milliken-Taylor theorem and many related results, although we stress that this section is self-contained.

The colouring argument needed to establish Theorem 3.2 is rather complicated, and it is perhaps worthwhile to describe why this is the case. As we will see, it turns out to be useful to 'change variables' to some other variables, the y_n, that satisfy a related condition. However, this related condition is not preserved by passing to subsequences. This is in contrast to the usual situations when one is finding a 'bad' colouring (see for example [15], [28], [39]), where the first step is always to pass to a subsequence or sequence of sums in which the supports of the elements, when written say in binary, are disjoint, and even more are ordered in the sense that one variable's support ends before the next one's begins. Since this step is not available to us here, we have to deal with the situation when the y_i do not have disjoint supports, and therefore we need to consider how the carry-digits behave when we add them to each other. This means that the colouring, and especially the proof that it works, is far more difficult than for other problems that superficially look similar.

The plan for this section is as follows. In Subsection 3.1.2 we show that Theorem 3.1 is implied by Theorem 3.2, and then in Subsection 3.1.3 we prove Theorem 3.2. Subsection 3.1.4 is devoted to some related problems that we have been unable to solve.

3.1.2 The link between the two main theorems

In this subsection we will show that Theorem 3.2 implies Theorem 3.1. As explained above, we know that given any finite colouring of X^*, any eventually periodic word has a suffix that admits a super-monochromatic factorisation. Therefore, we only need to prove the reverse implication of Theorem 3.1: given an aperiodic word x (i.e. x is not eventually periodic), we must construct a finite colouring of X^* for which no suffix of x has a super-monochromatic factorisation. This will be accomplished using the colouring given by Theorem 3.2.

Theorem 3.3. *Theorem 3.2 implies Theorem 3.1.*

Proof. By Theorem 3.2, there exists a finite colouring of $\mathbb{N}^{(2)}$, Λ, for which there is no increasing sequence $(x_k)_{k\geq 1}$ such that all edges of the form $(x_{k_1} - x_{k_2} + \cdots + x_{k_t}, x_{k_{t+1}})$ have the same colour, where $k_1 < k_2 < \cdots < k_{t+1}$ and t is odd. Let $\mathcal{C}$ be the set of colours of Λ.

Let x be an aperiodic word. We will use Λ to construct a finite colouring ϕ of X^* for which no suffix of x has a super-monochromatic factorization. We denote by x_i the i^{th} letter of x.

For any factor u of x, define $A_x(u) = \min\{n \in \mathbb{N} : u = x_n x_{n+1} \cdots x_{n+|u|-1}\}$ and $B_x(u) = A_x(u) + |u|$, where $|u|$ is the length of $|u|$. In other words, $A_x(u)$ is the start position of the first occurrence of u in x, while $B_x(u)$ is the first position after this first occurrence of u.
For an arbitrary factorisation $(u_i)_{i\geq 1}$ of x, we say $(w_i)_{i\geq 1}$ is a *block subfactorisation* of $(u_i)_{i\geq 1}$ if there exists a strictly increasing sequence of positive integers $(k_j)_{j\geq 1}$ such that $w_1 = u_1 u_2 \cdots u_{k_1}$ and $w_i = u_{k_{i-1}+1} \cdots u_{k_i}$ for each $i \geq 2$. Here by $(u_i)_{i\geq 1}$ being a factorisation of x we mean $x = u_1 u_2 u_3 \cdots$. We immediately note that a block subfactorisation of a super-monochromatic factorisation is still super-monochromatic.

Now we are ready to define a colouring $\phi : X^* \to (\mathcal{C} \times \{0,1\}) \cup \{2\}$ as follows:

1. If u is not a factor of x, then $\phi(u) = 2$.
2. If u is a factor of x and there exists a factorisation $u = vw$ such that $A_x(u) = A_x(v)$ and $B_x(u) = B_x(w)$ (in other words, the first occurrence of v in x is as the start of the first occurrence of u in x and also the first occurrence of w in x is as the end of the first occurrence of u in x), then $\phi(u) = (\Lambda(A_x(u), B_x(u)), 0)$.
3. Otherwise $\phi(u) = (\Lambda(A_x(u), B_x(u)), 1)$.

We claim that for this colouring ϕ no suffix of x has a super-monochromatic factorisation.
Suppose to the contrary that there is a suffix y of x having a super-monochromatic factorisation $y = u_1 u_2 \cdots$. Let u_0 be the (possibly empty) prefix of x so that $x = u_0 y$. It is important to remember that each factor u_i may occur in several places in y, not necessarily only in the place immediately following $u_1 u_2 \cdots u_{i-1}$. We call this place the *standard* position of u_i. Let the colour of all concatenations of the u_i be $c \in (\mathcal{C} \times \{0,1\}) \cup \{2\}$. Since $\phi(u_1) = c$, and since u_1 is a factor of x, we have $c \neq 2$. Thus, $c = (a, b)$ where $a \in \mathcal{C}$ and $b \in \{0,1\}$.

Claim 1. *By passing to a block subfactorisation, we may assume that for every $i \in \mathbb{N}$, the first occurrence of u_i in x is exactly the standard position of u_i.*

Equivalently, this means $A_x(u_i) = |u_0| + |u_1| + \cdots + |u_{i-1}| + 1$ for all $i \geq 1$.

Proof. We start by showing that we may assume $A_x(u_1) = |u_0| + 1$. If initially $A_x(u_1) < |u_0| + 1$, we consider all concatenations $u_1 u_2 \cdots u_k$. If $A_x(u_1 u_2 \cdots u_k) = |u_0| + 1$ for some k, we set our first factor to be $u_1 u_2 \cdots u_k$ and renumber the rest of them. Since concatenating consecutive factors does not change the super-monochromatic property,

the new factorisation is still super-monochromatic and the first factor now has the desired property.
If on the other hand $A_x(u_1u_2\cdots u_k) < |u_0|+1$ for all $k \geq 1$, then each concatenation $u_1u_2\cdots u_k$ first occurs in x starting at some position in u_0. Since there are infinitely many of them and only finitely many positions in u_0, there exists a position i, with $i \leq |u_0|$, at which infinitely many $u_1u_2\cdots u_k$ start. This immediately implies that the suffix of x starting at position i is exactly y. But this means that x has two suffices equal to y, which implies that x is eventually periodic. More precisely, we have $x_1x_2\cdots x_{i-1}y = x_1x_2\cdots x_{|u_0|}y$. Therefore $y_k = x_{i-1+k}$ and $y_k = x_{|u_0|+k}$ for any k. It follows that $x_{i-1+k} = x_{|u_0|+k}$ for any k, thus x is eventually periodic with period $|u_0|-i+1$, contradicting our initial assumption.
Therefore we may assume u_1 has the desired property. We now move on to u_2 and repeat the same argument, looking at concatenations of the form $u_2u_3\cdots u_k$: so u_2 may be assumed to have the same property too. It follows inductively that we may assume that all u_i have the property stated in the claim. □

We further observe that once we have the property that the first occurrence of each u_i in x is in the standard position, then any block subfactorisation has this property as well. For example, u_1u_2 cannot appear earlier or else u_1 would. Therefore, we can further assume that $|u_{n+1}| \geq |u_1u_2\cdots u_n|$ for all $n \geq 1$.

We now look at u_1u_2. Because of the above claim we certainly have $A_x(u_1u_2) = A_x(u_1)$ and $B_x(u_1u_2) = B_x(u_2)$. This means that u_1u_2 is a factor of x that satisfies the factorisation condition specified in the colouring rule. Thus $\phi(u_1u_2) = (a,b) = (\Lambda(A_x(u_1u_2), B_x(u_1u_2)), 0)$, and so the colour of the factorisation is $(a, 0)$ with $a \in \mathcal{C}$.

Claim 2. *The word $u_1u_2\cdots u_n$ is a suffix of u_{n+1}, for every $n \geq 1$.*

Proof. Consider the concatenation $u_1u_2\cdots u_nu_{n+2}$. Because our factorisation $u_1u_2\cdots$ is super-monochromatic we have that $\phi(u_1u_2\cdots u_nu_{n+2}) = (a, 0)$. This means that not only is $u_1u_2\cdots u_nu_{n+2}$ a factor of x, but also that $u_1u_2\cdots u_nu_{n+2} = vw$ for some v, w with $A_x(u_1u_2\cdots u_nu_{n+2}) = A_x(v)$ and $B_x(u_1u_2\cdots u_nu_{n+2}) = B_x(w)$.
We now have two possibilities: either v is a prefix of $u_1u_2\cdots u_n$ or $u_1u_2\cdots u_n$ is a prefix of v.
If v is a prefix of $u_1u_2\cdots u_n$ then u_{n+2} is a suffix of w. Therefore we immediately have that $A_x(v) \leq A_x(u_1u_2\cdots u_n)$ and $B_x(w) \geq B_x(u_{n+2})$. It follows that

$$B_x(u_1u_2\cdots u_nu_{n+2}) = B_x(w) \geq B_x(u_{n+2}) = B_x(u_1u_2\cdots u_nu_{n+1}u_{n+2})$$

and

$$A_x(u_1u_2\cdots u_nu_{n+2}) = A_x(v) \leq A_x(u_1u_2\cdots u_n) = A_x(u_1u_2\cdots u_nu_{n+1}u_{n+2}),$$

where the last equalities in each line follow from the property that, for each i, the first occurrence of u_i in x is at its standard position. Consequently, any consecutive concatenation of such factors has the same property.
Putting these two inequalities together, we obtain

$$B_x(u_1u_2\cdots u_nu_{n+2}) - A_x(u_1u_2\cdots u_nu_{n+2}) \geq$$
$$B_x(u_1u_2\cdots u_nu_{n+1}u_{n+2}) - A_x(u_1u_2\cdots u_nu_{n+1}u_{n+2}).$$

This is equivalent to $|u_1u_2\cdots u_nu_{n+2}| \geq |u_1u_2\cdots u_nu_{n+1}u_{n+2}|$, which is a contradiction. Hence $u_1u_2\cdots u_n$ is a prefix of v, and so w is a suffix of u_{n+2}. This implies that $B_x(w) \leq B_x(u_{n+2})$. Since u_{n+2} is a suffix of $u_1u_2\cdots u_nu_{n+2}$, the same argument gives

$$B_x(u_{n+2}) \leq B_x(u_1u_2\cdots u_nu_{n+2}) = B_x(w).$$

Therefore $B_x(u_{n+2}) = B_x(u_1u_2\cdots u_nu_{n+2})$. By Claim 1 we also have $B_x(u_{n+1}u_{n+2}) = B_x(u_{n+2})$. We conclude that $B_x(u_{n+1}u_{n+2}) = B_x(u_1u_2\cdots u_nu_{n+2})$ which, combined with $|u_{n+1}| \geq |u_1u_2\cdots u_n|$, gives that $u_1u_2\cdots u_n$ is a suffix of u_{n+1}. □

Claim 3. *The word $u_{k_1}u_{k_2}\cdots u_{k_m}$ is a suffix of u_n, for every $k_1 < k_2 < \cdots < k_m < n$.*

Proof. We prove the statement by induction on the number of factors.
From Claim 2 we get that u_t is a suffix of u_{t+1} for every $t \geq 1$. Since 'is a suffix of' is a transitive property, we obtain that u_t is a suffix of u_n for every $t < n$, thus the base case is proved.
Assume now that the result is true for all concatenations of at most s factors, and consider a concatenation $u_{k_1}u_{k_2}\cdots u_{k_s}u_{k_{s+1}}$ with all $k_i < n$. If the indices are consecutive numbers, Claim 2 guarantees that this is a suffix of $u_{k_{s+1}+1}$, which is a suffix of u_n. If that is not the case, then $k_i + 1 < k_{i+1}$ for some $i \leq s$. We take i to be the biggest such index and apply the induction hypothesis to obtain that $u_1u_2\cdots u_{k_i}$ is a suffix of u_{k_i+1}, which is a suffix of $u_{k_{i+1}-1}$. It then follows that $u_{k_1}u_{k_2}\cdots u_{k_s}u_{k_{s+1}}$ is a suffix of the consecutive concatenation of factors $u_{k_{i+1}-1}u_{k_{i+1}}\cdots u_{k_{s+1}}$, which is a suffix of $u_{k_{s+1}+1}$, thus a suffix of u_n. This finishes the inductive step and hence proves the claim. □

Combining Claim 1 and Claim 3, we obtain that $B_x(u_{k_1}u_{k_2}\cdots u_{k_t}) = B_x(u_{k_t})$. This is because repeatedly applying Claim 3 tells us that $u_{k_1}u_{k_2}\cdots u_{k_t}$ is a suffix of a consecutive concatenation of factors ending in u_{k_t}. Note that, by construction, we also have $A_x(u_{n+1}) = B_x(u_n)$.

We now return to our original colouring. By assumption, we have that

$$\Lambda(A_x(u_{k_1}u_{k_2}\cdots u_{k_t}), B_x(u_{k_1}u_{k_2}\cdots u_{k_t})) = a$$

for every $k_1 < k_2 < \cdots < k_t$. We also know that

$$\begin{aligned} A_x(u_{k_1}u_{k_2}\cdots u_{k_t}) &= B_x(u_{k_1}u_{k_2}\cdots u_{k_t}) - |u_{k_1}u_{k_2}\cdots u_{k_t}| \\ &= B_x(u_{k_t}) - |u_{k_1}| - |u_{k_2}| - \cdots - |u_{k_t}| \\ &= A_x(u_{k_t}) + A_x(u_{k_{t-1}}) + \cdots + A_x(u_{k_1}) - B_x(u_{k_{t-1}}) - \cdots - B_x(u_{k_1}), \end{aligned}$$

where we used the fact that $|u_{k_i}| = B_x(u_{k_i}) - A_x(u_{k_i})$.
Let $m_i = B_x(u_i)$ for each i. Clearly $(m_i)_{i\geq 1}$ is a strictly increasing sequence. We then have

$$A_x(u_{k_1}u_{k_2}\cdots u_{k_t}) = m_{k_t-1} + m_{k_{t-1}-1} + \cdots + m_{k_1-1} - m_{k_{t-1}} - \cdots - m_{k_1}.$$

It follows that for any choice of $k_1 < k_2 < \cdots < k_t$, we have that

$$\Lambda(m_{k_1-1} - m_{k_1} + m_{k_2-1} - m_{k_2} + \cdots + m_{k_{t-1}-1} - m_{k_{t-1}} + m_{k_t-1}, m_{k_t}) = a.$$

By choosing the k_i appropriately, it follows that that for any l odd and $i_1 < i_2 < \cdots < i_l < i_{l+1}$, we have $\Lambda(m_{i_1} - m_{i_2} + \cdots - m_{i_{l-1}} + m_{i_l}, m_{i_{l+1}}) = a$, which contradicts the choice of Λ. □

3.1.3 Constructing the colouring Λ

In this subsection we will construct a finite colouring of $\mathbb{N}^{(2)}$ with the property that for no infinite strictly increasing sequence $(x_n)_{n\geq 1}$ do all pairs of the form $(x_{k_1} - x_{k_2} + \cdots - x_{k_{t-1}} + x_{k_t}, x_{k_{t+1}})$ have the same colour, where $k_1 < k_2 < \cdots < k_{t+1}$ and t is odd.

We start with a simple observation. Let $y_1 = x_1$ and $y_n = x_n - x_{n-1}$ for each $n \geq 2$. So $x_n = y_n + y_{n-1} + \cdots + y_1$.

Now let t be odd and $k_1 < k_2 < \cdots < k_t$. We then have that $x_{k_1} - x_{k_2} + \cdots - x_{k_{t-1}} + x_{k_t} = x_{k_1} + (x_{k_3} - x_{k_2}) + \cdots + (x_{k_t} - x_{k_{t-1}})$. Thus $x_{k_1} - x_{k_2} + \cdots - x_{k_{t-1}} + x_{k_t} = y_1 + y_2 + \cdots + y_{k_1} + (y_{k_2+1} + \cdots + y_{k_3}) + \cdots + (y_{k_{t-1}+1} + \cdots + y_{k_t})$.

Let $1 < m_1 < \cdots < m_s$ be integers and set in the above expression $k_1 = 1$, $k_2 + 1 = k_3 = m_1, \cdots, k_{2s} + 1 = k_{2s+1} = m_s$. Then we obtain that $x_{k_1} - x_{k_2} + \cdots - x_{k_{t-1}} + x_{k_t} = y_1 + y_{m_1} + \cdots + y_{m_s}$. This shows that Theorem 3.2 is equivalent to:

Theorem 3.4. *There exists a finite colouring of* $\mathbb{N}^{(2)}$ *such that there does not exist a sequence of natural numbers* $(y_k)_{k\geq 1}$ *for which all pairs of the form* $(y_1 + y_{k_1} + y_{k_2} + \cdots + y_{k_t}, y_1 + y_2 + y_3 + \cdots + y_{k_{t+1}})$ *have the same colour, for all choices of* $1 < k_1 < k_2 < \cdots < k_{t+1}$.

Proof. Our construction of the colouring will be in several stages. At each stage, we add more colours, meaning that we take the product colouring of the colouring we have so far with a new colouring. The conditions on a supposed sequence $(y_n)_{n\geq 1}$ satisfying the conditions in Theorem 3.4 will thus become more and more stringent, eventually resulting in a contradiction.

As the colouring is rather complex, we give a brief overview of what each stage is supposed to achieve. We first need some notation. We work with natural numbers in their binary form, so strings of '0' and '1'. The *position* of a digit is the power of 2 it represents. The *first* digit of n in binary is at position i, where i is the greatest non-negative integer such that 2^i divides n. The *last* digit of n in binary is at position j, where $2^j \leq n < 2^{j+1}$. The *support* of n is the set of positions having the digit '1' in its binary expansion. For example, let $n = 2^7 + 2^6 + 2^3$. Below n is shown in binary, where the first row represents the position of each digit. The support of n is $\{3, 6, 7\}$.

Position number	7	6	5	4	3	2	1	0
Binary digit of n	1	1	0	0	1	0	0	0
	last digit				first digit			

In Stage 1 we will ensure that the supports (in binary) of the y_n do roughly 'go off to the left'. What we hope to achieve is a 'staircase' pattern. More precisely, writing n_i for the position of the first digit of y_i and m_i for the position of its last digit, we would like to ensure that the n_i and the m_i form strictly increasing sequences, with y_{i+2} starting to the left of where y_i ends for all i. This is the idea behind the definition of 'Type A' below. However, it will turn out that we cannot always achieve this, and so there is a residual case to deal with that we call 'Type B', which represents what happens when there is no way to pass to Type A. Of course, we cannot ignore this case, but somehow it has the feel of an annoying special case: the reader should perhaps view Type A as the 'main' case. Roughly speaking, the sequence is of Type B when the sequence, with y_1 removed, is of Type A, but also y_1 starts where y_2 starts, and $y_1 + y_2$ starts where y_3 starts, and $y_1 + y_2 + y_3$ starts where y_4 starts, and so on.

Then Stage 2 gives that the supports of the y_i, despite having the above staircase pattern, cannot be disjoint. And it then starts to deal with the unpleasant issues arising from 'carry digits', that arise when adding numbers whose supports are not disjoint. It will give that the carries must be short-range, in a certain sense. The fact that we forbid the carries to propagate arbitrarily far will actually show that Type B cannot occur. And finally, Stage 3 will eliminate these short-range carries as well.

We are now ready to turn to the proof itself. As stated above, we construct our colouring step by step. When colouring a pair (a, b), $a < b$, we will often look at just a, just b, or just $b - a$. At other times we will make full use of the fact that we are colouring pairs, not just numbers.

Stage 1. To start with, our colours are quadruples (c_0, c_1, c_2, c_3) where $c_0, c_1, c_2 \in \{0, 1, 2\}$ and c_3 is one of the four possible bit-strings of length 3 having '1' at their rightmost positions. We give the colour (c_0, c_1, c_2, c_3) to the pair (a, b), $a < b$, if the last digit of $b - a$ in binary is at position c_0 modulo 3, the first digit of $b - a$ in binary is at position c_1 modulo 3, the last digit of a in binary is at position c_2 modulo 3, and the first 3 digits of $b - a$ form, from left to right, c_3.
Assume now $(y_n)_{n\geq 1}$ is a sequence that satisfies the conditions of Theorem 3.4 for the above colouring. The pairs that are all the same colour are the pairs of the form $(y_1+y_{k_1}+\cdots+y_{k_{t-1}}+y_{k_t}, y_1+y_2+\cdots+y_{k_{t+1}})$ for some $t \geq 0$ and $1 < k_1 < \cdots < k_t < k_{t+1}$. The differences $b - a$, where (a, b) is a pair of the above form, are precisely the sums of the form $y_{k_1} + \cdots + y_{k_t}$ for some $t \geq 1$ and $1 < k_1 < \cdots < k_t$. It follows that any such finite sum must have the same first 3 digits, with the first digit at a fixed position modulo 3.

Claim 1. *There do not exist $j > i > 1$ such that y_j and y_i have their first digit at the same position.*

Proof. Assume that two such y_i and y_j exist. The position of their first digit is the same, say n_0, and by our colouring they have the same first 3 digits since the pairs $(y_1 + \cdots + y_{i-1}, y_1 + \cdots + y_i)$ and $(y_1 + \cdots + y_{j-1}, y_1 + \cdots + y_j)$ have the same colour.

On the other hand, for $j > i+1$, the colouring also requires the pair $(y_1 + \cdots + y_{i-1} + y_{i+1} + \cdots + y_{j-1}, y_1 + \cdots + y_j)$ to have the exact same colour, thus $y_i + y_j$ must have the first digit at a position congruent to n_0 modulo 3. If $j = i+1$, we consider the pair $(y_1 + \cdots + y_{i-1}, y_1 + \cdots + y_{i+1})$, thus $y_i + y_{i+1} = y_i + y_j$ must have the first digit at a position congruent to n_0 modulo 3 in this case too. However, adding two identical strings in binary shifts the support by exactly one to the left. Hence, when we add y_i and y_j, their first '1', which was at position n_0 for both of them, is moved to position $n_0 + 1 \not\equiv n_0$ modulo 3, a contradiction. □

We now know that, except for possibly y_1, no two terms of the sequence start at the same position.

Let $(z_n)_{n\geq 1}$ be a sequence of natural numbers. We call $(w_n)_{n\geq 1}$ a *full block subsequence* or simply a *block subsequence* of $(z_n)_{n\geq 1}$ if there exists an increasing sequence of natural numbers $(k_n)_{n\geq 1}$ such that $w_1 = z_1 + \cdots + z_{k_1}$ and $w_n = z_{k_{n-1}+1} + \cdots + z_{k_n}$ for $n \geq 2$. We stress that there are no 'gaps': every z_n appears as a summand in some w_m.

We observe that if the sequence $(y_n)_{n\geq 1}$ satisfies the conditions in Theorem 3.4 for a given colouring then so does any of its block subsequences. So, by passing to a block subsequence, we may assume that $(y_n)_{n\geq 1}$ is strictly increasing.

Let $(z_n)_{n\geq 1}$ be a sequence of natural numbers. Let n_i be the position of the first digit of z_i and m_i the position of the last digit of z_i, for all $i \geq 1$. We call the sequence $(z_n)_{n\geq 1}$ of *Type A* if for all $i \geq 1$, $n_i < n_{i+1}$, and $m_i < m_{i+1}$, and $m_i + 1 < n_{i+2}$. We call the sequence $(z_n)_{n\geq 1}$ of *Type B* if none of its block subsequences is of Type A, the sequence $(z_n)_{n\geq 2}$ is of Type A, and also $n_1 = n_2$, and $m_1 < m_2$, and $m_1 + 1 < n_3$. We remark that this definition of 'Type B' is more abstract that the one informally described in the proof overview above: the reason is that we want this definition to capture the idea of 'we cannot pass to Type A'.

Claim 2. *By passing to a block subsequence, we may assume that $(y_n)_{n\geq 1}$ is of either Type A or Type B.*

Proof. As above, let the positions of the first and last digits of y_i be n_i and m_i respectively.

Assume first that there is no k such that the first digit of y_k is at the same position as the first digit of y_1. We will prove that we can find a block subsequence of Type A. We start with y_1. By Claim 1, only finitely many terms have the position of their first digit at most the position of the first digit of y_1. Let y_{l_1} be the last one of them. We replace y_1 by the consecutive sum $y_1 + y_2 + \cdots + y_{l_1}$ and relabel the sequence accordingly. Now we move on to the second term. All terms after y_1 now have the position of their first digit greater than that of y_1. Again, only finitely many terms have their first digit at a position at most one plus the position of the last digit of y_1. Let y_{l_2} be the last one of them. We now replace y_2 by $y_2 + y_3 + \cdots + y_{l_2} + y_{l_2+1}$ and again relabel the sequence. Now all terms after y_2 have their first digit at a position greater than one plus the position of the last digit of y_1. Also, only finitely many have the position of their first digit at most one plus the position of the last digit of y_2. Let y_{l_3} be the last one of them. We replace y_3 by $y_3 + y_4 + \cdots + y_{l_3+1}$. Now continue inductively. Hence, we obtain a block subsequence of Type A.

We now assume that $(y_n)_{n\geq 1}$ does not have any block subsequence of Type A. Therefore, there is a k such that the first digit of y_1 is at the same position as the first digit of y_k. We now construct a block subsequence of Type B.
First we note that we may assume that there is no $i > 1$ such that $n_i < n_1$: if such an i exists, then we replace y_1 with $y_1 + y_2 + \cdots + y_i$ and relabel. This new block subsequence has the property that the first digits of its terms are all on different positions. Therefore, by the argument presented at the begining of the proof, we can construct a block subsequence of Type A, which is a contradiction.
We fix y_1. We know that no y_i starts before it and only y_k starts at the same position. Only finitely many y_i have their first digit at a position at most one plus the position of the last digit of y_1. Let y_s be the last of them and let $t = \max(s+1, k)$. We now replace y_2 with $y_2 + \cdots + y_t$. Note that in this block subsequence y_1 and y_2 start on the same position, and $m_1 < m_2$. Since from now on the terms start on different positions, we repeat the inductive construction presented at the beginning of the proof and thus obtain the desired block subsequence of Type B. □

In what follows, a property that will play a crucial role is the fact that any block subsequence of $(y_n)_{n\geq 1}$ still satisfies Claim 2. More precisely, as we now show, Type A sequences are invariant under taking block subsequences, and the same holds for Type B sequences. That is a direct consequence of binary addition and our colouring so far.

Claim 3. *If $(y_n)_{n\geq 1}$ satisfies the conditions in Theorem 3.4 for the above colouring and is of Type A, then the same also holds for each of its block subsequences, and similarly for Type B.*

Proof. Consider a sum $y_m + y_{m+1} + \cdots + y_n$, where $2 \leq m \leq n$. In any given position, at most two of the summands have a digit 1 and so the last digit of the sum is either at the same position as the last digit of y_n, or one greater. Because of the colour c_0, we conclude that the last digit of $y_m + y_{m+1} + \cdots + y_n$ is at the same position as the last digit of y_n.
If $n \geq 3$, the sum $y_1 + y_2 + \cdots + y_n$ has the last digit at the same position modulo 3 as $y_1 + y_n$, by c_2. Since y_n and y_1 have disjoint supports, the last digit of $y_1 + y_n$ is at the same position as the last digit of y_n. Similarly as above, the last digit of the sum $y_1 + y_2 + \cdots + y_n$ is either at the same position as the last digit of y_n, or one position greater. We conclude that the last digit of $y_1 + y_2 + \cdots + y_n$ is at the same position as the last digit of y_n, for $n \geq 3$.
Finally, we look at $y_1 + y_2$. Its last digit is either at the same position as the last digit of y_2, or one position greater. By c_2, the position of its last digit has to agree modulo 3 with the position of the last digit of $y_1 + y_3$, which is the position of the last digit of

y_3, by disjointness. However, by c_0, y_2 and y_3 have the last digit at the same position modulo 3. We conclude that the last digit of $y_1 + y_2$ is at the same position as the last digit of y_2.
Thus, we have that for any $1 \leq m \leq n$, the position of the last digit of $y_m + \cdots + y_n$ is the position of the last digit of y_n.
If the sequence $(y_n)_{n\geq 1}$ is of Type A, then the position of the first digit of $y_m + \cdots + y_n$ is the position of the first digit of y_m for all $1 \leq m \leq n$. Combining these two observations we obtain that by passing to a block subsequence, we also obtain a Type A sequence.
If the sequence is of Type B, the first digit of $y_m + \cdots + y_n$ is at the same position as the first digit of y_m if $m > 1$. If $m = 1$, then $y_1 + \cdots + y_n$ has to start at the same position as some other term $y_t + y_{t+1} + \cdots + y_{t+s}$, where $t \geq n+1$. This is because $(y_n)_{n\geq 1}$ is of Type B and thus cannot have any block subsequence of Type A, which can always be constructed from a sequence with terms starting at different positions. Since the last digit of this sum is at the same position as the last digit of y_n which is less than the position of the first digit of y_{n+2}, we must have that the first digit of $y_1 + \cdots + y_n$ is at the same position as the first digit of y_{n+1}. This shows that any block subsequence is of Type B. □

We note that Claim 3 also implies that the last digit of any sum is at the same position as the last digit of its biggest term.

Stage 2. Let a, b with $a < b$ be a pair of natural numbers. We write a and b in binary and we call a position i a '2' if both a and b have at position i the digit 1. We call a position i a '1' if exactly one of a and b has at position i the digit 1. We define the *number of '2 to 1'-jumps* of (a, b), denoted by $J(a, b)$, to be the number of transitions from a '2' to a '1' as we traverse the positions in increasing order, ignoring the positions where both numbers have a '0'. For example, if $a = 100000100$ and $b = 1101010111$, then the positions labelled '1' are 0,1,4,6 and 9, the positions labelled '2' are 2 and 8 and the positions ignored are 3,5 and 7. Thus the number of '2 to 1'-jumps is 2, namely the jump from position 2 to position 4 and the jump from position 8 to position 9.

Let c be a natural number and $c = l_p l_{p-1} \cdots l_1$ its binary representation. We call a binary string not containing a '0', $a_s \cdots a_1 = 1 \cdots 1$, an *interval* of c if there exits $1 \leq i \leq p - s + 1$ such that $l_{i+s-1} \cdots l_i = a_s \cdots a_1$, $l_{i-1} = 0$ or $i = 1$, and $l_{i+s} = 0$ or $i + s = p + 1$. We denote by $I(c)$ the number of intervals of c, counted with multiplicity. For example, if $c = 11101110010101$, then $I(c) = 5$ since c has two intervals of length 3 and three intervals of length 1.

We now incorporate this into the colouring: we define a new colouring by colouring (a, b) by $(c_0, c_1, c_2, c_3, c_4, c_5)$ where c_0, c_1, c_2 and c_3 are defined above, and $c_4 = J(a, b) \mod 2$, $c_5 = I(b - a) \mod 2$, with $c_4, c_5 = 0$ or 1. So if $(y_n)_{n\geq 1}$ satisfies the conditions in Theorem 3.4 for this new colouring then it has all the properties we have already established, in addition to any new properties that may be forced by the new part of the colouring.

We say that two numbers a and b, with $a < b$, have *right to left disjoint supports* if the last digit of a is at a position smaller than the position of the first digit of b.

Claim 4. *By passing to a block subsequence, we may assume that there is no $i \in \mathbb{N}$ such that both the pair y_i, y_{i+1} and the pair y_{i+1}, y_{i+2} have right to left disjoint supports.*

Proof. Assume that such an i exists. As the cases $i = 1$ and $i = 2$ are slightly different to the general case, we analyse them separately.

1. If $i = 1$, we look at the colour of the pairs $(y_1 + y_3, y_1 + y_2 + y_3 + y_4)$ and $(y_1 + y_2 + y_3, y_1 + y_2 + y_3 + y_4)$, which have to be the same colour. In particular, the value of $J \mod 2$ has to be the same. However, when we add y_2 to $y_1 + y_3$, we eliminate exactly one '2 to 1'-jump, namely the one where we moved from the last '2' in the support of y_1 to the first '1' in the support of y_2. By the disjointness of the supports, y_2 does not interact with y_1 or y_3, so indeed the value of J changes by exactly 1, a contradiction.

2. If $i = 2$, we do not necessarily have that the supports of y_1 and y_2 are right to left disjoint. However, we observe that our colouring requires that the position of the last digit of $y_1 + y_2 + y_3$ is the position of the last digit of y_3. This tells us that $y_1 + y_2 + y_3$ and y_4 have right to left disjoint supports. By replacing y_1 with $y_1 + y_2 + y_3$ and relabelling the rest of the sequence, we may assume that y_1 and y_2 have right to left disjoint supports. With this assumption, if y_2, y_3 and y_4 still have right to left disjoint supports, we see that (with this choice of block subsequence) we are back in Case 1.

3. If $i > 2$, then we look at the colour of the pairs $(y_1 + y_2 + \cdots + y_i + y_{i+2}, y_1 + y_2 + \cdots + y_{i+3})$ and $(y_1 + y_2 + \cdots + y_{i+2}, y_1 + y_2 + \cdots + y_{i+3})$. As we argued above, the position of the last digit of $y_1 + y_2 + \cdots + y_i$ is the position of the last digit of y_i. This means that $y_1 + y_2 + \cdots + y_i$, y_{i+1} and y_{i+2} have right to left disjoint supports, thus this case is analogous to Case 1.

Therefore, we cannot have 3 consecutive terms with right to left disjoint supports. □

Claim 5. *By passing to a block subsequence, we may assume that the sequence $(y_n)_{n\geq 1}$ contains no two consecutive terms with right to left disjoint supports.*

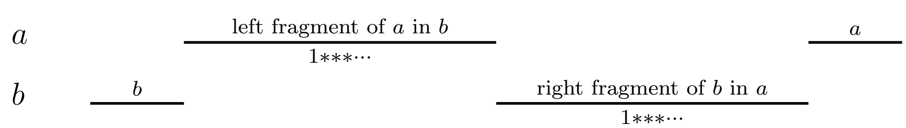
Proof. Using Claim 4, we construct the new block subsequence $(z_n)_{n\geq 1}$ inductively, with each term being either a y_i or a sum of two consecutive y_i. If y_1 and y_2 do not have right to left disjoint supports, we do not change them. If they do, then y_2 and y_3 do not have right to left disjoint supports. We then replace y_1 by $y_1 + y_2$ and relabel the sequence. Thus now the first and the second terms do not have right to left disjoint supports.

Assume we have built our block subsequence up to the i^{th} term: thus we have $(z_n)_{n=1}^{i}$, with z_i being the sum of at most two consecutive terms of our original sequence. Thus

$z_i = y_k + y_{k+1}$ or $z_i = y_{k+1}$, for some $k \in \mathbb{N}$. If y_{k+1} and y_{k+2} do not have right to left disjoint supports, then we let $z_{i+1} = y_{k+2}$. If on the other hand y_{k+1} and y_{k+2} do have right to left disjoint supports, then y_{k+2} and y_{k+3} do not have right to left disjoint supports, so we replace z_i by $z_i + y_{k+2}$ and set $z_{i+1} = y_{k+3}$. We note that since y_{k+1} and y_{k+2} have right to left disjoint supports, by Claim 4 this implies that y_k and y_{k+1} do not have right to left disjoint supports, thus, by our inductive construction we must have $z_i = y_{k+1}$. Hence, when we perform the inductive step in this case, z_i will be replaced by $y_{k+1} + y_{k+2}$, a sum of two consective terms of our original sequence. This gives us our block subsequence up to the $(i+1)^{th}$ term. □

Note that Claim 5 is true for any block subsequence of $(y_n)_{n\geq 1}$ as well. This is an immediate consequence of Claim 2 and the fact that Claim 2 is preserved by passing to any block subsequence.

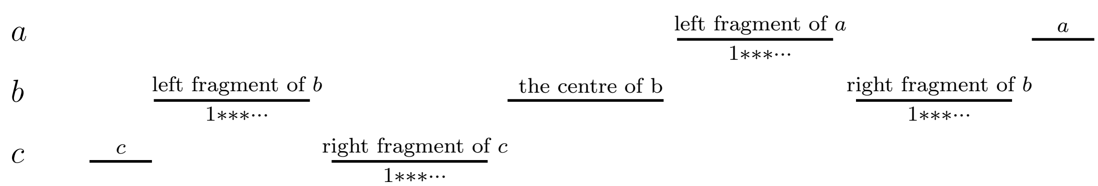
For a natural number n we denote the positions of its last and first digits by l_n and f_n, respectively. Let $a < b$ with $f_a < f_b$, $l_a < l_b$, and a and b having disjoint supports, but not right to left disjoint supports. We define a *fragment of b in a* to be a maximal binary string in $a+b$ that appears in b at the same positions, has the digit 1 at its last position, and is situated between l_a and f_b inclusive. More formally, let $a + b = r_{k_1} r_{k_1-1} \cdots r_1$, $b = b_{k_2} b_{k_2-1} \cdots b_1$ and $a = a_{k_3} a_{k_3-1} \cdots a_1$ be the binary representations of $a+b$, b and a. A binary string $s_k s_{k-1} \cdots s_1$ is called a *fragment of b in a* if there exists a positive integer t such that $k + t - 1 \leq l_a$, $t \geq f_b$, $r_{k+t-1} r_{k+t-2} \cdots r_t = b_{k+t-1} b_{k+t-2} \cdots b_t = s_k s_{k-1} \cdots s_1$, $r_{k+t-1} = b_{k+t-1} = 1$, $r_{t-1} \neq b_{t-1}$ or $t = f_b$, and there exists no binary string $w_d \cdots w_1$ with $w_d = 1$ such that $w_d \cdots w_1 = b_{k+t+d-1} \cdots b_{k+t} = r_{k+t+d-1} \cdots r_{k+t}$ and $k+t+d-1 \leq l_a$. We sometimes refer to these fragments as the *right fragments of b in a*. Similarly, a binary string $s_p s_{p-1} \cdots s_1$ is called a *fragment of a in b* if there exists a positive integer l such that $p + l - 1 \leq l_a$, $l \geq f_b$, $r_{p+l-1} r_{p+l-2} \cdots r_l = a_{p+l-1} a_{p+l-2} \cdots a_l = s_p s_{p-1} \cdots s_1$, $r_{p+l-1} = a_{p+l-1} = 1$, $r_{l-1} \neq a_{l-1}$ or $l = f_b$, and there exists no binary string $v_e \cdots v_1$ with $v_e = 1$ such that $v_e \cdots v_1 = a_{p+l+e-1} \cdots a_{p+l} = r_{p+l+e-1} \cdots r_{p+l}$ and $p + l + e - 1 \leq l_a$. We sometimes refer to these fragments as the *left fragments of a in b*. Note that there is always at least one left fragment of a in b and at least one right fragment of b in a, because a and b have disjoint supports but *not* right to left disjoint supports. The picture below illustrates this definition in the case where there is only one fragment.



Now let $a < b < c$ with the property that $l_a < l_b < l_c$, $f_a < f_b < f_c$, $l_a + 1 < f_c$, and such that they have disjoint supports, but the pairs (a, b) and (b, c) do not have right to left disjoint supports. The *fragments of b with respect to a and c* are the fragments of b in a together with the fragments of b in c. Whenever we count fragments, we count them with multiplicity – so for example, if the string 10110 occurs as a fragment in two different places, then we count this as two fragments. Note that fragments do not overlap by the maximality condition.

Let $p < r < s$ be three natural numbers with $f_p < f_r < f_s$, $l_p < l_r < l_s$, $l_p \geq f_r$, $l_r \geq f_s$ and $l_p + 1 < f_s$. We define *the centre of r with respect to p and s* to be the binary string in r situated strictly between l_p and f_s. We note that the centre of r cannot be the empty string, although, unlike a fragment in the disjoint case, it can certainly be a string of '0's.

The picture below illustrates the concepts we have just defined. The fragments are with respect to the three numbers a, b and c. For example, the centre of b is the centre of b with respect to a and c. Here there is only one left fragment of a and only one right fragment of b; in general, of course, there could be several, alternating from one to the other.



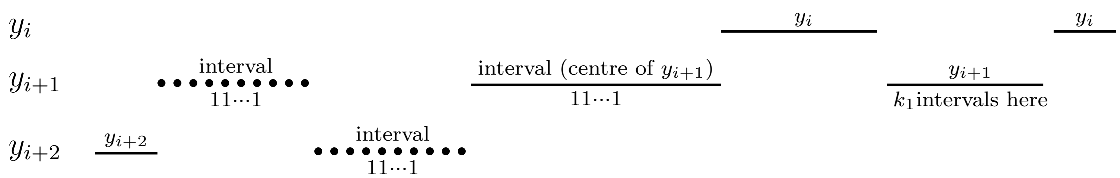
When working with a sequence $(y_n)_{n\geq 1}$, we consider the fragments or the centre of a term or of a consecutive sum of terms to be with respect to its neighbours. In other words, for any $1 < i < j$, the fragments and centre of $y_i + y_{i+1} + \cdots + y_{j-1}$ are implicitly understood to be with respect to y_{i-1} and y_j.

Let m and n be two natural numbers such that $l_m < l_n$ and $f_m < f_n$. For each i, let m_i, n_i and $(m+n)_i$ be the digits of m, n and $m+n$ at position i, respectively. When adding m and n in binary, it is convenient to refer to the minimal interval in which all binary carrying occur as the *carry region* or just the *carry*. More precisely, the carry region *starts* at the least i for which $m_i = n_i = 1$, and *stops* at position k, where k is the maximum i such that $(m+n)_i \neq m_i + n_i$. For example, if m is 100111010010 and n is 1010011011100, then the carry starts at position 4 and stops at position 9.

Claim 6. *There exists no $i \geq 1$ with the following property: y_i, y_{i+1}, y_{i+2}, y_{i+3} and y_{i+4} have pairwise disjoint supports and each of the centres of y_{i+1}, y_{i+2}, y_{i+3}, y_{i+4}, $y_{i+1} + y_{i+2}$ and $y_{i+2} + y_{i+3}$ are a string of '1's.*

Proof. Suppose for a contradiction that such an i exists. Let y_{i+1} have k_1 intervals (i.e. k_1 disjoint strings of '1's) between the position of the last digit of y_i and the position of its first digit inclusive, and k_2 intervals between the position of its last digit and the position of the first digit of y_{i+2} inclusive. Because we assumed the centre of $y_{i+1} + y_{i+2}$ is a string of '1's, we get that y_{i+1} and y_{i+2} complement each other between the position of the first digit of y_{i+2} and the position of the last digit of y_{i+1} inclusive. Therefore y_{i+2} has k_2 intervals between these 2 positions too. Similarly, if y_{i+2} has k_3 intervals between the position of its last digit and the position of the first digit of y_{i+3}, then so does y_{i+3}. Finally, let y_{i+3} have k_4 intervals between the position of its last digit and the position of the first digit of y_{i+4} inclusive.

The reader might find the diagram below helpful, where the two dotted fragments are intervals as a result of y_{i+1} and y_{i+2} complementing each other in order to have an interval as the centre of $y_{i+1} + y_{i+2}$. In the example below we have $k_2 = 1$, and only one right fragment of y_{i+1} in y_i that contains k_1 fragments. The number k_1 does not depend on the number of such fragments: it is the sum of the number of intervals in the fragments.



Since each centre is an interval and all numbers have disjoint supports, we get that y_{i+1} has $1 + k_1 + k_2$ intervals, y_{i+2} has $1 + k_2 + k_3$ intervals, y_{i+3} has $1 + k_3 + k_4$ intervals, $y_{i+1} + y_{i+2}$ has $1 + k_1 + k_3$ intervals, and $y_{i+1} + y_{i+3}$ has $k_1 + k_2 + k_3 + k_4 + 2$ intervals since y_{i+1} and y_{i+3} have disjoint right to left supports that are at least one position apart.
By looking at the I value of these numbers, c_5 tells us that

$$1 + k_1 + k_2 \equiv 1 + k_2 + k_3 \equiv 1 + k_3 + k_4 \equiv 1 + k_1 + k_3 \equiv k_1 + k_2 + k_3 + k_4 + 2 \mod 2.$$

The first four equations imply that k_1, k_2, k_3 and k_4 have the same parity. Hence $k_1 + k_2 + k_3 + k_4 + 2$ is even, which implies that $k_1 + k_2 + 1$ is even, a contradiction. □

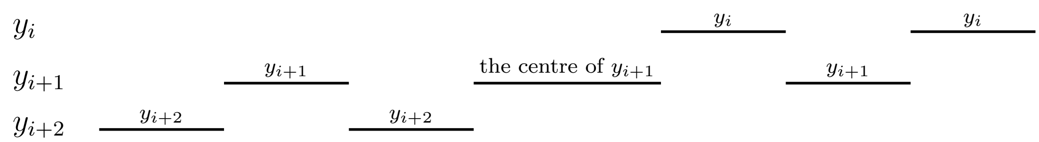
It is important to note that Claim 6 implies that our sequence $(y_n)_{n\geq1}$, and thus any of its block subsequences, cannot be of Type B. Indeed, if the sequence $(y_n)_{n\geq1}$ is of Type B, then so are all of its block subsequences, and so the first digit of $y_1+y_2+\cdots+y_k$ is at the same position as the first digit of y_{k+1} for all $k \geq 1$. If we first look at y_1, y_2 and y_3, we notice that the above conditions imply that the centre of y_2 has to be an interval, otherwise the carry in $y_1 + y_2$ would stop before the position of the first digit of y_3. Moreover, if we look at the block subsequence obtained by just replacing y_2 with $y_2 + y_3$, we must also have that the centre of $y_2 + y_3$ is an interval. This immediately implies that y_2 and y_3 must have disjoint supports, otherwise the first position they both have a '1' at will become a '0' in $y_2 + y_3$, as well as being part of the centre.

Recapping, we have shown that if $(y_n)_{n\geq1}$ is of Type B then the centre of y_2 (with respect to y_1 and y_3) is an interval, the centre of $y_2 + y_3$ (with respect to y_1 and y_4) is an interval, and y_2 and y_3 have disjoint supports. Passing to the block subsequence $y_1+y_2, y_3, y_4, \cdots$ and repeating the argument, we find that the centre of y_3 (with respect to $y_1 + y_2$ and y_4) is an interval, the centre of $y_3 + y_4$ (with respect to $y_1 + y_2$ and y_5) is an interval, and y_3 and y_4 have disjoint supports. By Claim 3, the position of the last digit of $y_1 + y_2$ is the same as that of y_2, so the centre of y_3 with respect to $y_1 + y_2$ and y_4 is the same as the centre of y_3 (with respect to y_2 and y_4), and similarly for $y_3 + y_4$. Continuing inductively, we obtain that for all $n \geq 2$ the centres of y_n and $y_n + y_{n+1}$ are intervals, and the terms y_n and y_{n+1} have disjoint supports, which contradicts Claim 6.

Therefore we can guarantees that in what follows all sequences are of Type A.

Claim 7. *There exists no $i \in \mathbb{N}$ such that y_i, y_{i+1}, y_{i+2}, $\dots$, y_{i+15} have pairwise disjoint supports.*

Proof. Suppose for a contradiction that such an i exists. We will find a block subsequence of $(y_n)_{n\geq 1}$ that will not satisfy the conditions in Theorem 3.4, a contradiction. By Claim 2 we know that if three consecutive terms have disjoint supports, then the positions between the first and the last digit of their sum inclusive can be partitioned into fragments such that each fragment corresponds to exactly one term y_i, as illustrated below.



We immediately observe that every fragment in the picture, except for the centre of y_{i+1}, has to contain the digit 1, by definition of fragments.
As we noted above, the centres can be strings of '0'. However, since the last digit of y_{i+1} is contained in the centre of $y_{i+1}+y_{i+2}$ that sits between y_i and y_{i+3}, we can replace y_{i+1} with $y_{i+1}+y_{i+2}$, y_{i+2} with $y_{i+3}+y_{y+4}$, y_{i+3} with $y_{i+5}+y_{i+6}$, $\dots$, y_{i+7} with $y_{i+13}+y_{i+14}$ and relabel the sequence. Thus, by passing to a block subsequence, we may assume that we can find 9 consecutive terms, y_k, y_{k+1}, y_{k+2}, y_{k+3}, y_{k+4}, $\dots$, y_{k+8}, such that they have disjoint supports and the centre of y_{k+1}, y_{k+2}, $\dots$, y_{k+7} all contain the digit 1.
The next step is to look at what happens with the sum $y_1+y_2+\cdots+y_k$. We know that, by disjointness, at the position of the first digit of y_{k+1}, y_k has a '0'. If the centre of y_k contains at least one '0', or $k=1$, then the sum $y_1+y_2+\cdots+y_k$ and y_{k+1} have the same fragment interaction as y_k and y_{k+1} (in other words, the fragments of y_k in y_{k+1} are the same as the fragments of $y_1+y_2+\cdots+y_k$ in y_{k+1}, and the fragments of y_{k+1} in y_k are the same as the fragments of y_{k+1} in $y_1+y_2+\cdots+y_k$) since the carry stops before the fragments start, and when $k=1$ there is no carry to consider as the above sum is just y_1. Here we used the fact that the last digit of $y_1+y_2+\cdots+y_{k-1}$ is at the same position as the last digit of y_{k-1} for $k\geq 2$.
However, Claim 6 tells us that amongst 5 consecutive terms with disjoint supports, we can always find one or a sum of two consecutive terms that does not have the centre a string of '1's (since Claim 6 is invariant under taking block subsequences). Therefore, by passing to a block subsequence or ignoring some previous terms, we can assume that the centre of y_k is not an interval, or $k=1$.
Finally, by passing to a block subsequence, we may assume that we can find 5 consecutive terms y_t, y_{t+1}, y_{t+2}, y_{t+3} and y_{t+4} such that the centres of y_{t+1}, y_{t+2} and y_{t+3} each contain at least one '1', $y_1+y_2+\cdots+y_t$ interacts with the fragments of y_{t+1} the same way y_t does, and all 5 terms have pairwise disjoint supports.

We now look at the value of J for the following pairs: $(y_1+\cdots+y_t+y_{t+3}, y_1+y_2+\cdots+y_{t+4})$, $(y_1+\cdots+y_t+y_{t+2}+y_{t+3}, y_1+y_2+\cdots+y_{t+4})$, $(y_1+\cdots+y_t+y_{t+1}+y_{t+3}, y_1+y_2+\cdots+y_{t+4})$ and $(y_1+\cdots+y_{t+3}, y_1+\cdots+y_{t+4})$. Let y_{t+1} have l_{t+1} fragments on its left and r_{t+1} fragments on its right. We define r_{t+2}, r_{t+3}, l_{t+2} and l_{t+3} similarly. We notice that $y_{t+1}+y_{t+2}$ has l_{t+2} fragments on its left and r_{t+1} fragments on its right. We also notice, by the definition of fragments, that $r_{t+2} = l_{t+1}$. If we look at the first pair above, the term $y_{t+1}+y_{t+2}$ is missing from the first sum. So the non-zero digits in its fragments will all be labelled '1'. Therefore, its right fragments will give $r_{t+1}+1$ jumps, while its left fragments will give l_{t+2} jumps. Hence, the missing term gives $r_{t+1}+l_{t+2}+1$ jumps. Similarly for the next two pairs, the missing terms give $r_{t+1}+l_{t+1}+1$ and $r_{t+2}+l_{t+2}+1$ jumps, respectively. For the last pair there is no missing term, so the jumps come from the interaction between y_{t+3} and y_{t+4}, which is identical for the other three pairs by disjointness. All the other digits in all four pairs remain unchanged.

The explanation above is summarised as follows:

$$J(y_1+\cdots+y_t+y_{t+3}, y_1+\cdots+y_{t+4}) - J(y_1+\cdots+y_{t+3}, y_1+\cdots+y_{t+4}) = r_{t+1}+l_{t+2}+1,$$
$$J(y_1+\cdots+y_t+y_{t+2}+y_{t+3}, y_1+\cdots+y_{t+4}) - J(y_1+\cdots+y_{t+3}, y_1+\cdots+y_{t+4}) = r_{t+1}+l_{t+1}+1,$$
$$J(y_1+\cdots y_t+y_{t+1}+y_{t+3}, y_1+\cdots+y_{t+4}) - J(y_1+\cdots+y_{t+3}, y_1+\cdots+y_{t+4}) = r_{t+2}+l_{t+2}+1.$$

Since our coloring asks for the J values to have same parity, we need $0 \equiv r_{t+1}+l_{t+2}+1 \equiv r_{t+1}+l_{t+1}+1 \equiv r_{t+2}+l_{t+2}+1 \mod 2$. Because $r_{t+2} = l_{t+1}$, the last equation tells us that l_{t+2} and l_{t+1} have different parities. However, by taking the difference of the first two equations, we must have that they have the same parity, a contradiction. □

Claim 8. *By passing to a block subsequence, we may assume that the sequence $(y_n)_{n\geq 1}$ contains no two consecutive terms with disjoint supports.*

Proof. The same as the proof of Claim 5. □

Claim 9. *By passing to a block subsequence, we may assume that for every $n \geq 1$ the carry in any sum where the biggest term is y_n, stops before the position of the first digit of y_{n+1}.*

Proof. As in Claim 7, it is enough to show that for every $n \geq 2$, the centre of every y_n contains at least one '0'. We will prove this by induction, replacing terms by consecutive sums and relabelling, and also bearing in mind that our initial sequence does not have any two consecutive terms with disjoint supports. Assume we have built the sequence with the desired property up to the n^{th} term. The terms y_{n+1} and y_{n+2} are consecutive terms of the original sequence, so their supports are not disjoint. If the centre of y_{n+1} contains a '0', then we have found the $(n+1)^{\text{th}}$ term. If it does not contain a '0', then we take $y_{n+1}+y_{n+2}$ to be the $(n+1)^{\text{th}}$ term. To see that this satisfies the claim, we notice that since y_{n+1} and y_{n+2} do not have disjoint supports, the first position at which both have a '1', becomes a '0' in $y_{n+1}+y_{n+2}$. As the sequence $(y_n)_{n\geq 1}$ is of Type A, we see that that position is part of the centre of $y_{n+1}+y_{n+2}$. Note that the base case $n=2$ is the same as the induction step. Thus the claim is proved. □

Note that the condition in Claim 9 is invariant under passing to a block subsequence.

We also note that the property that no two consecutive terms have disjoint supports is not necessarily preserved by passing to a block subsequence. We also observe that we have altered the sequence in Claim 8 that was assumed not to have two consecutive terms with disjoint supports, and obtained one such that the carry of any sum with biggest term y_n stops before the support of y_{n+1} begins. Further, this property is preserved by passing to a block subsequence. Therefore, starting with a sequence $(y_n)_{n\geq 1}$ with this property, we can repeat the process in Claim 7 and Claim 8 again and assume that $(y_n)_{n\geq 1}$ has both the property that the binary carry of any sum stops before the support of the next term starts, and also the property that no two consecutive terms have disjoint supports. These two properties together are invariant under our standard operation of passing to a block subsequence (noting that the property of 'consecutive terms do not have disjoint supports' is preserved because the carry resulting from any earlier additions is guaranteed to stop before the supports overlap).

For a sequence $(z_n)_{n\geq 1}$ that is of Type A and has the two properties we have stated in the previous paragraph, we define j_n, for $n \geq 2$, to be the maximum of the position of where the carry of $z_n + z_{n-1}$ stops (or equivalently any finite sum of the z_i with greatest terms z_n and z_{n-1}) and the position of the last digit of z_{n-1}. For completeness, we set j_1 to be one less than the position of the first digit of y_1. We also define the *middle* of z_n to be the (possible empty) binary string contained strictly between j_n and the position of the first digit of z_{n+1}. We call the middle of z_n *proper* if it is nonempty and it contains at least one nonzero digit. Finally, we define the *overlapping zone* of z_n and z_{n+1} to be the consecutive set of positions between the position of the first digit of z_{n+1} and j_{n+1} inclusive.

Claim 10. *By passing to a block subsequence, we may assume that the middle of y_n is proper for all $n \geq 2$.*

Proof. We prove the claim by induction. Assume that all terms up to y_{n-1}, $n \geq 3$, have a proper middle. If y_n has a proper middle, then we move on to the next term. If y_n does not have a proper middle, then $y_n + y_{n+1}$ has a proper middle with respect to y_{n-1} and y_{n+2}. This is because at position j_{n+1} in the sum $y_n + y_{n+1}$ we find the digit 1 by definition. Note that by Claim 9 the 'new j_n' (corresponding to $y_n + y_{n+1}$) is equal to the 'old j_n' (corresponding to y_n). Also, j_{n+1} is less than the position of the first digit of y_{n+2} and, by construction, $j_{n+1} > j_n$. Thus $y_n + y_{n+1}$ does have a proper middle. Therefore we take the n^{th} term to be $y_n + y_{n+1}$, and relabel the rest of the sequence, thus complete the induction step. We note that the same argument directly gives that the middle of y_2 can be assumed to be proper, which finishes the proof. □

Note that, given that a sequence satisfies the conditions of Claim 9, the conditions in Claim 10 are invariant under taking block subsequences. By earlier remarks, we may now therefore assume that out sequence satisfies Claim 8, Claim 9 and Claim 10.

Stage 3. We now add a final piece of notation. For positive integers a and b, that do not have disjoint supports, consider the positions where binary carries occur in the sum $a+b$. Those positions form some intervals which we call the *carry intervals* of a and b. For example, if $a = 110100010$ and $b = 10100111$, then the carry intervals are $\{1,2,3\}$, $\{5,6\}$ and $\{7,8,9\}$.

Let $m < n$ be two positive integers such that m and $n-m$ do not have disjoint supports. We label a position by '2' if it is not part of any carry interval of m and $n-m$, and both m and n have the digit 1 at that position. Also, we label a position by '1' if it is not part of any carry interval of m and $n-m$, and exactly one of m and n has a nonzero digit at that position. Let $\tilde{J}(m,n)$ be the number of jumps from a position labelled '2' to a position labelled '1', as we read the labels from right to left (ignoring the positions that do not have labels).

Returning to our sequence, let y'_n be the number obtained from y_n by changing all the digits in the carry intervals of y_n and y_{n-1}, and in the carry intervals of y_n and y_{n+1}, to 0, for each $n > 1$. Let also y'_1 be the number obtained from y_1 by changing all the digits in the carry interval of y_1 and y_2 to 0. Note that the new sequence $(y'_n)_{n\geq 1}$ is still increasing and of Type A as a consequence of Claim 10, and that its terms have pairwise disjoint supports.

With this in mind, our final colouring is: we colour (a,b) by $(c_0,c_1,c_2,c_3,c_4,c_5,c_6)$, where c_0, c_1, c_2, c_3, c_4, c_5 are defined above, and $c_6 = \tilde{J}(a,b) \mod 2$, with $c_6 = 0$ or 1, if a and $b-a$ do not have disjoint supports, and $c_6 = 3$ if a and $b-a$ have disjoint supports.

Let $(y_1 + y_{k_1} + \cdots + y_{k_t}, y_1 + \cdots + y_{k_{t+1}})$ be any of the pairs that have the same colour. We first observe that, by Claim 8 and Claim 9, $y_1 + y_{k_1} + \cdots + y_{k_t}$ and $y_1 + \cdots + y_{k_{t+1}} - (y_1 + y_{k_1} + \cdots + y_{k_t}) = y_2 + \cdots + y_{k_1 - 1} + \cdots + y_{k_t+1} + \cdots + y_{k_{t+1}}$ never have disjoint supports – for example, y_{k_t} and y_{k_t+1} do not have disjoint supports and, in the above sums, they are unchanged in their overlapping zone. We therefore have that $c_6 \neq 3$. Moreover, $\tilde{J}(y_1 + y_{k_1} + \cdots + y_{k_t}, y_1 + \cdots + y_{k_{t+1}}) = J(y'_1 + y'_{k_1} + \cdots + y'_{k_t}, y'_1 + \cdots + y'_{k_{t+1}})$, and so the same argument as in Claim 7 gives us a contradiction. This completes the proof of Theorem 3.4. $\square$

3.1.4 Conclusions and open problems

The colouring of $\mathbb{N}^{(2)}$ above, constructed in the previous section, involves colouring pairs. But can Theorem 3.4 be solved by a colouring that comes in a natural way just from a colouring of numbers? In particular, what happens if we promise that our colouring for Theorem 3.4 gives (a,b) a colour that depends only on the value of $a+b$?

In this case, the sum $a+b$, for a pair (a,b) as in the statement of Theorem 3.4, is exactly a sum $a_1y_1 + a_2y_2 + \cdots + a_ky_k$, where each a_i is 1 or 2 with $a_k = 1$ and $a_1 = 2$. Replacing y_1 with $2y_1$, this yields the following question.

Question 3.5. *Is it true that whenever $\mathbb{N}$ is finitely coloured, there exists a sequence $(y_n)_{n\geq 1}$ such that every sum $a_1y_1 + a_2y_2 + \cdots + a_ky_k$, for any choice of $a_i \in \{1,2\}$, with* *the only constrain that $a_1 = a_k = 1$, has the same colour?*

In general, such Ramsey-type statements, in which each coefficient can vary independently between some values, tend to be false. But here the fact that there are no 'gaps', in other words that the y_i in a given sum form an initial segment of the sequence $(y_n)_{n\geq 1}$, seems to perhaps make a difference.

We mention that if one allows a_k to be 1 or 2, then the result is easily seen to be false, because one sum will be forced to be roughly double another, which can be ruled out by a suitable colouring. And if one instead allows a_1 to be 1 or 2 then the result is also false, by considering the 2-colouring given by the least significant non-zero digit in the base 3 expansion of a number. Finally, if one allows 'gaps', so that some of the a_i are allowed to be zero, then it turns out that the result is again false, by using a colouring that examines the lengths of the jumps between successive elements of the support of a number: this is similar to the colourings considered in [15].

It is possible that Question 3.5 might be related to a problem considered by Hindman, Leader and Strauss [28]? They conjectured that whenever $\mathbb{N}$ is finitely coloured there exists a sequence $(y_n)_{n\geq 1}$ such that all finite sums of the y_i, and also all sums of the form $y_{n-1}+2y_n+y_{n+1}$, are the same colour. In each of these problems, it is the fact that the terms must be consecutive (in each sum for Question 3.5, and for the sums $y_{n-1} + 2y_n + y_{n+1}$ in the conjecture of Hindman, Leader and Strauss) that causes the difficulty. We mention that if one attempts to strengthen the conjecture of Hindman, Leader and Strauss in almost any significant way then the resulting statement turns out to be false: this is related to the 'inconsistency' of Milliken-Taylor systems (see [15] and the discussion in [28]).

Finally, returning to infinite words, what happens in Theorem 3.1 if we relax the condition that the factors u_n form an actual factorisation of our word x: what if we allow some gaps between them? Could it be that we can actually allow gaps, as long as they are bounded, and still find a bad colouring? This is a natural question to ask, in light of some variants of Hindman's theorem, such as Theorem 5.23 of [29].

Question 3.6. *Let x be an infinite word on alphabet X that is not eventually periodic. Must there exist a finite colouring of X^* such that there does not exist a sequence $u_1, u_2, \cdots$ of factors of x, with $0 \leq A_x(u_{n+1}) - B_x(u_n) \leq C$ for all n (for some C), such that all the words $u_{k_1}u_{k_2}\cdots u_{k_n}$, where $k_1 < k_2 \cdots < k_n$, have the same colour?*

Note that if we insist that $C = 0$ then this is precisely Theorem 3.1.

3.2 Monochromatic sums and products over the rationals

3.2.1 Introduction

Hindman's Theorem [26] states that whenever the natural numbers are finitely coloured there exists an infinite sequence all of whose finite sums are the same colour. By considering just powers of 2, this immediately implies the corresponding result for products: whenever the naturals are finitely coloured there exists a sequence all of whose products are the same colour. But what happens if we want to combine sums and products?

Hindman [27] showed that one cannot ask for sums and products, even just pairwise: there is a finite colouring of the naturals for which no (injective) sequence has the set of all of its pairwise sums and products monochromatic. The question of what happens if we move from the naturals to a larger space is of especial interest. Bergelson, Hindman and Leader [8] showed that if we have a finite colouring of the reals with each colour class measurable then there exist a sequence with the set of all of its finite sums and products monochromatic. (They actually proved a stronger statement: one may insist that the infinite sums are the same colour as well.) However, they also showed that there is a finite colouring of the dyadic rationals such that no sequence has all of its finite sums and products monochromatic. The questions of what happens in general for finite colourings, in the rationals or the reals, remain open.

The arguments in [8] do not extend beyond the dyadics. Our aim in this section is to go further. Let $\mathbb{Q}_{(k)}$ denote the set of rationals whose denominators (in reduced form) involve only the first k primes. Then we show that there is a finite colouring of $\mathbb{Q}_{(k)}$ such that no sequence has all of its finite sums and products monochromatic.

In fact, we strengthen this result in two ways. First of all, we insist that the number of colours does not grow with k, and more importantly we give one colouring that 'works for all $\mathbb{Q}_{(k)}$ at once', in the following sense: there is a finite colouring of the rationals such that no sequence for which the set of primes that appear in the denominators is finite has the set of its finite sums and products monochromatic. This is really made up of two separate results: one about just pairwise sums, asserting that no such *bounded* sequence can have all of its pairwise sums and products monochromatic, and the other about general finite sums, saying that no such *unbounded* sequence can have all of its finite sums and products monochromatic.

Our proofs are based on a careful analysis of the structure of addition and multiplication in $\mathbb{Q}_{(k)}$, and also on a result (Lemma 3.7 below) about colouring pairs of naturals that may be of independent interest. One application of this lemma is a new short proof of the result of Hindman mentioned above, about pairwise sums and products in the naturals.

We also prove various other related results. For example, we give a finite colouring of the reals such that no sequence that is bounded and bounded away from zero can have its pairwise sums and products monochromatic.

The plan of the section is as follows. In Subsection 3.2.2 we state and prove our lemma about colouring pairs of naturals, and use it in Subsection 3.2.3 to give a new proof of the result about pairwise sums and products in the naturals. In Subsection 3.2.4 we give the above result about the reals, which we then build on in Subsection 3.2.5 to prove the statement about pairwise sums and products in bounded sequences. Amusingly, it is not entirely clear that the colouring in Subsection 3.2.5 does not prevent monochromatic finite sums and products from *every* sequence in the rationals, and so we digress in Subsection 3.2.6 to exhibit such a sequence for this colouring. Finally in Subsection 3.2.7 we construct a colouring of the rationals such that if a sequence has the set of its finite sums and products monochromatic and the set of primes that appear in the denominators of its terms is finite, then the sequence has to be bounded – together with the results of Subsection 3.2.5 this establishes the main result.

Theorem. *There exists a finite colouring of the rational numbers with the property that there exists no sequence such that the set of its finite sums and products is monochromatic and the set of primes that divide the denominators of its terms is finite.*

Our notation is standard. We restrict our attention to the positive rationals and the positive reals (which we write as $\mathbb{Q}^+$ and $\mathbb{R}^+$ respectively), since in all situations either it would be impossible to use negative values (for example because the sums are negative but the products are positive) or because, if say we are dealing only with sums, then any colouring of the positive values could be reflected, using new colours, to the negative values. Throughout this section $\mathbb{N}$ is the set of positive integers.

We end this introduction by mentioning that in the case of finite sequences very little is known. The question of whether or not in every finite colouring of the naturals there exist two (distinct) numbers that, together with their sum and product, all have the same colour, remains tantalisingly open. Moreira [43] showed that we may find x and y such that all of $x, x+y, xy$ have the same colour, and in the rationals Bowen and Sabok [11] showed that we can indeed find the full set $x, y, x+y, xy$. But for example for sums and products from a set of size three or more almost nothing is known.

3.2.2 Some useful lemmas

In this subsection we prove the lemma mentioned above that we will make use of several times (Lemma 3.7). We will also need two slight variants of it, namely Lemma 3.8 and Lemma 3.9.

Lemma 3.7. *There exists a finite colouring Φ of $\mathbb{N}^{(2)} = \{(a,b) \in \mathbb{N}\times\mathbb{N} : a < b\}$ such that we cannot find two strictly increasing sequences of naturals, $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, such that $a_i < b_i$ for every i and $\{(a_n + a_m, b_n + b_m) : n < m\} \cup \{(a_n, b_m) : n < m\}$ is monochromatic.*

The way this will be of use to us is, roughly speaking, as follows. Suppose that we are trying to show that a certain kind of sequence cannot have its pairwise sums and products monochromatic (in the sense that there is a colouring that prevents this).

Then it is enough to find two 'parameters' a_n and b_n so that when we multiply two terms $n < m$ of the sequence we have that $a_{n \cdot m} = a_n + a_m$ and $b_{n \cdot m} = b_n + b_m$, but when we add them we have $a_{n+m} = a_n$ and $b_{n+m} = b_m$.

Before starting the proof, we need a little notation. When a natural number is written in binary we call the rightmost 1 the 'last digit' of the number (the end), and the leftmost 1 the 'first digit' of the number (the start). So for example the number 10001010 has start 7 and end 1. Also, we say that natural numbers a and b are 'right to left disjoint' if the end of b is greater than the start of a.

Proof. We colour a pair (a, b) by $(c_1, c_2, c_3, c_4, c_5)$, where c_1 is the position of the last digit of a mod 2 , c_2 is the position of the last digit of b mod 2 , and c_3 and c_4 are the digits immediately to the left of the last digits of a and b respectively. Finally c_5 is 0 if the supports of a and b are right to left disjoint, and 1 otherwise.

Suppose for a contradiction that we can find two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ as given in the statement of the lemma. Assume that for some $n < m$, a_n and a_m end at the same position. Say that position is i. Because (a_n, b_m) and (a_m, b_{m+1}) have to have the same colour, it follows that a_n and a_m have the same last 2 digits. This implies that the position of the last digit of $a_n + a_m$ is $i + 1$. On the other hand (a_n, b_m) and $(a_n + a_m, b_n + b_m)$ must have the same colour, but they have a different c_1, a contradiction. Therefore we know that all a_n have to end at different positions. By passing to subsequences, we may assume that the a_n have pairwise right to left disjoint supports.

Since (a_n, b_m) and (a_{n-1}, b_n) have the same colour, the same argument as above shows that for any $1 < n < m$, b_n and b_m must end at different positions. Thus by passing to subsequences we may assume that both a_n have right to left disjoint supports and b_n have right to left disjoint supports.
Finally, we can choose n large enough that a_1 and b_n have right to left disjoint supports and b_1 and a_n have right to left disjoint supports. Thus $c_5 = 0$ for the pair (a_1, b_n), but $c_5 = 1$ for the pair $(a_1 + a_n, b_1 + b_n)$ (as the right-hand side starts before the left-hand side finishes), a contradiction. □

We will also need two slight variants of this lemma.

Lemma 3.8. *There exists a finite colouring Ψ of $\mathbb{N}^{(2)}$ such that we cannot find two strictly increasing sequences of naturals, $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, such that $a_i < b_i$ for every i and $\{(a_n + a_m + 1, b_n + b_m) : n < m\} \cup \{(a_n, b_m) : n < m\}$ is monochromatic.*

Proof. Let Φ be the colouring in Lemma 3.7. Define Ψ by $\Psi(a, b) = \Phi(a, b + 1)$. Suppose we can find sequences $(a_n)_{n\geq 1}$ and $(b_n)_{\geq 1}$ with the above properties. Let $d_n = b_n + 1$. Then for $n < m$ we have $\Phi(a_n, d_m) = \Phi(a_n, b_m + 1) = \Psi(a_n, b_m)$ and $\Phi(a_n + a_m, d_n + d_m) = \Phi(a_n + a_m, b_n + b_m + 2) = \Psi(a_n + a_m, b_n + b_m + 1)$, contradicting Lemma 3.7. □

The next lemma is proved in a completely analogous manner; we omit the proof.

Lemma 3.9. *There exists a finite colouring Ψ' of $\mathbb{N}^{(2)}$ such that we cannot find two strictly increasing sequences of natural numbers, $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, such that $a_i < b_i$ for every $i \geq 1$ and $\{(a_n + a_m - 1, b_n + b_m) : n < m\} \cup \{(a_n, b_m) : n < m\}$ is monochromatic.* □

Finally, we note a simple fact that we will use repeatedly.

Lemma 3.10. *There exists a finite colouring $\varphi : \mathbb{Z} \to \{0, 1\}$ such that $\varphi(k+1) \neq \varphi(2k)$ and $\varphi(k+1) \neq \varphi(2k+1)$ for all $k \notin \{0, 1\}$, and $\varphi(0) \neq \varphi(1)$ and $\varphi(2) \neq \varphi(3)$.*

Proof. We build φ inductively. Let $\varphi(0) = \varphi(2) = 0$ and $\varphi(1) = \varphi(3) = 1$. We now assume that $l \leq -1$, $k \geq 2$ and that φ has been defined on $\{2l+2, 2l+3, \cdots, 2k-1\}$. Since $0 < k+1 \leq 2k-1$, $\varphi(k+1)$ is defined, thus we set $\varphi(2k) = \varphi(2k+1) = 1-\varphi(k+1)$. Similarly, since $2l + 2 \leq l + 1 \leq 0$, $\varphi(l + 1)$ is defined, so we set $\varphi(2l) = \varphi(2l + 1) = 1 - \varphi(l + 1)$, which finishes the induction step. □

3.2.3 Colouring the naturals

To illustrate the usefulness of Lemma 3.7, we use it here to give a short proof of the result of Hindman [27] about pairwise sums and products in the naturals. Because of the use of Lemma 3.7, what we are really doing is analysing the positions of the digits in binary of the numbers that are themselves the positions of the digits in binary of the terms of the sequence.

For a natural number a, we write $e_1(a)$ for the end of a, or the position of the rightmost significant digit in its binary expansion (the subscript is because later we will be looking at non-binary bases) and $s_1(a)$ for the start of a, or the position of the leftmost significant digit in its binary expansion. We also write g_a for the difference between the positions of the two most significant 1s of a in binary, and call it the 'gap' or 'left gap' of a. Thus for example 10001010 has gap 4.

Theorem 3.11. *There exists a finite colouring θ of $\mathbb{N}$ such that there is no injective sequence $(x_n)_{n\geq 1}$ of natural numbers with the property that all numbers $x_n + x_m$ and $x_n x_m$ for all $1 \leq n < m$ have the same colour.*

Proof. We begin by extending the colouring Φ from Lemma 3.7 to a colouring of $(\mathbb{N} \cup \{0\}) \times (\mathbb{N} \cup \{0\})$ by setting $\Phi(a, b)$ to be 0 if $a = 0$ or $b = 0$ or $a \geq b$. Now let a be a natural number. We define

$$\theta(a) = (p_a, e_1(a) \bmod 2, g_a \bmod 2, \Phi(e_1(a), s_1(a)), \Phi(e_1(a), s_1(a) + 1), \varphi((e_1(a)), t_a)$$

where p_a is 1 if a is a power of 2 and 0 otherwise, and $t_a = 0$ if $g_a = 1$ and 1 otherwise. Observe that φ ensures that there are no two numbers a and b of the same colour such that their end positions are $i + 1$ and $2i$ respectively, for some $i \neq 1$.

Suppose for a contradiction that there exists a strictly increasing sequence $(x_n)_{n\geq 1}$ such that all pairwise sums and products have the same colour with respect to θ. We observe that the first component of the colouring tells us that we cannot have two distinct powers of 2 in our sequence, and so we may assume that no term is a power of 2. Let a_n be the position of the last digit of x_n (i.e. $a_n = e_1(x_n)$). Note that the position of the last digit of $x_n x_m$ is $a_n + a_m$. Similarly, let b_n be the position of the first digit of x_n (i.e. $b_n = s_1(x_n)$). We know that there will either be infinitely many x_n such that $x_n < 2^{b_n}\sqrt{2}$, or infinitely many x_n such that $x_n > 2^{b_n}\sqrt{2}$. By passing to a subsequence we may assume that either $x_n < 2^{b_n}\sqrt{2}$ for all n, or $x_n > 2^{b_n}\sqrt{2}$ for all n. In the first case, the position of the first digit of $x_n x_m$ is $b_n + b_m$, while in the second case it is $b_n + b_m + 1$.

Assume first that all elements of the sequence end at position 1. We either have infinitely many terms with the same gap, or infinitely many terms with pairwise distinct gaps. If the latter is true we may assume that $(x_n)_{n\geq 1}$ has pairwise distinct gaps. Therefore we can find two m and n such that $x_n = 2 + 2^i + \cdots$ and $x_m = 2 + 2^j + \cdots$ where $2 < i < j$. In this case the gap of the sum is $i - 2$, while the gap of the product is $i - 1$, a contradiction. Therefore we may assume that all x_n end at position 1 and they have the same gap g'.

If $g' > 1$ then by the pigeonhole principle (and passing to a subsequence) we may assume that all terms have the same digit in position $g' + 2$. Now it is easy to see that the sum of any two terms has gap g', while the product has gap $g' + 1$, a contradiction. Hence we must have $g' = 1$.

In other words, we may assume that all terms end $2+2^2+\cdots$, and by the pigeonhole principle we may further assume that the digit in position 3 is the same for all terms. A simple computation shows that the sum of any two terms has gap 1, while the product does not, a contradiction.

This shows that we must have infinitely many terms that do not end at position 1. Then, by passing to a subsequence, we may assume that no term of the sequence ends at position 1. If two terms x_n and x_m end at the same position, say $i \neq 1$, then they cannot have the same gap. Indeed, if that were the case, the position of the last digit of $x_n + x_m$ is $i+1$, while the position of the last digit of $x_n x_m$ is $2i$, a contradiction. Thus we have $x_n = 2^i + 2^{i+k_1} + \cdots$ and $x_m = 2^i + 2^{i+k_2} \cdots$ for some $0 < k_1 < k_2$ (without loss of generality). The gap of the product is k_1. If $k_1 \neq 1$ then the gap of the sum is $k_1 - 1$, a contradiction. But among any three terms that have the same end positions (and thus different gaps), we must always have two with gaps not equal to 1. In other words, for any end position there are at most two terms that end there. By passing to a subsequence we may assume that the terms have right to left disjoint supports.

To sum up, by passing to a subsequence, we may assume that the terms x_n are strictly increasing and have pairwise left to right disjoint supports. Thus the start and end positions form two increasing sequences, and since for $n < m$ we have $e_1(x_n+x_m) = a_n$ and $s_1(x_n + x_m) = b_m$, we are done by Lemma 3.7 or Lemma 3.8. □

3.2.4 Colouring the reals

In this subsection we prove the result about the reals mentioned in the introduction, that there is a colouring for which no sequence that is bounded and bounded away from zero has all of its pairwise sums and products monochromatic. There is a fair amount of notation, which will also be used in later sections, but all of it is very simple and self-explanatory. The aim is to analyse carefully how the 'starting' few 1s (in binary) of the numbers behave, and especially how close together those first few 1s are.

For $x \in \mathbb{R}^+$, we define $a(x)$ to be the unique integer such that $2^{a(x)} \leq x < 2^{a(x)+1}$. Moreover, for $x \in \mathbb{R}^+ \setminus \{2^k : k \in \mathbb{Z}\}$, we define $b(x) = a(x - 2^{a(x)})$. In other words, for x not an integer power of 2, $b(x)$ is the unique integer such that $2^{a(x)} + 2^{b(x)} \leq x < 2^{a(x)} + 2^{b(x)+1}$. For $x \in \mathbb{R}^+ \setminus \{2^k : k \in \mathbb{Z}\}$ we also define $c(x)$ to be the unique integer such that $2^{a(x)+1} - 2^{c(x)+1} \leq x < 2^{a(x)+1} - 2^{c(x)}$.

Note that if $x \in \mathbb{N}$ then $a(x)$ is what we called the start of x in Subsections 3.2.2 and 3.2.3. If x is not a power of 2, then $b(x)$ is the position of the second most significant digit 1 in the base 2 expansion of x, and $c(x)$ is the position of the leftmost zero when x is written in binary without leading 0s.

We now define $A_0 = \{x \in \mathbb{R}^+ : 2^{a(x)} < x < 2^{a(x)+\frac{1}{2}}\}$, $A_1 = \{x \in \mathbb{R}^+ : 2^{a(x)+\frac{1}{2}} < x < 2^{a(x)+1}\}$, $C_1 = \{2^k : k \in \mathbb{Z}\}$ and $C_2 = \{2^{k+\frac{1}{2}} : k \in \mathbb{Z}\}$. We observe that A_0, A_1, C_1, and C_2 are pairwise disjoint sets that partition $\mathbb{R}^+$, and A_0 and A_1 are open in $\mathbb{R}^+$, while C_1 and C_2 are countable.

Recalling the colouring φ in Lemma 3.10, define $G_i = \{x \in \mathbb{R}^+ \setminus C_1 : \varphi(a(x)) = i\}$ for $i \in \{0, 1\}$. Since G_i is the union of all the open intervals $(2^k, 2^{k+1})$ where $k \in \mathbb{Z}$ and $\varphi(k) = i$, we see that G_i is open in $\mathbb{R}^+$. Moreover, C_1, G_0 and G_1 also form a partition of the positive reals, where C_1 is countable and G_0 and G_1 are open.

Next we define $C_3 = \{2^k + 2^l : k, l \in \mathbb{Z} \text{ and } l < k\}$, and $H_i = \{x \in \mathbb{R}^+ \setminus (C_1 \cup C_3) : a(x) - b(x) \equiv i \mod 3\}$ for $i \in \{0, 1, 2\}$. By writing H_i as the union of all open intervals $(2^k + 2^l, 2^k + 2^{l+1})$ where $k, l \in \mathbb{Z}$, $l < k$ and $k - l \equiv i \mod 3$, we have that H_i is open in $\mathbb{R}^+$ for $i \in \{0, 1, 2\}$. As before, C_1, C_3, H_0, H_1 and H_2 partition the positive reals.

Define now $C_4 = \{2^k - 2^l : k, l \in \mathbb{Z} \text{ and } l < k\}$, and $J_i = \{x \in \mathbb{R}^+ \setminus C_4 : a(x) - c(x) \equiv i \mod 3\}$ for $i \in \{0, 1, 2\}$. Note that $C_1 \subset C_4$ and $C_3 \cap C_4 = \{2^{k+1} + 2^k : k \in \mathbb{Z}\} \neq \emptyset$. By writing J_i as the union of all open intervals $(2^{k+1} - 2^{l+1}, 2^{k+1} - 2^l)$ where $k, l \in \mathbb{Z}$, $l < k$ and $k - l \equiv i \mod 3$, we see that J_i is open in $\mathbb{R}^+$ for $i \in \{0, 1, 2\}$. Also, C_4, J_0, J_1 and J_2 partition the positive reals.

Finally, we define $C_5 = \{2^{k+1}(1 - 2^{l-k})^{\frac{1}{2}} : k, l \in \mathbb{Z} \text{ and } l < k\}$, and $B_i = \{x \in \mathbb{R}^+ \setminus (C_1 \cup C_5) : x < 2^{a(x)+1}(1 - 2^{c(x)-a(x)})^{\frac{1}{2}} \text{ and } a(x) - c(x) \equiv i \mod 3, \text{ or } x > 2^{a(x)+1}(1 - 2^{c(x)-a(x)})^{\frac{1}{2}} \text{ and } a(x) - c(x) \equiv i + 1 \mod 3\}$ for $i \in \{0, 1, 2\}$. Note that $C_2 \subset C_5$. Since B_i can be written as the union of all the sets of the form $(2^{k+1} - 2^{l+1}, 2^{k+1}(1 - 2^{l-k})^{\frac{1}{2}})$ where $l, k \in \mathbb{Z}$, $l < k$ and $k - l \equiv i \mod 3$, and all the sets of the form $(2^{k+1}(1 - 2^{l-k})^{\frac{1}{2}}, 2^{k+1} - 2^l)$ where $k, l \in \mathbb{Z}$, $l < k$ and $k - l \equiv i + 1 \mod 3$, we see that B_i is open in $\mathbb{R}^+$ for all $i \in \{0, 1, 2\}$. Also, C_1, C_5, B_0, B_1 and B_2 partition the positive reals.

We are now ready to define our colouring ν. To start with, we let C_1, C_2, $C_3 \setminus C_4$,

$C_4 \setminus C_1$ and $C_5 \setminus C_2$ be five colour classes of ν. If $x \in \mathbb{R}^+ \setminus (C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5)$, then we set $\nu(x) = (w_1, w_2, w_3, w_4, w_5)$, where $w_i = i$ if $x \in A_i$, $w_2 = i$ if $x \in G_i$, $w_3 = i$ if $x \in H_i$, $w_4 = i$ if $x \in J_i$ and $w_5 = i$ if $x \in B_i$.

It is important to note that, with the exception of the five countable classes defined first, the colour classes of ν are open (as a consequence of $C_1 \cup \cdots \cup C_5$ being closed).

Theorem 3.12. *Let $(x_n)_{n\geq 1}$ be an injective sequence of positive reals with the property that all numbers $x_n + x_m$ and $x_n x_m$ for all $1 \leq n < m$ have the same colour. Then $(x_n)_{n\geq 1}$ cannot be bounded and bounded away from zero.*

Proof. The colour class of the pairwise sums and products of $(x_n)_{n\geq 1}$ cannot be any of C_1, C_2, $C_3 \setminus C_4$, $C_4 \setminus C_1$ and $C_5 \setminus C_2$. Indeed, the proofs for C_1 and C_2 are an easy exercise. The proofs for C_3 and C_5, while routine, are lengthy, and so are presented in the Appendix at the end of this section. The proof for C_4 is very similar to the one for C_3, and so we omit it. Therefore $x_n + x_m$ and $x_n x_m$ are all in $\mathbb{R}^+ \setminus (C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5)$ for all $n < m$.

Suppose for a contradiction that $(x_n)_{n\geq 1}$ is bounded and bounded away from zero. This immediately implies that the sequence of integers $(a(x_n))_{n\geq 1}$ is bounded. By passing to a subsequence, we may assume that $(a(x_n))_{n\geq 1}$ is constant, and thus equal to some fixed integer k. Moreover, by the pigeonhole principle and passing to another subsequence, we may assume that either $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all n or $2^{a(x_n)+\frac{1}{2}} \leq x_n$ for all n.

Let n and m be two distinct natural numbers. Since $a(x_n) = a(x_m) = k$ we have that $2^{k+1} < x_n + x_m < 2^{k+2}$ and $2^{2k} < x_n x_m < 2^{2k+2}$. This implies that $a(x_n + x_m) = k + 1$ and that either $a(x_n x_m) = 2k$, or $a(x_n x_m) = 2k + 1$. Let $i \in \{0, 1\}$ be such that $x_n + x_m \in G_i$ and $x_n x_m \in G_i$. In other words we must have $\varphi(a(x_n + x_m)) = \varphi(a(x_n x_m))$, which implies that $\varphi(k + 1) = \varphi(2k)$ or $\varphi(k + 1) = \varphi(2k + 1)$, and thus $k \in \{0, 1\}$.

We consider first the case when $k = 1$. This means that $2 < a_n < 4$ and $a(x_n x_m) = a(x_n + x_m) = 2$ for all distinct naturals n and m. Hence we must have $2 < x_n < 2^{\frac{3}{2}}$ for all n.

We first assume that the integer sequence $(b(x_n))_{n\geq 1}$ is bounded. By passing to a subsequence, we may assume that $(b(x_n))_{n\geq 1}$ is constant and equal to a fixed integer $l < k = 1$. Since $x_n \geq 2^{a(x_n)} + 2^{b(x_n)}$ for all n, we cannot have $l = 0$, or else $x_n \geq 2 + 1 = 3 > 2^{\frac{3}{2}}$, and so $l \leq -1$.

Let m and n be two distinct natural numbers. By the above we have that $x_n = 2 + 2^l + u$ and $x_m = 2 + 2^l + v$ for some $0 \leq u, v < 2^l$. Next we have that $x_n + x_m = 4 + 2^{l+1} + u + v$ and $0 \leq u + v < 2^{l+1}$, thus $b(x_n + x_m) = l + 1$, and consequently $a(x_n + x_m) - b(x_n + x_m) = 2 - (l + 1) = 1 - l$.

On the other hand, $x_n x_m = 4 + 2^{l+2} + (2^l + 2)(u + v) + uv + 2^{2l}$. The sum of terms involving the variables u and v can be bounded as follows: $(2^l + 2)(u + v) + uv + 2^{2l} < (2^l + 2)2^{l+1} + 2^{2l} + 2^{2l} = 2^{2l+2} + 2^{l+2}$. Therefore we trivially have $4 + 2^{l+2} < x_n x_m$ and $x_n x_m < 4 + 2^{l+2} + 2^{2l+2} + 2^{l+2} = 4 + 2^{l+3} + 2^{2l+2} < 4 + 2^{l+4}$. This tells us that either $b(x_n x_m) = l + 2$, or $b(x_n x_m) = l + 3$, thus either $a(x_n x_m) - b(x_n x_m) = -l$, or

$a(x_nx_m)-b(x_nx_m) = -l-1$. In both cases $a(x_nx_m)-b(x_nx_m)$ and $a(x_n+x_m)-b(x_n+x_m)$ are not congruent $\mod 3$, a contradiction.

Therefore we must have that $(b(x_n))_{n\geq 1}$ is unbounded and, by passing to a subsequence, we may assume that $(b(x_n))_{n\geq 1}$ is strictly decreasing.

Let n be a natural number and $l = b(x_n)$. We know that there exists u such that $0 \leq u < 2^l$ and $x_n = 2+2^l+u$. We now pick an integer $s < l$ such that $u+2^s < 2^l$, and then a natural number m such that $b(x_m) < s$. Let $t = b(x_m)$ and $x_m = 2+2^t+v$, where $0 \leq v < 2^t$. It follows that $x_n + x_m = 4+2^l+u+2^t+v$. By all the above we have that $u+2^t+v < u+2^{t+1} \leq u+2^s < 2^l$. Thus $b(x_n+x_m) = l$ and $a(x_n+x_m)-b(x_n+x_m) = 2-l$.

Finally, since $2+2^l \leq x_n < 2+2^{l+1}$ and $2+2^t \leq x_m < 2+2^{t+1}$, we first have that $4+2^{l+1} < 4+2^{l+1}+2^{t+1}+2^{l+t} \leq x_nx_m$. Moreover, $x_nx_m < 4+2^{l+2}+2^{t+2}+2^{l+t+2} < 4+2^{l+3}$. Putting these together we see that either $b(x_nx_m) = l+1$ or $b(x_nx_m) = l+2$. Thus either $a(x_nx_m)-b(x_nx_m) = 1-l$, or $a(x_nx_m)-b(x_nx_m) = -l$, neither of which is congruent to $a(x_n+x_m)-b(x_n+x_m) \mod 3$, a contradiction. This concludes the case when $k = 1$.

We must therefore have $k = 0$. In other words $a(x_n) = 0$, $2^{\frac{1}{2}} \leq x_n < 2$, and $a(x_n+x_m) = a(x_nx_m) = 1$ for all distinct natural numbers n and m. Since there is at most one n such that $x_n = 2^{\frac{1}{2}}$, by passing to a subsequence we may assume that $2^{\frac{1}{2}} < x_n < 2$ for all n.

We observe that if $2^{\frac{1}{2}} < x_n < \frac{3}{2}$ and $2^{\frac{1}{2}} < x_n < \frac{3}{2}$ for two distinct m and n, then $2\cdot 2^{\frac{1}{2}} = 2^{\frac{3}{2}} < x_n+x_m < 3$, thus $x_n+x_m \in A_1$, while $2 < x_nx_m < 9/4 < 2^{\frac{3}{2}}$, so $x_nx_m \in A_0$, a contradiction. Therefore, by passing to a subsequence, we may assume that $\frac{3}{2} \leq x_n < 2$. This immediately implies that $x_n \geq 2^1-2^{-1} = 2^{a(x_n)+1}-2^{-2+1}$, and so $c(x_n) \leq -2$ for all n.

We first assume that the integer sequence $(c(x_n))_{n\geq 1}$ is bounded. Thus by passing to a subsequence we may assume that it is constant and equal to a fixed integer $l \leq -2$. Let m and n be two distinct natural numbers. Then we have that $2-2^{l+1} \leq x_n < 2-2^l$ and $2-2^{l+1} \leq x_m < 2-2^l$. Summing the above we obtain $4-2^{l+2} \leq x_n+x_m < 4-2^{l+1}$, and thus $c(x_n+x_m) = l+1$ and consequently $a(x_n+x_m)-c(x_n+x_m) = -l$. On the other hand, multiplying the above gives $4-2^{l+3}+2^{2l+2} \leq x_nx_m < 4-2^{l+2}+2^{2l}$. The lower bound is trivially greater than $4-2^{l+3}$, and $2^{l+2}-2^{2l} > 2^{l+1}$, so $4-2^{l+2}+2^{2l} < 4-2^{l+1}$. This means that $c(x_nx_m)$ is either $l+1$ or $l+2$. Since $c(x_nx_m) = l+2$ implies $a(x_nx_m)-c(x_nx_m) = -l-1$ which is not congruent to $-l = a(x_n+x_m)-c(x_n+x_m) \mod 3$, we conclude that $c(x_nx_m) = l+1$ for all $n \neq m$, which can be written as $4-2^{l+2} \leq x_nx_m < 4-2^{l+1}$ for all $n \neq m$.

Observe that if $x_n < 2(1-2^l)^{\frac{1}{2}}$ and $x_m < 2(2-2^l)^{\frac{1}{2}}$ for two distinct positive integers m and n, then $x_nx_m < 4(1-2^l) = 4-2^{l+2}$, which contradicts $c(x_nx_m) = l+1$. Therefore, by passing to a subsequence, we may assume that $x_n \geq 2(1-2^l)^{\frac{1}{2}}$ for all n.

Let $n \neq m$ be two natural numbers. Then $x_n+x_m \geq 4(1-2^l)^{\frac{1}{2}} = 4(1-2^{c(x_n+x_m)-a(x_n+x_m)})^{\frac{1}{2}}$. Let $i \in \{0,1,2\}$ such that $-l = a(x_n+x_m)-c(x_n+x_m) \equiv i+1 \mod 3$. This means that $x_n+x_m \in B_i$, and consequently $x_nx_m \in B_i$. On the other hand, since $x_n < 2-2^l$ and $x_m < 2-2^l$, it is easy to check that the product $x_nx_m <$

$4-2^{l+2}+2^{2l} = 4(1-2^l+2^{2l-2}) < 4(1-2^l)^{\frac{1}{2}}$. Since $a(x_nx_m)-c(x_nx_m) = 1-(l+1) = -l$ we have that $x_nx_m < 4(1-2^{c(x_nx_m)-a(x_nx_m)})^{\frac{1}{2}}$, and thus $x_nx_m \in B_j$ where $j \in \{0,1,2\}$ and $j \equiv -l \mod 3 \equiv i+1 \mod 3$. But this is a contradiction since it implies that $i \neq j$, so that the sum and the product are in different B-classes.

Therefore we must have that the sequence $(c(x_n))_{n\geq 1}$ is unbounded and, by passing to a subsequence, we may assume that it is strictly decreasing.

Let us first assume that there exist $n < m$ such that $x_n = 2 - 2^{c(x_n)+1}$ and $x_m = 2-2^{c(x_m)+1}$. Then we have that $x_n+x_m = 4-2^{c(x_n)+1}-2^{c(x_m)+1}$, and since $c(m) < c(n)$ we get that $4-2^{c(x_n)+2} < x_n+x_m < 4-2^{c(x_n)+1}$, so $c(x_n+x_m) = c(x_n)+1$ and consequently $a(x_n+x_m)-c(x_n+x_m) = -c(x_n)$. On the other hand, $x_nx_m = 4-2^{c(x_n)+2}-2^{c(x_m)+2}+2^{c(x_n)+c(x_m)+2} = 4-2^{c(x_n)+2}-2^{c(x_m)+2}(1-2^{c(x_n)})$. Hence we have that $4-2^{c(x_n)+3} < 4-2^{c(x_n)+2}-2^{c(x_m)+2} < x_nx_m < 4-2^{c(x_n)+2}$. It follows that $c(x_nx_m) = c(x_n)+2$, so$a(x_nx_m)-c(x_nx_m) = -c(x_n)-1$, a contradiction.

Finally, after passing to a subsequence, we may assume that for every n there exists u_n such that $0 < u_n < 2^{c(x_n)}$ and $x_n = 2-2^{c(x_n)+1}+u_n$. Let n be a natural number and let $s \in \mathbb{Z}$ be such that $u_n + 2^s < 2^{c(x_n)}$. Since the sequence $(c(x_n))_{n\geq 1}$ is strictly decreasing and unbounded, we can find $m > n$ such that $c(x_m) < \min\{s, \log_2 u_n - 1\}$. It then follows that $x_n+x_m = 4-2^{c(x_n)+1}+u_n-2^{c(x_m)+1}+u_m$. We observe that $0 < u_n-2^{c(x_m)+1}+u_m < u_n-2^{c(x_m)+1}+2^{c(x_m)} < u_n+2^{c(x_m)} < u_n+2^s < 2^{c(x_n)}$. This means that $4-2^{c(x_n)+1} < x_n+x_m < 4-2^{c(x_n)+1}+2^{c(x_n)} = 4-2^{c(x_n)}$, and so $c(x_n+x_m) = c(x_n)$ and consequently $a(x_n+x_m)-c(x_n+x_m) = 1-c(x_n)$.

We are now going to analyse the product x_nx_m. We have that $2-2^{c(x_n)+1} < x_n < 2-2^{c(x_n)}$ and $2-2^{c(x_m)+1} < x_m < 2-2^{c(x_m)}$. By multiplying the above inequalities we obtain that $4-2^{c(x_n)+2}-2^{c(x_m)+2}+2^{c(x_n)+c(x_m)+2} < x_nx_m$ and $x_nx_m < 4-2^{c(x_n)+1}-2^{c(x_m)+1}+2^{c(x_n)+c(x_m)}$. We consider these two inequalities separately.

First we have that $4-2^{c(x_n)+1}-2^{c(x_m)+1}+2^{c(x_n)+c(x_m)} = 4-2^{c(x_n)+1}-2^{c(x_m)}(2-2^{c(x_n)}) < 4-2^{c(x_n)+1}$, thus $x_nx_m < 4-2^{c(x_n)+1}$.

Secondly we have that $4-2^{c(x_n)+2}-2^{c(x_m)+2}+2^{c(x_n)+c(x_m)+2} > 4-2^{c(x_n)+2}-2^{c(x_m)+2} > 4-2^{c(x_n)+3}$, since $c(x_m) < c(x_n)$.

Putting everything together we get that $4-2^{c(x_n)+3} < x_nx_m < 4-2^{c(x_n)+1}$, and thus either $c(x_nx_m) = c(x_n)+1$ or $c(x_nx_m) = c(x_n)+2$. This means that either $a(x_nx_m)-c(x_nx_m) = -c(x_n)$, or $a(x_nx_m)-c(x_nx_m) = -c(x_n)-1$, neither of which is congruent to $a(x_n+x_m)-c(x_n+x_m) = 1-c(x_n) \mod 3$, a contradiction. □

It is important to point out that the colouring ν cannot be used to rule out similar statements about sums and products from a sequence $(x_n)_{n\geq 1}$ that tends to zero. Indeed, since each colour class of ν is measurable (being either countable or open), the result of [8] tells us that there is a sequence with all of its products and all of its sums (even infinite sums) having the same colour for ν.

3.2.5 Combining an extension of θ over the rationals with ν

In this subsection we will build a colouring of the positive rationals via an 'extension' of the colouring θ that also incorporates ν. This colouring will force any bounded sequence with monochromatic pairwise sums and products to have the set of primes which divide the denominators of the terms of the sequence to be infinite.

Roughly speaking, we will be concerned with how a number ends, not just how it starts, and therefore we will be considering numbers written not in binary (of course) but rather in the smallest base for which they terminate. The analysis is considerably more complicated than it would be for binary. There is also the issue that different numbers will have different 'smallest bases', but it turns out that this will not cause too much of a problem.

Let $(p_n)_{n\geq 1}$ be the enumeration of all primes in increasing order, and $P_n = \prod_{k=1}^{n} p_k$ for all $n \in \mathbb{N}$. Let also $T_n = \mathbb{Q}_{(n)} \cap (0,1)$. In other words, T_n consists of all the rationals between 0 and 1 for which, in reduced form, the denominator does not have any p_t with $t > n$ as a factor. For completeness, define $T_0 = \emptyset$. If $x \in T_n \setminus T_{n-1}$ we may say that P_n is the 'minimal base' of x.

For $n \in \mathbb{N}$ and $x \in T_n$, we define $s_n(x)$ to be the position of the leftmost significant digit and $e_n(x)$ the position of the rightmost significant digit in the base P_n expansion of x. For example, if x has the base 6 expansion $405.00213 = 4\cdot 6^2 + 2\cdot 6^{-3} + 6^{-4} + 3\cdot 6^{-5}$ then $s_2(x) = 2$ and $e_2(x) = -5$. For $x \in \mathbb{N}$, so that $e_1(x)$ and $s_1(x)$ are the position of the rightmost significant digit and leftmost significant digit respectively in the binary expansion of x, we set $d(x)$ to be the digit in position $e_1(x)+1$. Finally, for $x, y \in \mathbb{N}$, define $g(x,y) = 0$ if $e_1(y) > s_1(x)$ and $g(x,y) = 1$ if $e_1(y) \leq s_1(x)$.

The colouring Φ of $\mathbb{N}^{(2)}$ defined previously can be rewritten as follows: $\Phi(x,y) = (e_1(x) \mod 2, e_1(y) \mod 2, d(x), d(y), g(x,y))$. We also define the colouring Ψ' of $\mathbb{N}^{(2)}$, which is very similar in spirit to the previously defined colouring Ψ, by $\Psi'(x,y) = \Phi(1,2)$ if $x = 1$, and $\Psi'(x,y) = \Phi(x-1,y)$ if $x > 1$.

We are now ready to define a colouring μ of $\mathbb{Q}$ as follows. If $x \geq 1$, let $\mu(x) = \nu(x)$. Otherwise, for any $x \in \mathbb{Q} \cap (0,1)$, there exists a unique $n \in \mathbb{N}$ such that $x \in T_n \setminus T_{n-1}$. Consequently, define

$$\mu(x) = (\nu(x), \Phi(-s_n(x), -e_n(x)), \Psi'(-s_n(x), -e_n(x))).$$

The following is what we wish to prove.

Theorem 3.13. *Let $(x_n)_{n\geq 1}$ be a bounded sequence of positive rationals such that the set $\{x_n + x_m, x_n x_m : n \neq m\}$ is monochromatic with respect to μ. Then for any $k \in \mathbb{N}$ there exist l and n such that $x_n \in T_l \setminus T_k$.*

Proof. Because the sequence $(x_n)_{n\geq 1}$ is monochromatic with respect to μ it is also monochromatic with respect to ν. Since $(x_n)_{n\geq 1}$ is bounded, Theorem 3.12 tells us that $(x_n)_{n\geq 1}$ must converge to 0, and so we may assume that all terms are less than 1.

Assume for a contradiction that there exists $k \in \mathbb{N}$ such that $x_n \in T_k$ for all $n \in \mathbb{N}$. By passing to a subsequence, we can assume that for all n $x_n \in T_t \setminus T_{t-1}$ for some $t \leq k$. In other words, the minimal base of the form P_s for x_n is P_t, for all $n \geq 1$. Since $(x_n)_{n\geq 1}$ converges to 0, $s_t(x_n)$ and $e_t(x_n)$ must tend to $-\infty$. In particular, we may assume from now on that $s_t(x_n) < -1$ for all n. Moreover, by passing to a subsequence, we may assume that the sequence is strictly decreasing and that all of its terms have pairwise left to right disjoint support – in other words, if $n < m$ then $e_t(x_n) > s_t(x_m)$. Also, by the pigeonhole principle, there exists a subsequence for which all terms have the same last digit, say $0 < d < P_t$, and by passing to that subsequence we may assume that this is the case for $(x_n)_{n\geq 1}$ itself.

Let $n < m$ be positive integers. Then, because x_n and x_m have disjoint supports in base P_t, which is their minimal base, $x_n + x_m$ also has minimal base P_t. Furthermore, $s_t(x_n + x_m) = s_t(x_n)$ and $e_t(x_n + x_m) = e_t(x_m)$. It is also easy to see that if both x_n and x_m have minimal base P_t then so does $x_n x_m$.

We note that if $x \in T_t \setminus T_{t-1}$, then $-e_t(x)$ is the smallest positive integer u such that $x(P_t)^u \in \mathbb{N}$. Clearly $x_n x_m (P_t)^{-e_t(x_n)-e_t(x_m)} \in \mathbb{N}$, and thus $e_t(x_n x_m) \geq e_t(x_n) + e_t(x_m)$.

Now suppose that there exists $k' \in \mathbb{N}$ smaller than $-e_t(x_n) - e_t(x_m)$ such that $x_n x_m (P_t)^{k'} \in \mathbb{N}$. It follows that $x_n (P_t)^{-e_t(x_n)} x_m (P_t)^{-e_t(x_m)} (P_t)^{k'+e_t(x_n)+e_t(x_m)} \in \mathbb{N}$. But $x_n (P_t)^{-e_t(x_n)} \equiv x_m (P_t)^{-e_t(x_m)} \equiv d \mod P_t$. Because the power of P_t is negative, we must have that P_t divides d^2, and since P_t is a product of distinct primes, we must in fact have that P_t divides d, a contradiction. Therefore $e_t(x_n x_m) = e_t(x_n) + e_t(x_m)$.

Finally, for $x \in T_t \setminus T_{t-1}$, $s_t(x)$ is the unique integer l such that $(P_t)^{l+1} > x \geq (P_t)^l$. By the pigeonhole principle we either have $x_n \geq \sqrt{P_t}(P_t)^{s_t(x_n)}$ for infinitely many n or $x_n < \sqrt{P_t}(P_t)^{s_t(x_n)}$ for infinitely many n. By passing to a subsequence we may assume that we are either in the first case for all n or in the second case for all n. In the first case $s_t(x_n x_m) = s_t(x_n) + s_t(x_m) + 1$ for all $m \neq n$, while in the second case $s_t(x_n x_m) = s_t(x_n) + s_t(x_m)$.

Let $a_n = -s_t(x_n) > 1$ and $b_n = -e_t(x_n) > a_n$ for all $n \in \mathbb{N}$. Note that both $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ are strictly increasing sequences of natural numbers. Then μ tells us that either $\Phi(a_n, b_m) = \Phi(a_n + a_m, b_n + b_m)$ for all $n < m$, or $\Phi(a_n - 1, b_m) = \Phi(a_n + a_m - 2, b_n + b_m)$ for all $n < m$, which contradicts Lemma 3.7 or Lemma 3.9. □

3.2.6 Exploring μ further

It turns out that for μ we can find an injective sequence with all pairwise sums and products monochromatic, and actually even all finite sums and products monochromatic. This shows that neither θ nor ν nor their product can provide a counterexample for the 'finite sums and products' problem in the set of all rationals.

We say that the sequence $(y_n)_{n\geq 1}$ is a *product subsystem* of the sequence $(x_n)_{n\geq 1}$ if there exists a sequence $(H_n)_{n\geq 1}$ of finite sets of natural numbers such that for every $n \geq 1$, $\max H_n < \min H_{n+1}$ and $y_n = \prod_{t\in H_n} x_t$.

Theorem 3.14. *There exists a sequence $(y_n)_{n\geq 1}$ in $\mathbb{Q}\cap(0,1)$ such that all of its finite sums and finite products are monochromatic with respect to μ.*

Proof. Starting with $r_1=2$, we may inductively choose an increasing sequence $(r_n)_{n\geq 1}$ of natural numbers such that for all $n\in\mathbb{N}$ we have $\displaystyle\sum_{i=1}^{n}\frac{1}{p_{r_i}}<1$. By the Finite Sums Theorem (or rather a simple corollary of it – see Corollary 5.15 in [29]) we can choose a product subsystem $(x_n)_{n\geq 1}$ of $\displaystyle\left(\frac{1}{p_{r_i}}\right)_{n\geq 1}$ such that all finite products of $(x_n)_{n\geq 1}$ are monochromatic with respect to ν – in other words, they are all members of a colour class of ν, say U.

The colouring ν of $\mathbb{R}^+$ consists of five countable classes and several classes that are open in $\mathbb{R}^+$. Recall that the countable colour classes are $C_1=\{2^k:k\in\mathbb{Z}\}$, $C_2=\{2^{k+\frac{1}{2}}:k\in\mathbb{Z}\}$, $C_3\setminus C_4=\{2^k+2^l:k,l\in\mathbb{Z}\text{ and }l<k\}\setminus C_4$, $C_4\setminus C_1=\{2^k-2^l:k,l\in\mathbb{Z}\text{ and }l<k\}\setminus C_1$, and $C_5\setminus C_2=\{2^{k+1}(1-2^{l-k})^{\frac{1}{2}}:k,l\in\mathbb{Z}\text{ and }l<k\}\setminus C_2$. It is easy to see that C_2 contains only irrational numbers. Observe also that C_5 contains only irrational numbers, because $\left(1-\frac{1}{2^n}\right)^{\frac{1}{2}}$ is irrational for any $n\in\mathbb{N}$. (Indeed, suppose $\left(1-\frac{1}{2^n}\right)^{\frac{1}{2}}=\frac{p}{q}$ for some coprime $p,q\in\mathbb{N}$; we then get that $\frac{2^n-1}{2^n}=\frac{p^2}{q^2}$, so 2^n and 2^n-1 have to be perfect squares, but no two perfect squares in $\mathbb{N}$ differ by 1.)

The classes C_1, $C_3\setminus C_4$, and $C_4\setminus C_1$ consist of rational number that have denominator (in reduced form) a power of 2, and thus none of them can be in U as no x_n has this property since $r_1=2$. Furthermore, C_2 and $C_5\setminus C_2$ consist of irrational numbers, so $C_2\neq U$ and $C_5\setminus C_2\neq U$. We conclude that U is an open colour class of ν that contains all the finite products of $(x_n)_{n\geq 1}$.

We are now going to find a subsequence $(y_n)_{n\geq 1}$ of $(x_n)_{n\geq 1}$ such that all its finite sums are in U as well. We proceed by induction. Let $y_1=x_1$. Now assume $n\geq 1$ and that we have chosen $y_1>y_2>\cdots>y_n$ such that $y_i\in\{x_j:j\in\mathbb{N}\}$ for all $1\leq i\leq n$, and that for any finite non-empty set A of $\{1,2,\cdots,n\}$ we have $\displaystyle\sum_{i\in A}y_i\in U$.

Because U is open in $\mathbb{R}^+$, we can pick $\epsilon_A>0$ such that $\displaystyle\left(\sum_{i\in A}y_i,\sum_{i\in A}y_i+\epsilon_A\right)\subset U$ for any finite non-empty set A of $\{1,2,\cdots,n\}$. Let $\epsilon=\min\{\epsilon_A,y_i:\emptyset\neq F\subseteq\{1,2,\cdots,n\},1\leq i\leq n\}$. Pick m such that for all $j\geq m$ we have $x_j<\epsilon$, and set $y_{n+1}=x_m$. This finishes the induction step. Therefore, all the finite sums and all the finite products of the sequence $(y_n)_{n\geq 1}$ are in U, and so are monochromatic for the colouring ν.

To complete the proof we show that if z is either a finite sum or a finite product of $(y_n)_{n\geq 1}$ and $z\in T_k\setminus T_{k-1}$, then $e_k(z)=-1$, and consequently $s_k(z)=-1$.

First assume that z is a finite product of elements of $(y_n)_{n\geq 1}$. This implies that z is a finite product of elements of $\displaystyle\left(\frac{1}{p_n}\right)_{n\geq 1}$. Therefore there exists a finite set A of natural numbers such that $z = \prod_{i \in A} \frac{1}{p_i}$, and thus $z \in T_k \setminus T_{k-1}$, where $k = \max A$. We observe that $zP_k = \prod_{i \in \{1,2,\ldots,k\} \setminus A} p_i < P_k$, so $z = \frac{z'}{P_k}$ for some $1 \leq z' < P_k$, which implies that $e_k(z) = s_k(z) = -1$.

Finally, let $z = \sum_{i \in A} y_i$ for some finite set $A = \{j_1, j_2, \cdots, j_s\}$ of natural numbers of size $s > 1$, where $j_1 < j_2 < \cdots < j_s$. Since $(y_n)_{n \geq 1}$ is a subsequence of $(x_n)_{n \geq 1}$, which is a product subsystem of $\left(\frac{1}{p_{r_n}}\right)_{n \geq 1}$, for each $i \in \{1, 2, \ldots, s\}$ there exists a finite set F_i of natural numbers such that $\max F_i < \min F_{i+1}$ if $i < s$, and $y_{j_i} = \prod_{t \in F_i} \frac{1}{p_{r_t}}$. Denote by m_i the maximum of F_i for all $i \in \{1, 2, \ldots, s\}$, and let $k = r_{m_s}$, so that $z \in T_k \setminus T_{k-1}$.

We first note that $\sum_{i=1}^{s} \frac{1}{p_{r_{m_i}}} < 1$, and thus $\sum_{i=1}^{s} \frac{p_k}{p_{r_{m_i}}} < p_k$. We now see that $zP_k = \left(\sum_{i=1}^{s} y_{j_i}\right) \prod_{m=1}^{k} p_m = \left(\sum_{i=1}^{s} \prod_{t \in F_i} \frac{1}{p_{r_t}}\right) \prod_{m=1}^{k} p_m \leq \left(\sum_{i=1}^{s} \frac{1}{p_{r_{m_i}}}\right) \prod_{m=1}^{k} p_m = \left(\sum_{i=1}^{s} \frac{p_k}{p_{r_{m_i}}}\right) \prod_{m=1}^{k-1} p_m < P_k$, by the above observation. Therefore, as before, $z = \frac{z''}{P_k}$ for some $1 \leq z'' < P_k$, which implies $e_k(z) = s_k(z) = -1$. □

3.2.7 Unbounded sequences in the rationals

In this subsection we give a finite colouring of the rationals such that no unbounded sequence whose denominators contain only finitely many primes can have the set of all its finite sums and products monochromatic.

The general aim is to write numbers as an integer part (which will be considered in binary) and a fractional part (which will be considered in the 'minimal base' as in Subsection 3.2.5), although actually we will also make use of the integer part written in that minimal base of the fractional part. By using the finite sums, we hope to show that the 'centres clear out', meaning that the fractional parts tend to 0 (or 1) and the integer parts have ends that tend to infinity. This will then give us the disjointness of support that we need to apply results conceptually similar to Lemma 3.7. For example, if the fractional parts tend to 0 and the integer parts have ends that tend to infinity then we will consider the relationship between quantities like the left gap of the integer part and the end of the fractional part – the key point being that we will be able to control how the integer parts behave under sum and product, because the fractional parts will be 'too small to interfere'.

Theorem 3.15. *There exists a finite colouring α of the positive rationals such that there exists no unbounded sequence $(x_n)_{n\geq 1}$ that has the set of all its finite sums and products* *monochromatic with respect to α, with the set of primes that divide the denominators of its terms being finite.*

Proof. Let $S_n = \{x \in \mathbb{Q}^+ : x \text{ has a terminating base } P_n \text{ expansion}\}$ for all $n > 0$, and $S_0 = \emptyset$. We first define the colouring α' of $\mathbb{Q}^+ \setminus (\mathbb{N} \cup \{2^k : k \in \mathbb{Z}\} \cup (0, 2])$ as follows: for $x \in S_r \setminus S_{r-1}$ we set

$$\alpha'(x) =$$
$$(a(x) \bmod 2, a(\text{frac}(x)) \bmod 2, \epsilon(\text{frac}(x)) \bmod 2, e_r(\lfloor x \rfloor) \bmod 2, e_1(\lfloor x \rfloor) \bmod 2,$$
$$e_r(\lfloor x \rfloor + 1) \bmod 2, e_1(\lfloor x \rfloor + 1) \bmod 2, a(r(x)) \bmod 3, p(x), q(x), q'(x), s(x), s'(x)),$$

where $1 - 2^{\epsilon(\text{frac}(x))} \leq \text{frac}(x) < 1 - 2^{\epsilon(\text{frac}(x))-1}$, and as before $e_r(x)$ is the position of the rightmost significant digit in base P_r and $e_1(x)$ is the position of the rightmost significant digit in binary, and also $r(x) = \dfrac{x - 2^{a(x)}}{2^{a(x)}}$, $p(x)$ is 0 if $\lfloor x \rfloor$ is a power of 2 and 1 otherwise, $q(x)$ is 0 if $a(x) - b(x) > e_r(\lfloor x \rfloor)$ and 1 otherwise, $q'(x)$ is 0 if $a(x) - b(x) > e_r(\lfloor x \rfloor + 1)$ and 1 otherwise, $s(x)$ is 0 if $a(x) - c(x) > e_r(\lfloor x \rfloor)$ and 1 otherwise, $s'(x)$ is 0 if $a(x) - c(x) > e_r(\lfloor x \rfloor + 1)$ and 1 otherwise. Here $\lfloor x \rfloor$ and $\text{frac}(x)$ are the integer and the fractional parts of x respectively.

We are now ready to define the colouring α. Let $x \in \mathbb{Q}^+$. Then $\alpha(x) = (0, \theta(x))$ if $x \in \mathbb{N}$, $\alpha(x) = 1$ if $x \in \{2^k : k \in \mathbb{Z}, k < 0\}$, $\alpha(x) = 2$ if $x \leq 2$, $x \notin \mathbb{N}$ and $x \notin \{2^k : k \in \mathbb{Z}, k < 0\}$, and $\alpha(x) = (1, \alpha'(x))$ otherwise.

Suppose for a contradiction that a sequence as specified in the statement of the theorem exists. Since it is unbounded, we may assume that all its terms are greater than 2. Since θ prevents any sequence of natural numbers from having monochromatic pairwise sums and products, we may assume, by passing to a subsequence, that none of the x_n are natural numbers – and hence, since the set of the finite sums and products is monochromatic, also no finite sum or product of the x_n is a natural number. Moreover, by looking at sums of two terms, it is easy to see that p prevents the integer parts from being powers of 2, and thus we can assume that no x_n has its integer part a power of 2. By assumption, and after passing to a subsequence, we may assume that there exists $r \in \mathbb{N}$ such that $x_n \in S_r \setminus S_{r-1}$ for all n. Since $S_r \setminus S_{r-1}$ is closed under multiplication, all the finite products are in $S_r \setminus S_{r-1}$ too.

Let $x_n = y_n + z_n$, where $y_n \in \mathbb{N}$ is the integer part of x_n and $0 < z_n < 1$ is its fractional part. By passing to a subsequence we may assume that the sequence $(y_n)_{n\geq 1}$ is strictly increasing and tending to infinity. Suppose that the sequence $(z_n)_{n\geq 1}$ is bounded away from both 0 and 1, which is equivalent to saying that $a(z_n)$ and $\epsilon(z_n)$ are both bounded. Therefore, by passing to a subsequence, we may assume that there exist fixed integers $k < 0$ and $l < 1$ such that $a(z_n) = k$ and $\epsilon(z_n) = l$ for all n. We either have $z_n < \frac{1}{2}$ for infinitely many n or $z_n \geq \frac{1}{2}$ for infinitely many n.

In the first case, if z_n and z_m are less than $\frac{1}{2}$ then $\text{frac}(x_n + x_m) = z_n + z_m$, and thus $a(\text{frac}(x_n + x_m)) = k + 1 \neq a(\text{frac}(x_n)) \mod 2$, a contradiction. In the second case, if z_n and z_m are at least $\frac{1}{2}$ then $\text{frac}(x_n + x_m) = z_n + z_m - 1$, so that $1 - \text{frac}(x_n + x_m) = 1 - z_n + 1 - z_m$ which implies that $\epsilon(\text{frac}(x_n + x_m)) = l + 1 \neq \epsilon(\text{frac}(x_n)) \mod 2$, a contradiction. This tells us that, by passing to a subsequence, we may either assume that z_n converges to 0 or that it converges to 1.

By passing to a subsequence we may assume that either $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all n or $x_n \geq 2^{a(x_n)+\frac{1}{2}}$ for all n. In the first case $a(x_nx_m) = a(x_n) + a(x_m)$, while in the second case $a(x_nx_m) = a(x_n) + a(x_m) + 1$ (for all $n \neq m$). Since, for $x \in \mathbb{R}^+ \setminus (\mathbb{N} \cup C_1)$, $r(x)$ is the unique number strictly between 0 and 1 such that $x = 2^{a(x)}(1 + r(x))$, a simple computation shows that in the first case $r(x_nx_m) = r(x_n) + r(x_m) + r(x_n)r(x_m)$, while in the second case $r(x_nx_m) = \dfrac{r(x_n) + r(x_m) + r(x_n)r(x_m) - 1}{2}$ for all $n \neq m$.

Suppose that $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all n and that $r(x_n)$ is bounded away from 0. Then $a(r(x_n))$ is bounded, so by passing to a subsequence we may assume that there is an integer $l < -1$ such that $a(r(x_n)) = l$ for all n (Recall that we are in the case where $r(x_n) + r(x_m) + r(x_n)r(x_m) < 1$ and thus $a(r(x_n)) < -1$). Since $2^l \leq r(x_n) < 2^{l+1}$ and $2^l \leq r(x_m) < 2^{l+1}$, we have that $2^{l+1} < 2^{l+1} + 2^{2l} \leq r(x_n) + r(x_m) + r(x_n)r(x_m) < 2^{l+1} + 2^{2l+2} < 2^{l+3}$. Thus $a(r(x_nx_m))$ is $l+1$ or $l+2$, neither of which is congruent to l mod 3, a contradiction. Therefore in this first case (namely when $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all n), we must have that $r(x_n)$ converges to 0, which immediately implies that $a(x_n) - b(x_n)$ (the 'left gap') goes to infinity.

Suppose instead that we are in the second case (namely that $x_n \geq 2^{a(x)+\frac{1}{2}}$ for all n), so that $r(x_nx_m) = \dfrac{r(x_n) + r(x_m) + r(x_n)r(x_m) - 1}{2}$ for all $n \neq m$. Suppose that $a(x_n) - c(x_n)$ is bounded. By passing to a subsequence, we may assume that there exists a fixed $l \in \mathbb{N}$ such that $a(x_n) - c(x_n) = l$ for all n. Let $2k - 2 < d \in \mathbb{N}$ be such that $\dfrac{(2^{k+1}-1)^d}{2^{(k+1)d}} < \dfrac{1}{2}$, and look at the first d terms. We have that $x_j < 2^{a(x_j)+1} - 2^{c(x_j)} = 2^{a(x_j)+1} - 2^{a(x_j)-k} = 2^{a(x_j)}\dfrac{2^{k+1}-1}{2^k}$, so that we have $x_1x_2\cdots x_d < 2^{a(x_1)+\ldots+a(x_d)}\dfrac{(2^{k+1}-1)d}{2^{kd}} < 2^{a(x_1)+\ldots+a(x_d)+k-1}$. On the other hand, by assumption, the product is at least $2^{a(x_1)+\cdots+a(x_d)+\frac{d}{2}} > 2^{a(x_1)+\cdots+a(x_d)+k-1}$, a contradiction. Therefore we may assume that $a(x_n) - c(x_n)$ is strictly increasing and goes to infinity, which is equivalent to $r(x_n)$ converging to 1.

To summarise, we either have $r(x_n)$ converging to 0, which is equivalent to $a(x_n) - b(x_n)$ going to infinity, or $r(x_n)$ converging to 1, which is equivalent to $a(x_n) - c(x_n)$ going to infinity. We distinguish these two cases.

Case 1. The sequence $(z_n)_{n\geq1}$ converges to 0. In this case, by passing to a subsequence we may assume that the terms of the sequence $(z_n)_{n\geq1}$ have pairwise left to right disjoint supports in base P_r – note that this implies that all finite sums of $(x_n)_{n\geq1}$ also have minimal base P_r. By passing to a subsequence we may assume that all y_n have the same digit in position $e_r(y_n)+1$ in base P_r, and that $z_n < \frac{1}{P_r}$ for all n. Suppose that there exist P_r terms such that their integer parts end at the same position in base P_r, call it p. It is easy to see that the integer part of their sum is the sum of their integer parts, which ends at position $p+1$, a contradiction. Therefore we may assume that the terms of the sequence $(y_n)_{n\geq 1}$ have left to right disjoint supports in base P_r. By exactly the same argument (looking at a sum of two terms only) we can further deduce that the terms of the sequence $(y_n)_{n\geq 1}$ have left to right disjoint supports in binary as well.

Assume first that $r(x_n)$ converges to 0. We fix x_1 and look at x_1+x_n. For n sufficiently large we have $q(x_1+x_n)=0$, because the left gap of the sum is the left gap of x_n, while the end position of $\lfloor x_1+x_n \rfloor$ in base P_s is fixed, namely the end position of y_1 in base P_r. On the other hand, if the fractional part of x_1 has end position $a<0$ in base P_r and n is large enough, then $\lfloor x_1x_n \rfloor$ has end position $e_r(y_n)+a$ in base P_r, which tends to infinity as n tends to infinity. However, due to the fact that the left gap of x_n goes to infinity, we see that for n large enough the left gap of x_nx_1 equals the left gap of x_1, which will eventually be less than $e_r(y_n)+a$. So $q(x_1x_n)=1$, a contradiction.

Assume now that $r(x_n)$ converges to 1. As before, we fix x_1 and look at x_n+x_1 for n large enough. Since x_n and x_1 have disjoint supports in binary, we have that $a(x_n+x_1)=a(x_n)$, and thus $r(x_n+x_1)=\dfrac{x_n+x_1-2^{a(x_n)}}{2^{a(x_n)}}$ which converges to 1. Therefore, as n tends to infinity, $a(x_n+x_1)-c(x_n+x_1)$ also tends to infinity – thus it will eventually be greater that the end position of $\lfloor x_n+x_1 \rfloor$ in base P_r (which is the end position of y_1 in base P_r), so $s(x_n+x_1)=0$ for all n large enough. On the other hand, it is a straightforward computation to show that $a(x_nx_1)-c(x_nx_1)$ is either $a(x_1)-c(x_1)$ or $a(x_1)-c(x_1)+1$, and thus is bounded. However, we have seen above that $e_r(\lfloor x_nx_1 \rfloor)$ is unbounded. We conclude that for all n sufficiently large we have $a(x_nx_1)-c(x_nx_1)<e_r(\lfloor x_nx_1 \rfloor)$, and thus $s(x_nx_1)=1$ for all n sufficiently large, a contradiction. This concludes Case 1.

Case 2. The sequence $(z_n)_{n\geq 1}$ converges to 1. In this case we have that $x_n=y_n+1-(1-z_n)$ and the sequence $(1-z_n)_{n\geq 1}$ converges to 0. With the same type of argument as the one presented above, we may assume that the terms of the sequence $(y_n+1)_{n\geq 1}$ have pairwise left to right disjoint supports in binary and in base P_r, and the sequence is strictly increasing (it suffices to show that we cannot have infinitely many terms ending at the same place in binary or in base P_r). Since the full argument for base P_r has been given above, here we just include the argument for binary. So suppose that we have $n\neq m$ such that $e_1(y_n+1)=e_1(y_m+1)=p$ and y_n+1 and y_m+1 have the same binary digit in position $p+1$ (which we can achieve by passing to a subsequence). Then $e_1(\lfloor x_n \rfloor+1)=p$, while $e_1(\lfloor x_n+x_m \rfloor+1)=e_1(y_n+y_m+2)=p+1$, a contradiction.

We observe that for any $n>1$, $e_r(\lfloor x_n+x_1 \rfloor+1)=e_r((y_n+1)+(y_1+1))=e_r(y_1+1)$. Let $e_r(x_1)=u<0$ and pick n such that $1-z_n<\frac{1}{x_1}$ and $e_r(y_n+1)=v_n>-u$. This implies that $0<1-(1-z_n)x_1<1$ and that $(y_n+1)x_1\in\mathbb{N}$. Therefore $x_nx_1=((y_n+1)-(1-z_n))x_1=(y_n+1)x_1-(1-z_n)x_1$, and thus $\lfloor x_nx_1 \rfloor+1=\lfloor x_nx_1+1 \rfloor=\lfloor (y_n+1)x_1+1-(1-z_n)x_1 \rfloor=(y_n+1)x_1$. This means that $e_r(\lfloor x_nx_1 \rfloor+1)=v_n+u$ for all n sufficiently large, so that the sequence $(e_r(\lfloor x_nx_1 \rfloor))_{n\geq 1}$ is unbounded.

To complete the proof, we show that if $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all $n \geq 1$ then for sufficiently large n we have $q'(x_n + x_1) = 0$ and $q'(x_n x_1) = 1$, while if $x_n \geq 2^{a(x_n)+\frac{1}{2}}$ for all $n \geq 1$ then for sufficiently large n we have $s'(x_n + x_1) = 0$ and $s'(x_n x_1) = 1$.

Assume first that $x_n < 2^{a(x_n)+\frac{1}{2}}$ for all $n \geq 1$. As we have seen above, this implies that $a(x_n)-b(x_n)$ tends to infinity (and we may also assume that it is strictly increasing and $a(x_1) - b(x_1) > 2$). Consequently $a(x_n + x_1) - b(x_n + x_1)$ also tends to infinity, and so is eventually larger than $e_r(\lfloor x_n + x_1 \rfloor + 1)$, whence $q'(x_n + x_1) = 0$ for n large enough. On the other hand, since $2^{a(x_n)} + 2^{b(x_n)} \leq x_n < 2^{a(x_n)} + 2^{b(x_n)+1}$ and $2^{a(x_1)} + 2^{b(x_1)} \leq x_1 < 2^{a(x_1)} + 2^{b(x_1)+1}$, we have that $2^{a(x_n)+a(x_1)} + 2^{a(x_n)+b(x_1)} < x_n x_1 < 2^{a(x_n)+a(x_1)} + 2^{a(x_n)+b(x_1)+1} + 2^{a(x_1)+b(x_n)+1} + 2^{b(x_n)+b(x_1)+2} < 2^{a(x_n)+a(x_1)} + 2^{a(x_n)+b(x_1)+2}$. This is because $b(x_n) + b(x_1) + 2 < a(x_1) + b(x_n) + 1 < a(x_n) + b(x_1) + 1$. Therefore $b(x_n x_1)$ is either $a(x_n) + b(x_1)$ or $a(x_n) + b(x_1) + 1$, and thus $a(x_n x_1) - b(x_n x_1) \leq a(x_1) - b(x_1)$. Since $e_r(\lfloor x_1 x_n \rfloor + 1)$ will eventually be greater than $a(x_1) - b(x_1)$, we have that $q'(x_n x_1) = 1$ for n large enough, a contradiction.

Finally, assume that $x_n \geq 2^{a(x_n)+\frac{1}{2}}$ for all $n \geq 1$. Thus $a(x_n) - c(x_n)$ goes to infinity (and as above we may assume it to be strictly increasing and such that $a(x_1)-c(x_1) > 2$), and consequently so does $a(x_n+x_1)-c(x_n+x_1)$. This means that $a(x_n+x_1)-c(x_n+x_1) > e_r(\lfloor x_n + x_1 \rfloor + 1)$ for n large enough, and so $s'(x_n + x_1) = 0$ for n large enough. On the other hand, $2^{a(x_n)+1} - 2^{c(x_n)+1} \leq x_n < 2^{a(x_n)+1} - 2^{c(x_n)}$ and $2^{a(x_1)+1} - 2^{c(x_1)+1} \leq x_1 < 2^{a(x_1)+1} - 2^{c(x_1)}$. This implies that $2^{a(x_n)+a(x_1)+2} - 2^{a(x_n)+c(x_1)+3} \leq 2^{a(x_n)+a(x_1)+2} - 2^{a(x_n)+c(x_1)+2} - 2^{a(x_1)+c(x_n)+2} + 2^{c(x_n)+c(x_1)+2} < x_n x_1 < 2^{a(x_n)+a(x_1)+2} - 2^{a(x_n)+c(x_1)+1}$. Here the first inequality holds because $a(x_n)+c(x_1)+2 > a(x_1)+c(x_n)+2$, which implies that $2^{a(x_n)+c(x_1)+2}+2^{a(x_1)+c(x_n)+2} < 2^{a(x_n)+c(x_1)+3}$. Therefore $c(x_n x_1)$ is either $a(x_n)+c(x_1)+1$ or $a(x_n)+c(x_1)+2$, and so $a(x_n x_1)-c(x_n x_1) \leq a(x_1)-c(x_1)$. Since $e_r(\lfloor x_1 x_n \rfloor + 1)$ will eventually be greater than $a(x_1)-c(x_1)$, we have that $s'(x_n x_1) = 1$ for n large enough, a contradiction. This concludes Case 2. □

Note that Theorem 3.15, together with Theorem 3.13, completes the proof of our main result.

Theorem 3.16. *There exists a finite colouring of the rational numbers with the property that there exists no sequence such that the set of its finite sums and products is monochromatic and the set of primes that divide the denominators of its terms is finite.*

3.2.8 Concluding remarks

The first remaining problem is of course to understand what happens with finite sums and products in the rationals. The above colourings of $\mathbb{Q}_{(k)}$ do rely heavily on the representation of numbers in a suitable base, and so do not pass to sequences from the whole of $\mathbb{Q}$. It would be very good to find 'parameters' a and b that would allow Lemma 3.7 to be applied, or perhaps some variant like Lemma 3.8. We have tried to find such parameters in the rationals in general, but have been unsuccessful. It would be extremely interesting to decide whether or not such parameters do exist.

3.2.9 Appendix

Here we provide the cases in the proof of Theorem 3.12 when the colour class is C_3 or C_5.

Proposition 3.17. *There does not exist an injective sequence $(x_n)_{n\geq 1}$ in $\mathbb{R}^+$ such that the set of all its pairwise sums and products is contained in $C_3 = \{2^k + 2^l : k, l \in \mathbb{Z}$ and $l < k\}$.*

Proof. Assume for a contradiction that such a sequence $(x_n)_{n\geq 1}$ exists. It is easy to see that if $x < y < z$ are three positive real such that $\{x+y, x+z, y+z\} \subseteq C_3$ then $\{x, y, z\} \subseteq \mathbb{Q}_{(2)}$, and so $x_n \in \mathbb{Q}_{(2)}$ for all $n \geq 1$.

We know that the set $\{x_n : n \in \mathbb{N}\} \cap \{2^k : k \in \mathbb{Z}\}$ is finite, otherwise we get a contradiction as the product of two powers of 2 does not lie in C_3. We may therefore assume that no x_n is a power of 2.

Suppose first that $x_n \in (0, 1)$ for all $n \geq 1$. Suppose that $\{s_1(x_n) : n \in \mathbb{N}\}$ is infinite. We may pick n such that $s_1(x_n) < e_1(x_1)$, but then the binary expansion of $x_1 + x_n$ has at least four nonzero digits, and thus $x_1 + x_n \notin C_3$, a contradiction. We may therefore assume (after passing to a subsequence) that there exists $k \in \mathbb{Z}$ (with $k < 0$) such that $s_1(x_n) = k$ for every $n \geq 1$. Then each $x_n = 2^k + y_n$ where $s_1(y_n) < k$. Since there are only finitely many numbers with given values of $s_1(x)$ and $e_1(x)$, by passing to a subsequence we may also assume that $e_1(y_n) > e_1(y_{n+1})$ for all $n \geq 1$. We now observe that if $n < m$ then $s_1(x_n + x_m) = k + 1$ and $e_1(x_n + x_m) = e_1(x_m)$, so $x_n + x_m$ has a nonzero digit at positions $k + 1$ and $e(x_m)$, and thus, since it is in C_3, we have $x_n + x_m = 2^{k+1} + 2^{e(x_m)}$. But then $x_1 + x_3 = x_2 + x_3$, a contradiction.

We may therefore assume that $x_n > 1$ for all $n \geq 1$. By Ramsey's theorem for pairs, we may assume either that for all $n \neq m$ we have $x_n + x_m \in \{2^k + 2^l : k, l \in \mathbb{Z}$ and $0 \leq l < k\}$ or that for all $n \neq m$ we have $x_n + x_m \in \{2^k + 2^l : k, l \in \mathbb{Z}$ and $l < 0 < k\}$.

Case 1. For all $n \neq m$ we have $x_n + x_m \in \{2^k + 2^l : k, l \in \mathbb{Z}$ and $0 \leq l < k\}$.

Let $y_n = \lfloor x_n \rfloor$ and $\alpha_n = x_n - y_n$ for all $n \geq 1$. Given $n \neq m$, we have $x_n + x_m = y_n + y_m + \alpha_n + \alpha_m$, and so $\alpha_n + \alpha_m \in \{0, 1\}$. If n, m and r are pairwise distinct and $\alpha_n, \alpha_m, \alpha_r \notin \{0, \frac{1}{2}\}$, then some two are in $(0, \frac{1}{2})$ or some two are in $(\frac{1}{2}, 1)$, a contradiction. Hence, for all but at most two values of n, we have $\alpha_n \in \{0, \frac{1}{2}\}$. If $n \neq m$ and $\alpha_n = \alpha_m = \frac{1}{2}$, then $x_n \cdot x_m \notin \mathbb{N}$, again a contradiction. We may therefore assume that $\alpha_n = 0$ for all $n \geq 1$.

Since no x_n is a power of 2, $\{e_1(x_n) : n \in \mathbb{N}\}$ is finite. The reasoning is similar to that presented above: if $e_1(x_n) > s_1(x_1)$ then the binary expansion of $x_1 + x_n$ has at least four nonzero digits. We may therefore assume that there exists k such that $e_1(x_n) = k$ for all $n \geq 1$. By passing to a subsequence, we may further assume that either each x_n end in 01 or each x_n ends in 11, so that $e_1(x_n + x_m) = k + 1$. Moreover, we may also assume that $s_1(x_n) < s_1(x_{n+1})$ for all $n \geq 1$.

We now see that if $n < m$ then $s_1(x_n + x_m) = s_1(x_m)$ or $s_1(x_n + x_m) = s_1(x_m) + 1$. Pick $i \neq j$ in $\{1, 2, 3\}$ and $t \in \{0, 1\}$ such that $s_1(x_i + x_4) = s_1(x_4) + t$ and $s_1(x_j + x_4) =$

$s_1(x_4)+t$. Since $k+1 < s_1(x_4)+t$ are two positions of nonzero digits, we must have $x_i + x_4 = x_j + x_4 = 2^{s_1(x_4)+t} + 2^{k+1}$, a contradiction

Case 2. For all $n \neq m$ we have $x_n + x_m \in \{2^k + 2^l : k, l \in \mathbb{Z}$ and $l < 0 < k\}$.

In this case, for all $n \neq m$, $x_n + x_m$ has one nonzero digit to the right of the decimal point and one nonzero digit to the left of the decimal point.

Suppose first that $\{e_1(x_n) : n \in \mathbb{N}\}$ is unbounded. By passing to a subsequence, we may assume that $0 > e_1(x_1) > e_1(x_2) > e_1(x_3)$. This implies that $x_1 + x_3$ and $x_2 + x_3$ each have a nonzero digit in position $e_1(x_3)$ and $x_1 + x_2$ has a nonzero digit in position $e_1(x_2)$. Thus there exist $y, z, w \in \mathbb{N}$ such that $x_1 + x_3 = y + 2^{e_1(x_3)}$, $x_2 + x_3 = z + 2^{e_1(x_3)}$, and $x_1 + x_2 = w + 2^{e_1(x_2)}$. Clearly we have that $y \neq z$. If $z > y$, then $x_2 - x_1 = z - y$ so $2x_2 = z - y + w + 2^{e_1(x_2)}$, whence $e_1(x_2) = e_1(2x_2) = e(x_2) + 1$, a contradiction. If $y > z$, then $x_1 - x_2 = y - z$, so $2x_1 = y - z + w + 2^{e_1(x_2)}$, giving $e_1(x_2) = e_1(2x_1) = e_1(x_1) + 1 > e_1(x_2)$, again a contradiction.

Hence $\{e_1(x_n) : n \in \mathbb{N}\}$ is bounded. Thus $\{s_1(x_n) : n \in \mathbb{N}\}$ has to be unbounded. We may therefore assume that there exists $k < -1$ such that $e_1(x_n) = k$ for all $n \geq 1$. (If $e_1(x_n) = e_1(x_m) = -1$ then $x_n + x_m \in \mathbb{N}$.) By passing to a subsequence, we may also assume that all terms of the sequence have the same digit in position $k+1$, and for all $n \neq m$ we have $e_1(x_n + x_m) = k + 1$.

We may further assume that $s_1(x_1) < s_1(x_2) < s_1(x_3) < s_1(x_4)$. For $i \in \{1, 2, 3\}$, $x_i + x_4$ has a nonzero digit in position $s_1(x_4)$ or in position $s_1(x_4) + 1$. Pick $i \neq j$ in $\{1, 2, 3\}$ and $t \in \{0, 1\}$ such that $x_i + x_4$ and $x_j + x_4$ each have a nonzero digit in position $s_1(x_4) + t$. Then $x_i + x_4 = x_j + x_4 = 2^{s_1(x_4)+t} + 2^{k+1}$, a contradiction. $\square$

Proposition 3.18. *There does not exist an injective sequence $(x_n)_{n\geq 1}$ in $\mathbb{R}^+$ such that the set of all its pairwise sums and products is contained in $C_5 = \{2^{k+1}(1 - 2^{l-k})^{\frac{1}{2}} : k, l \in \mathbb{Z}$ and $l < k\}$.*

Proof. Assume for a contradiction that such a sequence $(x_n)_{n\geq 1}$ exists. Let α, β, γ be three numbers in C_5 such that $x_1 + x_2 = \alpha$, $x_1 + x_3 = \beta$ and $x_2 + x_3 = \gamma$. Let also $x_1 x_2 = \mu$, $x_1 x_3 = \nu$ and $x_2 x_3 = \eta$, where μ, ν and η are in C_5. We therefore have $x_1^2 = \frac{\mu \cdot \nu}{\eta}$, whence x_1^4 is rational.

Case 1. Suppose that $\alpha \cdot \beta$, $\alpha \cdot \gamma$ and $\beta \cdot \gamma$ are all irrational. Since α^2, β^2 and γ^2 are rational, α/β, α/γ and β/γ are all irrational as well. It is easy to show that if K and R are two fields such that $\mathbb{Q} \subset K \subset R$ and $\delta \in R \setminus K$ is such that $\delta^2 \in \mathbb{Q}$, then $K(\delta) = \{a + b \cdot \delta : a, b \in K\}$. Using this fact, it is straightforward to show that $\beta \notin \mathbb{Q}(\alpha)$, $\alpha \notin \mathbb{Q}(\beta)$ and $\gamma \notin \mathbb{Q}(\alpha, \beta)$.

Now, we know that x_1^4 is rational. On the other hand, $x_1 = \frac{\alpha+\beta-\gamma}{2}$, and so $16 \cdot x_1^4 = (\alpha + \beta - \gamma)^4 = r_0 + r_1 \cdot \alpha \cdot \beta - r_2 \cdot \alpha \cdot \gamma - r_3 \cdot \beta \cdot \gamma$, where r_0, r_1, r_2, and r_3 are positive rationals. It then follows that $\gamma \cdot (r_2 \cdot \alpha + r_3 \cdot \beta) = -16 \cdot x_1^4 + r_0 + r_1 \cdot \alpha \cdot \beta$, which implies that γ is in $\mathbb{Q}(\alpha, \beta)$, a contradiction. (For the conscientious reader, the coefficients are $r_0 = \alpha^4 + \beta^4 + \gamma^4 + 6 \cdot \alpha^2 \cdot \beta^2 + 6 \cdot \alpha^2 \cdot \gamma^2 + 6 \cdot \beta^2 \cdot \gamma^2$, $r_1 = 4 \cdot \alpha^2 + 4 \cdot \beta^2 + 12 \cdot \gamma^2$,

$r_2 = 4 \cdot \alpha^2 + 4 \cdot \gamma^2 + 12 \cdot \beta^2$ and $r_3 = 4 \cdot \beta^2 + 4 \cdot \gamma^2 + 12 \cdot \alpha^2$.)

Case 2. Suppose now that $\alpha \cdot \beta$ is a rational number, say q. It is clear that $q > 0$. We then have $(x_1 + x_2)(x_1 + x_3) = q = x_1^2 + x_1x_3 + x_1x_2 + x_2x_3 = x_1^2 + \mu + \nu + \eta$. We now observe that, by the definition of C_5, all of its elements are square roots of positive rational numbers. Hence there exist three positive rational numbers q_1, q_2 and q_3, such that $\mu = \sqrt{q_1}$, $\nu = \sqrt{q_2}$ and $\eta = \sqrt{q_3}$. Moreover, since $x_1^2 = \frac{\mu \cdot \nu}{\eta}$, it follows that x_1^2 is also a square root of a positive rational. More precisely $x_1^2 = \sqrt{q_4}$ where $q_4 = \frac{q_1 \cdot q_2}{q_3}$.

We therefore have $q = \sqrt{q_1} + \sqrt{q_2} + \sqrt{q_3} + \sqrt{q_4}$. Let $M = \mathbb{Q}(\sqrt{q_1}, \sqrt{q_2}, \sqrt{q_3}, \sqrt{q_4})$, and let d be its degree over $\mathbb{Q}$. On the one hand, the trace of q is $d \cdot q$, and on the other had it is the sum of $d \cdot \sqrt{q_i}$ for those q_i that are perfect squares. This is because, for any positive rational t, the trace of $\sqrt{t}$ is 0 if t is not a perfect square, and $d\sqrt{t}$ if t is a perfect square. The only way to have equality in the above is if all the q_i are perfect squares, but then $x_1x_2 \in C_5$ is rational, a contradiction. $\square$

4 Constructible Graphs and Pursuit

4.1 Introduction

The game of cops and robbers is played on a fixed graph G, which for the moment we will assume is finite. The cop picks a vertex to start at, and the robber then does the same. Then they move alternately, with the cop moving first: at each turn the player moves to an adjacent vertex or does not move. The game is won by the cop if he lands on the robber. We say that G is *cop-win* if the cop has a winning strategy. Needless to say, if the graph is not connected then this game is a rather trivial robber win, so we assume from now on that all graphs are connected.

The cop-win graphs were characterised by Nowakowski and Winkler [44]. It is an easy exercise to see that if the graph contains a dominated vertex, say x, then G is cop-win if and only if $G-x$ is cop-win. (Here as usual we say that a vertex y *dominates* a vertex x if the set of x and all neighbours of x is contained in the set of y and all neighbours of y.) It is also easy to see that if no vertex is dominated then the robber has a winning strategy, so that G is not cop-win – on each turn, the robber moves to a vertex not adjacent to the cop. Putting these together, we see that a finite graph G is cop-win if and only if it is constructible, meaning that it can be built up from the one-point graph by repeatedly adding dominated vertices. More precisely, we say that G is *constructible* if there is an ordering of its vertices, say $x_1, \ldots, x_n$, such that, for every $k > 1$, in the graph $G[x_1, \ldots, x_k]$ the vertex x_k is dominated by x_i for some $i < k$. We often refer to the given ordering of the vertices as the *construction ordering*, and the map sending x_k to its dominating x_i as the *domination map* for this ordering. Note that the construction ordering, and the domination map for a given ordering, are typically not unique.

We mention briefly that there is also the 'reverse' notion of a graph being *dismantleable*, meaning that we may start with the graph and repeatedly remove dominated vertices and arrive at the one-point graph. This is of course the same as being constructible (for finite graphs – it turns out that in the infinite setting this is not a useful notion, which is why work on cops and robbers in infinite graphs tends to focus on concepts to do with constructibility). See the book of Bonato and Nowakowski [10] for general background and a wealth of other results in the finite case, where there are many questions about the generalisation where there is more than one cop.

Let us now turn to infinite graphs. The game of cops and robbers has the exact same rules as before. We remark in passing that if the cop does not have a winning strategy then the robber has one, for example because the game is an open game (see eg. [37]).

What about constructibility? The right generalisation of the finite situation is to allow the vertices to be added recursively, in other words along a well-ordering. So we say that G is *constructible* if there is an ordinal β such that its vertices may be listed as the x_α, each $\alpha < \beta$, so that for every $\alpha > 0$ the vertex x_α is dominated in the induced graph $G[\{x_\gamma : \gamma \leq \alpha\}]$. We say that this well-ordering of the vertices is a *construction order*, with *domination map* as before. The *rank* or *construction time* of G is then the least β for which there is a construction order of order-type β. If the rank is ω then we say that that G is *naturally* constructible.

It is easy to find examples of constructible graphs that are not cop-win. For example, a one-way infinite path clearly has this property. However, there is a related notion of 'weak cop win', introduced by Lehner [40] (after earlier work by Chastand, Laviolette and Polat [13]). A graph G is a called a *weak cop win* if there is a strategy for the cop that guarantees that either the cop catches the robber or the robber has to eventually leave (and never return to) every finite set – in other words, for every vertex the robber only visits that vertex finitely often. In the usual language of infinite graphs, one could say that the cop either catches the robber or traps him in one end of the graph (although interestingly, as we will see later, this intuition is not really correct). For example, the one-way infinite path is a weak cop win.

Lehner [40] gave an elegant argument to show that every constructible graph is a weak cop win. He asked if the converse also holds. This was answered by Evron, Solomon and Stahl [18], who gave examples to show that, interestingly, this need not be the case. But none of those examples are cop-win, only weak cop-win. In this chapter we give an example of a graph that is actually cop-win and yet is not constructible. We also give a variant of this graph which is a weak cop win, with two ends, but where the robber never has to commit to being in one of these ends. This shows that in some sense the notion of a weak cop win is more subtle than it might appear.

One of the ingredients of our construction is a finite graph that acts as a kind of 'one-way valve'. This graph, that we call K, has the property that the cop can chase the robber out of it, but 'only in one direction'. It is by putting together copies of K in a certain way that we obtain our desired graph.

This method has an unexpected 'spin-off'. In all known examples of finite constructible graphs, the construction order and domination map could be chosen in such a way that the domination map was a homomorphism (meaning that if x and y are adjacent then their images are adjacent or equal). Chastand, Laviolette and Polat [13] asked if this is always the case. By putting together two copies of K in a certain way, we give a simple counterexample.

Before the paper of Evron, Solomon and Stahl [18], there were no known examples of graphs that were constructible but not naturally constructible. We stress to the reader how remarkable this lack of examples was: the problem is that when a graph is constructible there seem to always be 'many' ways to construct it, starting from almost anywhere in the graph, and this seems to lead to a construction in time ω. This relates to the general reason why cops and robbers on infinite graphs is not so well understood: it seems hard to produce graphs that are cop-win but for an 'interesting' (non-trivial) reason, and similarly for weak cop wins. Indeed, Lehner [40] proved that if G is locally finite and constructible then it must be naturally constructible. Evron, Solomon and Stahl gave examples of graphs whose rank is greater than ω, and indeed they showed that the set of ranks of constructible graphs are unbounded (in the countable ordinals).

They were unable to show that every countable ordinal arises as a rank, and they asked whether or not this holds. We show that this is indeed the case: our starting-point is again based on building up a graph from copies of K.

Another part of this chapter is concerned with a weakening of the notion of constructibility to 'local constructibility'. Returning to weak cop wins, one would naturally imagine that the following generalisation of Lehner's result holds: any graph that is *locally constructible* (meaning that every finite set is contained in a finite constructible set) should be a weak cop win. This should especially be true in the locally finite case. The intuition is that the cop can force the robber out of any finite set using the finite constructible superset of that finite set – perhaps with some compactness argument to make these strategies 'consistent' over different finite sets. We mention in passing that the notion of 'locally constructible' is somehow more tangible that that of 'constructible'. For example, it is clear that we can test whether or not a countable graph is locally constructible in time ω, whereas we see no way to test for constructibility even in any (countable) ordinal time.

We are able to prove this generalisation under a small strengthening of local constructibility: any graph that is 'consistently' locally constructible (which we define below) is indeed a weak cop win. Remarkably, though, some such condition is indeed necessary: our final example is a locally constructible graph that is not a weak cop win. In fact, this graph can even be taken to be *locally finite*, by which we mean that the degree of every vertex is finite. These are by far the most delicate and involved constructions in this chapter.

The plan of this chapter is as follows. In Section 4.2 we introduce the graph K, and as a 'warm-up' we use this to build a finite graph that is constructible but for which no construction order has a domination map that is a homomorphism. Although this result is not one of the main ones of the chapter, we prove it here so as to get the reader used to the properties of K. In Section 4.3 we exhibit a graph that is cop-win but not constructible. Then in Section 4.4 we turn to locally constructible graphs, showing that a consistently locally constructible graph (whether or not locally finite) must be weak cop-win. Section 4.5 contains our examples of locally constructible graphs that are not weak cop-win. We return to general constructibility in Section 4.6, where we find graphs whose ranks are any given countable ordinal. We conclude in Section 4.7 with several open problems.

For general background on cops and robbers, see [10]. For results particularly dealing with infinite graphs, see (apart from the papers mentioned above) Bonato, Golovach, Hahn and Kratochvil [9] for results about capture times, Polat [45][46][47] for material about dismantleability and related aspects, and Hahn, Laviolette, Sauer and Woodrow [25] for other structural properties. For some very attractive results on the computability aspects of constructibility and pursuit see Stahl [52].

Our notation is standard. Our graphs are undirected and loopless. For a subset U of the vertices of a graph G, we write $G[U]$ for the graph induced by U. For two vertices x and y we sometimes write $x \sim y$ if x and y are either adjacent or equal. We often talk informally about vertices being 'added' in a construction order, or 'removed' for dismantling. The chosen vertex that dominates a vertex x in a construction order (in other words, the image of x under the domination map) is often referred to as the 'parent' of x. Finally, for a constructible graph with given construction order and given domination map δ, the *trail* of a vertex x is the (necessarily finite) sequence $x, \delta(x), \delta(\delta(x)), \ldots$ that starts at x and terminates at the *root* (the initial vertex) of the construction order.

4.2 The graph K and and a finite application

In this section we introduce a finite constructible graph that is going to be pivotal for our later constructions. We call this graph K, pictured below. Note that x is the unique dominated vertex and y is its unique dominating vertex. In particular, in any construction ordering the vertex x must come last. To see that K is constructible, or

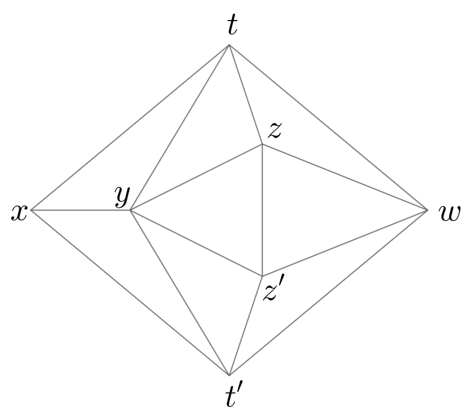


Figure 1: The graph K.

equivalently dismantleable, we observe that the vertex x is dominated by y. Once x is removed t and t' are dominated by z and z' respectively. Once they are removed, z is joined to everything so it dominates all remaining vertices. Thus the graph is dismantleable.

We note that from any vertex in the graph the robber can guarantee to either get to x without being caught or to survive forever. For example, if the robber is at w, he waits until the cop comes to one of t, z, t', z'. If the cop is at t or z the robber goes to t', and if the cop is at t' or z' the robber goes to t. After that he either stays at t, goes back to w or goes to x. (Alternatively, as the robber can obviously avoid being caught whenever he is at a non-dominated vertex, it follows that he can only be caught at x.)

The following lemma is one of the main results about K which we use in our constructions.

Lemma 4.1. *Let G be a constructible graph that has K as an induced subgraph. Moreover, let all the edges between K and $G \setminus K$ have their K-end at one of x and y. Then, in any construction order, x must be the last vertex of K added, and its parent must be y.*

Proof. Suppose that $v \neq x$ is the last vertex of K added. Then v, at this point in the construction, must be dominated by some vertex already added, either in K or the part of G so far constructed. However, since $v \neq x$, v is not dominated by any vertex in K, and v has neighbours in $K \setminus \{x, y\}$, so it is not dominated by any vertex outside K.

Therefore the last vertex of K added must be x. Since x and t are adjacent, the parent of x cannot be outside K, and so its parent must be y. □

To see how these properties of K may be used, we give a simple example of a (finite) constructible graph in which the domination order cannot be chosen to be a homomorphism.

Theorem 4.2. *There exists a finite constructible graph for which no domination map is a homomorphism.*

Proof. We construct the graph G by taking two disjoint copies of K, K_1 and K_2, and identifying the x of the first with the y of the second. The graph is pictured below.

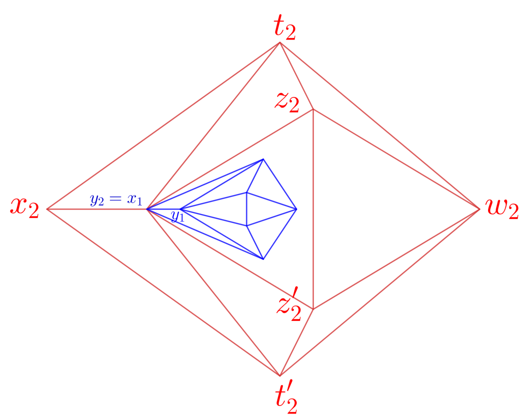


Figure 2: The graph G from Theorem 4.2 showing the two copies of K, K_1 in blue and K_2 in red.

First of all, the above graph is constructible. To see this, we prove that it is dismantleable. We can first remove x_2 as it is dominated by $x_1 = y_2$, then t_2 and t_2' as they are dominated by z_2 and z_2' respectively. Now we can remove w_2 (dominated by z_2) and then z_2 and z_2'. Now we are left with K_1 which we know is dismantleable. This finishes the proof that G is constructible.

We now show that regardless of construction, the domination map is not a homomorphism. In other words, for any construction order of G, there exist two adjacent vertices in G such that their parents cannot be chosen to be adjacent or equal.

Note first that $x_1 = y_2$ must have parent y_1: by Lemma 4.1, x_1 has to be the last vertex added in K_1, which implies that its parent must be y_1.

If the domination map were a homomorphism, then all the neighbours of $x_1 = y_2$ in K_2 would have to have parents that are adjacent to (or equal to) y_1, and the only possible vertex for this is y_2. However, if the vertices t_2, t_2', z_2 and z_2' all have parent y_2, then we cannot construct w_2: it has to be constructed before the last of these four neighbours is constructed, but all these four neighbors are adjacent to w_2, while y_2 is not. This shows that, no matter what the construction order is, the domination map cannot be a homomorphism. □

4.3 A non-constructible cop-win graph

In this section we show that there exists a non-constructible graph on which the cop can always win, meaning as before that he can always catch the robber in finite time.

We begin with an infinite sequence of copies of K, $K_1, K_2, \ldots$, where we identify y_i with x_{i+1}. Finally we add a new vertex which we call 0 and join it to all x_i and y_i. We call this graph, which is pictured below, G. Note that the line of copies of K extends 'to the right and not to the left': this will be crucial.

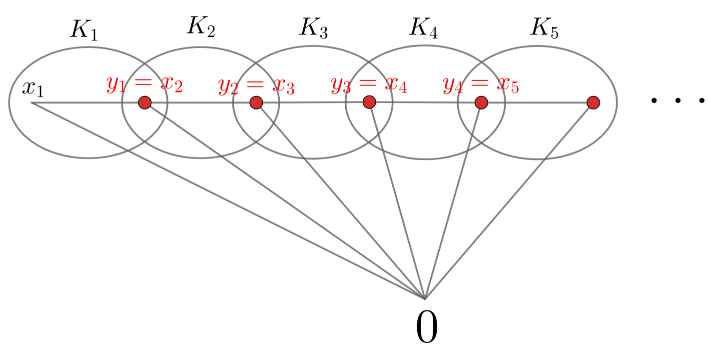


Figure 3: The graph G for Theorem 4.3.

Theorem 4.3. *The graph G is cop-win, but is not constructible.*

Proof. To show that the graph is not constructible, suppose for a contradiction that we have a construction order for it.

Now, by Lemma 4.1 we have that the parent of x_1 must be y_1. But $y_1 = x_2$, and again by Lemma 4.1 the parent of x_2 must be y_2. So in fact the parent of x_i is x_{i+1} for all i, and this contradicts the fact that the construction order is a well-order.

We are left to show that G is a cop-win graph. Whatever the cop's initial position, he can move in at most 2 steps to 0 (or indeed he may just start at 0). After the cop reaches 0, the robber makes his move, and now the robber must be inside some K_i. If the robber is at x_i or y_i then he is caught immediately as these vertices are adjacent to 0. So assume the robber is at some other vertex in K_i. Note that the only vertices in K_i with any neighbour outside K_i are x_i and y_i, so the robber cannot leave K_i until he reaches one of those two vertices.

Now the cop moves to y_i, which will be the start of a chain of forced moves for the robber. Since y_i is adjacent to all vertices in K_i except w_i, the robber has to move to w_i. Next the cop moves to z_i, forcing the robber to t'_i. Then the cop moves to z'_i forcing the robber to x_i. Finally the cop moves to y_i, forcing the robber to leave K_i – and not go to 0 since y_i and 0 are adjacent. The robber must thus move into K_{i-1}, and the cop follows him by moving to $x_i = y_{i-1}$, so that the process can repeat in K_{i-1}.

Continuing in this way, the cop forces the robber out of each copy of K in turn until the robber reaches x_1, where he cannot avoid getting caught. □

There are some variants of the above graph that have some interesting properties. For example, if we remove the vertex 0 then we have a rather simple example of a non-constructible graph that is not cop-win, but is weak cop-win. Indeed, if the cop is in a block to the right of the robber, then the cop can force the robber out of each block in turn as we have see above, and the robber gets caught. However, if the cop is on the robber's left, then the cop runs to the right and the only way the robber can avoid being to the left of the cop at some point is by also running to the right.

In terms of ends of graphs (see [16] for general background), it is natural to assume that in a weak cop win graph the cop can 'force a robber into one end', in the sense that the set of possible ends to which the robber's eventual path can belong, after say time n, should shrink down to one end as n tends to infinity. But, surprisingly, this is not the case. Indeed, consider the variant of the above graph G in which we have a two-way infinite line of copies of K. This graph has two ends. But when the cop chases the robber off to the right, then at every time the robber is always free to 'change direction' by going past the cop (without being caught) and then running off to the left. So the set of possible ends remains both ends of the graph, for all time.

4.4 Locally constructible graphs

We have seen that there are non-constructible graphs that are weak cop wins and even an example that is a cop win. In this section we introduce a weaker notion that captures many of the key properties of constructibility.

Recall that we call a graph G locally constructible if, for any finite set of vertices V, there is a finite set of vertices U containing V such that $G[U]$ is constructible. (We remark that actually one could omit the condition that U is finite, since if U is infinite then the union of the trails in U of all vertices in V is a finite constructible graph.)

One motivation for this definition is that it easily implies that the cop can force the robber to leave any finite set of vertices (although the robber may return later). Indeed, given a finite set of vertices V, take U as in the definition of locally constructible. If the robber stays on V then he necessarily stays on U, and since $G[U]$ is constructible, the standard finite result shows that the cop catches him.

However, this does not show that the game is a weak cop win as nothing in the above argument prevents the robber from returning to the finite set at some later stage. One may naturally feel that some form of compactness argument would yield, at least for locally finite graphs, some way of combining the local strategies from these local constructions into a global strategy. Rather surprisingly, as we shall see in the next section, this is not the case.

First, though, we prove that the cop does have a weak winning strategy in the locally constructible case with an extra condition, which is that the notion of parent is consistent. More precisely, we say a graph G is *consistently* locally constructible if there is a nested sequence of finite induced subgraphs G_i with $\bigcup_i G_i = G$, vertices $v_i \in G_i$, and maps $\delta_i \colon G_i \setminus \{v_i\} \to G_i$ such that:

1. Each G_i is constructible with domination map given by δ_i and root v_i;

2. The maps are consistent: if $v \in (G_i \setminus \{v_i\}) \cap (G_j \setminus \{v_j\})$ then $\delta_i(v) = \delta_j(v)$.

Note that we do not require that the construction orders are consistent, just that the notion of parent is.

We remark that in our example of a graph that is a weak cop win but not constructible above, the graph is consistently locally constructible in a very natural way: just take any finite block of the Ks and construct it starting from the right.

We will need the following finitary result of Isler, Kannan and Khanna [30]. We provide a short proof for the reader's convenience. We also remark that this result actually applies to any constructible graph G equipped with a fixed construction order.

Lemma 4.4. *Let G be a finite constructible graph. Consider the following cop strategy: he starts at the root, and then on each turn he moves to the maximal vertex on the trail of the robber's current position that he is adjacent to. Then*

1. *This strategy is well defined: there is always a neighbour of the cop's current position that is on the trail of the robber.*

2. *If the robber is at some vertex v and returns there at some later time, then the cop is strictly closer to the robber on the second occasion.*

In particular, this strategy is winning for the cop.

Proof. We start by fixing a construction ordering $<$ and associated domination map δ.

Suppose that u and v are any two vertices which are joined in G. We claim that any vertex u' on the trail of u is joined to the maximal vertex v' on the trail of v with $v' \leq u'$, and vice versa. We prove this by reverse induction on the set of vertices in the union of the trails of u and v.

Suppose that it holds for some vertex u' in the trail of u: thus u' is joined to v' where v' is the maximal vertex in the trail of v with $v' \leq u'$. It is immediate from the definition of domination and the domination map that v' is adjacent to (or equal to) $u'' = \delta(u')$. Now, v' is the greatest vertex in the trail of v which is at most u', and u'' is the greatest vertex in the trail of u that is less than u'. Therefore one of v' and u'' is the next-largest vertex in the union of the trails, and the other is the greatest vertex less than that in the other trail. In either case we have the inductive step, which establishes our claim.

Now suppose the robber is at x, the cop is at x' on the trail of x, and the robber moves to y. We see that x' is joined to the greatest vertex v on the trail of y with $v \leq x'$. In particular, x' is joined to a vertex on the trail of y, and therefore the strategy is well defined.

For the second part, suppose that $x' = \delta^k(x)$, and the cop moves to $y' = \delta^\ell(y)$. While the cop's position may decrease (i.e., we may have $y' < x'$) we claim that the position one step closer to the robber on the trail does increase, i.e., $\delta^{\ell-1}(y) > \delta^{k-1}(x)$. Indeed, let $y'' = \delta^{\ell-1}(y)$. Applying the above to x' we see that y' is at least the greatest vertex on the trail of y which is at most x', and thus $y'' > x'$. Now applying the above to y'' we see that y'' is joined to the greatest vertex v on the trail of x with $v \leq y''$. Since the cop did not move to y'' we know that x' is not joined to y''; in particular, $v \neq x'$. Since $y'' > x'$, we see that $v \geq x'$. Combining these, we have $v > x'$, so $v \geq x''$. Thus $y'' \geq v \geq x''$, as required.

Finally, note that these two conditions, together with the trivial observation that the root is a common ancestor of the whole graph, imply that the graph is a weak cop win: each time the robber returns to a vertex the cop is strictly closer, and each vertex only has finitely many ancestors. □

We now prove the main result of this section.

Theorem 4.5. *Let G be a consistently locally constructible graph. Then G is a weak cop win.*

Proof. Define the 'domination' map $\delta : G \to G$ by $\delta(v) = \delta_i(v)$ for any of the local domination maps that occur in the definition of consistently locally constructible that are defined at v. In particular, this means we can talk about the 'trail' of any vertex (although now there is no reason why the trail should be finite).

The cop strategy is as follows. Suppose that the robber is at x and the cop at z.

- Case 1. There exists a neighbour of the cop's current position that is on the trail of the robber. In this case, the cop moves to the most recent ancestor of the robber that he can reach: that is, he moves to the vertex $z' = \delta^k(x)$ with minimal k with $z \sim z'$.

- Case 2. Otherwise, the cop moves to the parent of his current position: that is, he moves to $\delta(z)$.

Suppose that the cop is ever in Case 1. Then we claim that he remains in Case 1. Indeed, after the cop's move the cop is at z' on the trail of the robber at x. The robber moves to some vertex x'. If we take any G_i containing all of x, x', z', then Lemma 4.4 tells us that there is a neighbour of z' on the trail of x' in G_i. Since trails in G_i are the same as trails in G, the claim follows.

Furthermore, if Case 1 ever occurs then the robber can only visit any vertex finitely many times. Indeed, suppose that the robber has a sequence of moves $x_1, x_2, \ldots x_n, x_1$ starting and finishing at x_1, and the cop's sequence under this strategy is $y_1, y_2, \ldots, y_n, y_{n+1}$. Let G_i be chosen to contain all the x_i and y_i. Since the parent maps are consistent, we see that the cop is following exactly the winning strategy in G_i, and so in particular, by Lemma 4.4, y_{n+1} is a (strictly) more recent ancestor of x than y_1 was.

Hence, each time the robber return to x_1, the cop is closer on the robber's trail, and after some finite number of loops he catches the robber.

If Case 1 never happens, then the robber is not caught, but he is eventually forced out of a finite set of vertices forever, otherwise the cop would be able to get on his trail after a finite set of moves as described in Case 2. This finishes the proof. $\square$

It is clear that, while the consistency condition makes this proof work, it is not the 'right' condition. Indeed, even our example of a cop win that is not constructible is actually not consistently constructible, since in any subgraph containing the root the rightmost vertex in the Ks, and only that vertex, has the root as its parent.

4.5 Locally constructible graphs may not be weak cop-win

In this section we first exhibit a locally constructible graph that is not weak cop-win. Our graph is not locally finite, but by carefully modifying the way it is built up we are able to find a locally finite locally constructible graph that is not weak cop-win. The lemma below is at the heart of our construction: it allows us to pass from any graph to a constructible one.

We write P_n for the path of length n, and view its vertices as $0, 1, \cdots, n$. For a finite graph G and a positive integer n, we write $G * P_n$ for the graph with vertex set $G \times P_n$ in which (x, j) is joined to (x', j') if either $x \sim x'$ and $j \sim j'$, or $j = j' = n$.

Lemma 4.6. *For any finite graph G, the graph $G * P_n$ is constructible. Moreover, if G is not constructible, then in $G * P_n$ the cop can be forced to visit a vertex of the form (x, n) for some x before catching the robber.*

Proof. Note that the vertices with second coordinate n form a complete graph, and so can be constructed first. Once we have these vertices, we observe that the vertex $(x, n-1)$ is dominated by the vertex (x, n), and so we can now add all of the vertices with second coordinate $n-1$. Continuing in this way, we can add all the vertices, and thus the graph is constructible.

Now suppose that the graph G is not constructible. This means that the robber can avoid being caught on G. Thus, if the cop never visits a vertex with second coordinate n, we can pretend by projection that the chase happens on G, so that the robber can avoid being caught. We conclude that the cop is be forced to visit a vertex (x, n), for some x, before catching the robber. □

The next step is to observe that if we have a graph G, and we attach disjoint 4-cycles to all of its vertices, the robber will always win in this new graph regardless of the starting position, by staying on the 4-cycle associated with his starting vertex.

More precisely, let C_4 be a 4-cycle, say on vertices $0, 1, 2, 3$, and let G be any graph. We define the graph $G \cdot C_4$ on vertex set $V(G) \times \{0, 1, 2, 3\}$ by joining (x, y) to (x', y') if either $x = x'$ and y is adjacent to y', or x is adjacent to x' and $y = y' = 0$. As explained above, this graph is clearly a robber win regardless of the starting position.

Therefore, if we start with the graph C_4, which is not constructible, and look at $C_4, C_4 * P_n, (C_4 * P_n) \cdot C_4, ((C_4 * P_n) \cdot C_4) * P_n, \ldots$, then we are alternating between constructible and non-constructible graphs. To achieve locally constructibility without being a weak cop win, it is reasonable to take the union of these graphs. The intuition behind this is that, although the cop can win on each of the constructible stages, namely the ones after taking a product with P_n, he has to go a long way from the robber, as shown in Lemma 4.6. This gives the robber time to get back to the origin and then head off into an extra coordinate.

Now we make this idea precise. The reader should bear in mind that the graph $\mathcal{G}$ constructed below is precisely the 'nested union' of the above sequence of graphs.

Theorem 4.7. *There exists a graph $\mathcal{G}$ which is locally constructible, but is not a weak cop win.*

Proof. We define the graph $\mathcal{G}$ as follows. The vertex set is $C_4 \times P_6 \times C_4 \times P_6 \times C_4 \times P_6 \times \ldots$, where we insist that all but finitely many of the coordinates are 0. Let $\widehat{0}$ be the vertex where all coordinates are 0. In order to define the edges we consider three cases. Let v and v' be two vertices.

If all their C_4 coordinates are 0 and there is no P_6 coordinate in which both vertices are 6, then v is adjacent to v' if and only if they differ by at most 1 in all P_6 coordinates.

Otherwise let m_1 be the maximal C_4 coordinate in which v and v' are not both 0, and m_2 the maximal P_6 coordinate in which both v and v' are 6 (and we set $m_i = 0$ if the corresponding coordinate does not exist).

If $m_1 < m_2$, then v is adjacent to v' if and only if, after the m_2^{th} coordinate, all their P_6 coordinates differ by at most 1 – note that after the m_2^{th} coordinates all their C_4 coordinates are 0 by definition.

If $m_1 > m_2$, then v is adjacent to v' if and only if they agree on all coordinates less than m_1, differ by at most 1 in the m_1^{th} coordinate, and differ by at most 1 in all the P_6 coordinates greater that m_1.

Claim 4.8. *The graph $\mathcal{G}$ is locally constructible.*

Proof. We observe that the graph we get if we restrict to all vertices which are always zero after some particular P_6 coordinate a finite graph of the form $(\cdots(C_4 * P_6)\cdot C_4 * \cdots P_6)$ and, by Lemma 4.6, is constructible. Another way to see that this graph is constructible is to show that the cop wins on this graph. Indeed, the cop goes to level 6 (the maximum level) in the final P_6 coordinate. Let that coordinate be m. He is then able to immediately move to a vertex that agrees with the robber's vertex on the rest of the coordinates. Then, after each robber move, if the robber is at the same level as, or one below, the cop, then the cop immediately catches him. Here level means the value of the m^{th} coordinate. Otherwise the cop moves to stay above the robber on the rest of the coordinates, while reducing the m^{th} coordinate by 1. In this way the cop must catch the robber by the time the cop reaches level 0. $\square$

Claim 4.9. *The graph $\mathcal{G}$ is not weak cop-win.*

Proof. The robber's strategy is to always have all coordinates zero with at most one exception, and that exception is in a cycle coordinate. It is clear that after the cop chooses his starting position, the robber can choose a large cycle coordinate m_0 and start at the vertex with 2 in the m_0 coordinate and 0 elsewhere. Note that this implies that the robber is distance at least 2 from the cop. The robber commits to stay in this cycle (that is, all coordinates except the m_0^{th} coordinate are zero) until he reaches $\widehat{0}$, after which he enters a different cycle and the whole process repeats.

We define 3 stages of the strategy which are characterised by the state after a robber move, where m is the cycle coordinate the robber is currently committed to stay in before he gets to $\widehat{0}$.

Stage 1. The robber is not at $\widehat{0}$ and the cop's vertex has no path coordinate 6. Furthermore, either the m^{th} coordinate of the cop's vertex is 2 different from the m^{th} coordinate of the robber's vertex, or it is 1 different and the cop's vertex has a non-zero earlier coordinate.
Stage 2. The robber is not at $\widehat{0}$ and the cop's vertex has a 6 in some path coordinate.
Stage 3. The robber is at $\widehat{0}$ and the cop is at least distance 2 away from the robber.

Suppose we are in Stage 1 of the strategy, the cop is at v and the robber is at w. By the definition of the edges of $\mathcal{G}$, in one move the cop can go to a vertex v' that differs from v either in the m^{th} coordinate or in some coordinate less than m – these two cases are disjoint by construction. In either case the coordinates of v' greater than m may differ from those of v. If v' has a 6 in some P_6 coordinate then the robber does not move and we are now in Stage 2. Thus assume v' does not have a 6 in any P_6 coordinate.

If the vertex v' differs from v in some coordinates greater than m, then they have the same m^{th} coordinate and, in particular, v' and w differ in the m^{th} coordinate by at least 1. Thus the robber moves (if necessary) to a vertex w' that differs from v' by at least 2 in the m^{th} coordinate. In this case we are either in Stage 1 or, if the robber has reached $\widehat{0}$, in Stage 3.

Finally, if v and v' differ in the m^{th} coordinate, then the robber moves to a vertex w' such that the difference between the m^{th} coordinates of v' and w' is the same as the difference between the m^{th} coordinates of v and w. Again, we are either in Stage 1 or, if the robber has reached $\widehat{0}$, in Stage 3.

If we are in Stage 2 of the strategy, we observe that the cop is distance at least 6 from $\widehat{0}$. Indeed, let the cop be at v_0, and fix a minimal path from v_0 to $\widehat{0}$. Consider the maximal P_6 coordinate that is ever 6 on this path. This coordinate needs to become 0 and can only change by at most 1 at each step along the path. Thus the path has length at least 6.

It follows that the robber can reach $\widehat{0}$ without being caught in at most 2 steps – this is because his vertex has all coordinates 0 except for one cycle coordinate, and the cop, who is originally distance 6 away from $\widehat{0}$, will end up being distance at least 2 from him. We are now in Stage 3. Note that the cop being distance 5 from the origin would have sufficed – in other words, the construction could use P_5 instead of P_6.

Finally, suppose we are in Stage 3 and the cop moves somewhere. He must still be at least distance 1 from $\widehat{0}$. The robber picks a new cycle coordinate m' where m' is greater than any of the non-zero coordinates of the cop's vertex, and moves to 1 in this cycle coordinate. We are now back to Stage 1.

This strategy ensures the robber is never caught. Moreover, the robber either stays in Stage 1 after some point, which means he stays on one particular cycle forever, or he reaches Stage 3 infinitely often, so visiting $\widehat{0}$ infinitely often. We conclude that this graph is not a weak cop win. $\square$

This concludes the proof of Theorem 4.7. $\square$

The above construction gives us a locally constructible graph that is not weak cop-win. However, this graph is not locally finite, as for example the degree of $\widehat{0}$ is infinite. It is natural to ask what happens if we insist that the graph is locally finite. Does this, together with the condition that it is locally constructible, guarantee that the graph is weak cop-win? Below we answer this question negatively by modifying the previous construction so that the graph is locally finite and yet not weak cop-win.

The key extra idea is to obtain locally finiteness by attaching the (iterated) graphs $G * P_6$ from the previous construction along the vertices of an infinite path rather than all to the same vertex. However, this means that it takes the robber longer and longer to return to the origin, so rather than using $G * P_6$ each time we will have to use a more involved construction, and in particular we will need to use an increasing sequence of path lengths rather than always using P_6 when constructing the graphs.

First we make precise what we mean by the description above of 'attaching graphs along the vertices of an infinite path'. Let $(G_n)_{n\geq 0}$ be any nested sequence of finite graphs – in other words G_n is a fixed induced subgraph of G_m for all $m \geq n$. We define the *union graph* $\bigsqcup G_n$ to be the graph with vertex set the disjoint union of the vertex sets of all G_n, which we view as pairs (n, x) where $n \in \mathbb{N}$ and $x \in G_n$, with (n, x) adjacent to (n', x') if $|n - n'| \leq 1$ and $x \sim x'$.

We observe that if a particular G_k is constructible then the subgraph of $\bigsqcup G_n$ given by the vertices (m, x) with $m \leq k$ is constructible. Indeed, we first construct the graph with vertices (k, x) which, because it is isomorphic to G_k, is constructible. As before, each vertex $(k-1, x)$ is now dominated by (k, x), and so we can add the entire $k-1$ layer. Continuing in this way we add all the layers, and so the graph is constructible.

Next we define an important step in our construction of each of the graphs G_n. This is analogous to Lemma 4.6, but modified to our new setting. Let G be a finite graph. We say that G' is the *hive graph* of G of *height* n if G' has vertex set $G \times \{0, 1, \cdots, n\}$ together with a special vertex v called the *hive vertex* that is adjacent to all vertices of the form (x, n), and (x, i) is adjacent to (x', i') if $x \sim x'$ and $|i - i'| \leq 1$.

The key points of the hive construction are that, for any G, the graph G' is constructible (just start from the hive vertex, then construct layer n, then layer $n-1$, and so on down to layer 0 in turn), and that if G is not constructible then the cop cannot win without visiting the hive vertex – which is a long way from the 0-layer.

We are now in a position to define our example $\mathcal{H}$ of a locally finite locally constructible graph that is not a weak cop win. We start with G_0 as a single vertex 0. Given G_{n-1}, we form H_n by adding a new copy of C_4 at 0 (in other words, we take the disjoint union of our graph with C_4 and then identify the two vertices called 0). We then set G_n to be the hive graph of H_n of height $l_n = 2n + 5$ with hive vertex v_n. The graphs G_n are naturally nested with G_{n-1} a subset of H_n which in turn is a subset of the 0-level of G_n. Finally, we define $\mathcal{H}$ to be the union graph $\bigsqcup G_n$. We call the vertex $(0, 0)$ the *origin* and the set $S = \{(n, 0) : n \in \mathbb{N}\}$ the *spine*.

Figure 4 below shows how the graph G_2 is built up (but with $l_1 = l_2 = 3$ for readability). We start with G_0, which is the single purple vertex. Next we form H_1 by adding the blue 4-cycle. From H_1 we form the red hive graph G_1 with hive vertex v_1 and height 3. We then form H_2 by attaching the green 4-cycle to the origin (the purple vertex). Finally, we form G_2, the hive graph of H_2 with height 3 and hive vertex v_2. The dotted lines are drawn to indicate that there are edges between the 4-cycles, between the copies of H_2, and so on.

Theorem 4.10. *The graph $\mathcal{H}$ is locally finite and locally constructible, but is not a weak cop win.*

Proof. Certainly the graph is locally finite, as a vertex (n, x) is only adjacent to vertices (n, y), $(n-1, x')$ and $(n+1, x'')$, which form a finite set as G_{n-1}, G_n and G_{n+1} are finite graphs.

Claim 4.11. *The graph $\mathcal{H}$ is locally constructible.*

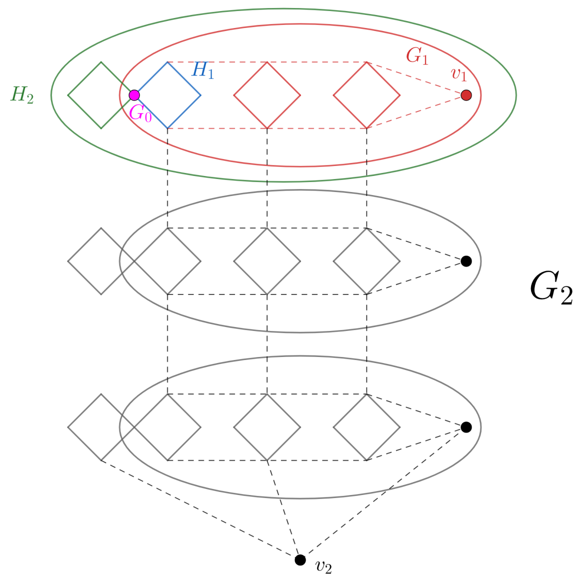


Figure 4: The graph G_2 showing G_0, H_1, G_1, H_2 as subgraphs. (Note that to keep the picture manageable we have set $l_1 = l_2 = 3$).

Proof. We saw above that any hive graph is constructible, and hence the graphs G_n are all constructible. This, combined with the above observation (about what happens when a G_k is constructible), tells us that for every n the subgraph of $\mathcal{H}$ induced by the vertices (m, x) with $m \leq n$ is constructible. Thus $\mathcal{H}$ is locally constructible. □

Claim 4.12. *The graph $\mathcal{H}$ is not a weak cop win.*

Proof. The rough idea is that if the robber is in one of the 4-cycles, say the one that appears first in G_n, then the cop can force the robber out of this cycle, but in order to do so he has to go to some hive vertex of some G_m with $m \geq n$, which means he is a long distance away from the robber. This gives the robber time to go to the origin and back out further than the cop.

However, as stated this is not correct: the cop can force the robber out of a 4-cycle by going to any of the copies of a hive vertex in later hive constructions, and these vertices can be arbitrarily far from the origin. Therefore, instead of looking at the cop's position itself, we look at how it 'projects' onto G_m. To make this idea precise we need a better understanding of the hive graphs and their properties.

Since we are dealing with several different graphs, many of which have vertices in common, in what follows, for a graph G, we denote by $d_G(x, y)$ and $d_G(z, A)$ the distance in G between two vertices x and y, and between a vertex z and a set of vertices A respectively.

Let H be a finite graph and H' a hive graph of H with hive vertex h. We define the *hive map* to be the function from $H' \setminus \{h\}$ to H that projects the vertices to the base layer – in other words (x, m) is mapped to $(x, 0)$ (and we view $(x, 0)$ as identified with x). We note that the hive map is a graph homomorphism (but is not defined for the hive vertex).

Returning to the graphs G_n used in the construction of $\mathcal{H}$, we define the *one-step projection* Q_n to be the function mapping $G_n \setminus \{v_n\}$ to G_{n-1} by first applying the hive map $G_n \setminus \{v_n\} \to H_n = G_{n-1} \cup C_4$, followed by the map $G_{n-1} \cup C_4 \to G_{n-1}$ that sends all the vertices in the C_4 to 0. It is easy to see that the one-step projection Q_n on $G_n \setminus \{v_n\}$ is a graph homomorphism.

We inductively define the *n-projection* J_n to be the map $\mathcal{H} \to G_n \cup \{v_k : k > n\}$ such that:

$$J_n((m,x)) =$$
$$\begin{cases} x & \text{if } m \leq n, \\ (m,x) & \text{if } m > n \text{ and } (m,x) \text{ is a hive vertex,} \\ J_n((m',x')) & \text{otherwise, where } (m',x') \text{ is the one-step projection of } (m,x). \end{cases}$$

It is important to note that the map J_n is almost a graph homomophism, in the sense that it only fails to be a homomorpism for vertices that reach a hive vertex in the definition; in other words J_n restricted to $J_n^{-1}(G_n)$ is a graph homomorphism. With this in mind, we classify the exceptional vertices, calling the vertices in $J_n^{-1}(v_n)$, *hive-type vertices of order* n. Note that if $J_n(x) = v_n$ then $J_m(x) = v_n$ for all $m \leq n$.

The map J_n is a projection onto G_n. At other points in the proof we will want a projection onto H_n instead of G_n, so we define $J'_n : \mathcal{H} \to H_n \cup \{v_k : k \geq n\}$ to be J_n followed by the hive map.

Lemma 4.13. *Let x be a hive-type vertex of order n and S the spine. Then $d_{\mathcal{H}}(x, S) \geq l_n + 1$.*

Proof. Fix a path from x to S. Let y be a vertex on the path P of maximum hive-type order, and suppose it has order m. Since x itself has hive-type order n we see $m \geq n$. By our choice of m the path P is in $J_m^{-1}(G_m)$, so $J_m(P)$ is a path in G_m. Since any hive vertex of order m maps to v_m under J_m, and any vertex on the spine maps to 0 under J_m, we see that the path $J_m(P)$ contains both v_m and 0. However, it is easy to see that $d_{G_m}(v_m, 0) = l_m + 1 \geq l_n + 1$, as the 'level' in the hive graph can decrease by at most 1 at each step. The result follows. □

Lemma 4.14. *Let x be a hive-type vertex of order n and suppose P is a path of length at most l_n not containing any hive-type vertex of order greater than n. Then $J'_{n+1}(P)$ does not contain the vertex 0 or any vertex of the C_4 first added in H_{n+1}.*

Proof. Since P does not contain any hive-type vertex of order greater than n, the projection map J'_{n+1} is a graph homomorphism on P: that is, $J'_{n+1}(P)$ is a path in H_{n+1}. Using Lemma 4.13 (or directly), we see that $d_{H_{n+1}}(v_n, 0) \geq d_{\mathcal{H}}(v_n, S) \geq l_n + 1$. The path starts at v_n so the result follows. □

We are now in a position to define the robber's strategy. As in the previous construction, we have several stages of this strategy that we cycle through. In other words, the strategy allows the game to move through the different stages, or eventually remain in Stage 1. Below we explain what the stages are and, given the fact that the cop and robber are in a particular stage, how the robber can force the game into a different stage (or not leave Stage 1). We view each turn as being the cop moving followed by the robber responding.

Stage 1. The robber is at vertex y in the cycle C_4 that first appears in H_m, the cop is at vertex x, and $d_{H_m}(J'_m(x), y) \geq 2$. The cop moves to a vertex x'. If x' is a hive-type vertex of order at least m we move to Stage 2. Otherwise, we see that $J'_m(x)$ and $J'_m(x')$ are neighbours in H_m. The robber stays on the cycle C_4 that first appears in H_m, moving to a vertex y' with $d_{H_m}(J'_m(x'), y') \geq 2$. In particular, the robber is not caught, and we remain in Stage 1.

Stage 2. The robber is at vertex y in the cycle C_4 that first appears in H_m, or at a point on the spine $(0, l)$ with $l < m$, and the cop is at a hive-type vertex of order $k \geq m$. The robber now goes to the spine in at most two steps, then to the origin in a further m steps. When the robber reaches the origin we move to stage 3. By Lemma 4.13 the cop's distance from the spine is at least $l_k + 1 > m + 2$, so the robber is not caught during this stage.

Stage 3. The robber is at the origin. Let k' be the maximal order of any hive-type vertex the cop visited during Stage 2. Since at the start of Stage 2 the cop was at a hive vertex of order k, we have $k' \geq k$. The robber sets off for the vertex $(k'+1, 0)$ in $H_{k'+1}$, and then to the point opposite the spine in the C_4 added at stage $k'+1$. This would take time $k'+1+2$. However, if at any point during this the cop reaches a hive vertex of order at least $k'+1$, the robber immediately switches back to Stage 2.

If the cop does not go to any such vertex then let P be the path followed by the cop during Stages 2 and 3. Since Stages 2 and 3 together take at most time $m+2+k'+1+2 \leq 2k'+5 \leq l_{k'}$, the path P has length at most $l_{k'}$. Thus, since P does not contain any hive vertex of order greater than k, Lemma 4.14 implies that $J_{k'+1}(P)$ does not contain 0 or any vertex of the C_4 first added in $H_{k'+1}$. This shows that the robber is not caught, and that this stage finishes with the robber at vertex y and the cop at vertex x with $d_{H_{k'+1}}(J_{k'+1}(x), y) \geq 2$, and we move back to Stage 1.

The game starts by the cop picking a vertex y, then the robber chooses a vertex satisfying the conditions for Stage 1. In the above strategy either the robber stays in Stage 1 after some time, which means the robber eventually stays in the same 4-cycle forever, or Stage 2 occurs infinitely often, which means the robber visits the origin infinitely often. We conclude that the graph $\mathcal{H}$ is not weak cop win. □

This concludes the proof of Theorem 4.10. □

4.6 The construction time of constructible graphs

In this section we turn to the possible ranks of a constructible graph. Recall that the rank (or construction time) of a constructible graph G is the least order-type of any construction ordering for G. It is easy to find graphs with construction time n, where n is any positive integer–for example a path with n vertices. By taking an infinite path we can also achieve construction time ω.

The next step is to ask if there exists a graph with construction time $\omega + 1$ – in other words we need to make infinitely many extensions and then one more at the end to be able to finish the construction.

This was achieved by Evron, Solomon and Stahl [18]. In fact, they showed that the set of construction times (of countable graphs) is unbounded in the countable ordinals. They asked, more generally, which countable ordinals can be the construction time of a graph? In this section we answer this question by constructing a graph with construction time γ, where γ is any countable ordinal.

We start by giving a graph of rank $\omega + 1$. We mention that this result will be contained in our general result below (and in that general result the construction will actually be slightly different) – we include it here to illustrate in a simpler setting how the graph K can be used.

We define the graph G as follows. We take countably infinitely many disjoint copies of K, say K_i for each positive integer i, and two additional vertices which we call A and B. Let x_i and y_i be the vertices of K_i corresponding to x and y in K. We join A to x_i and y_i for every i, and B to x_i for every i. We also join A and B. The graph G is pictured below.

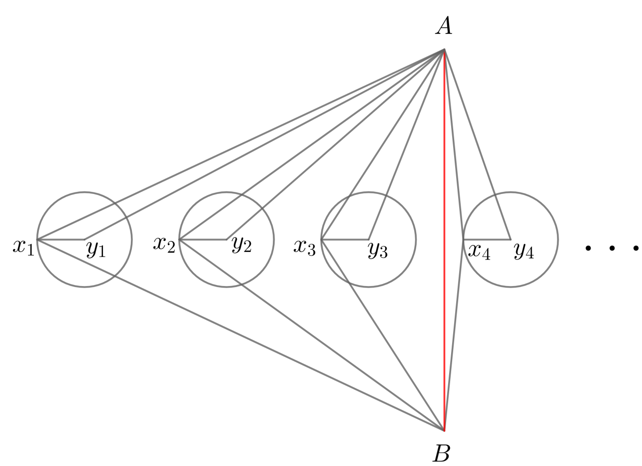


Figure 5: The graph G for Theorem 4.15.

Theorem 4.15. *The construction time of G is $\omega + 1$.*

Proof. To see that G is constructible in time $\omega + 1$ we begin with A, then add each copy of K in turn in the following construction order: first y, then z and z' (both with parent y), then w and t with parent z and t' with parent z', and finally x with parent y. This is valid even with A already present, since x has parent y and both of these vertices in a copy of K are joined to A.
Finally, after doing all the above we add B with parent A: this is allowed since all neighbours of B are also neighbours of A (and B and A are adjacent).

On the other hand, we cannot construct G in time ω. Indeed, suppose for a contradiction that there is a way to construct the graph in time ω. This implies that the vertex B must be constructed at some time t, where t is a natural number. Since t is finite, at time t we must have some copy of K with no vertices constructed yet. Let K_i be such a copy. By Lemma 4.1, x_i must be the last vertex in K_i to be added, and its parent must be y_i. But this is impossible since B is already present, and B is a neighbour of x_i while y_i is not. □

The above result tells us that any ordinal less or equal than $\omega + 1$ can be the construction time of some graph. We now prove that any countable ordinal can be achieved. We need the following simple lemma.

Lemma 4.16. *Let α be a (non-zero) countable limit ordinal. Then there are pairwise disjoint subsets S_i of α of order type α_i for all i, where $\alpha_1, \alpha_2 \cdots$ are the ordinals less than α.*

Proof. Since α is a limit, we know that $\alpha = \omega \cdot \beta$ for some ordinal β. Since ω contains infinitely many disjoint copies of itself, it follows that $\omega \cdot \beta$ contains infinitely many disjoint copies of $\omega \cdot \beta$ too. We conclude that α contains infinitely many disjoint sets of order type α. Let $Q_1, Q_2, \cdots$ be such a collection. Since Q_i has order type α, it has an initial segment S_i of order type α_i, as required. □

The following is the key result.

Lemma 4.17. *Let λ be an ordinal of the form $\lambda = \alpha + 6n + 1$ where n is a non-negative integer and α is a (possibly zero) countable limit ordinal. Then there exists a constructible graph G_λ with construction time λ. Moreover, G_λ has two vertices, A_λ and B_λ, such that in any construction order B_λ must be added last, and there exists a construction order of time λ that starts with A_λ. Furthermore, A_λ is joined to B_λ and, provided $n \geq 1$, B_λ is not dominated by A_λ.*

Proof. We proceed by induction. First we note that if we have found $G_{\alpha+6n+1}$ with the above properties, except possibly for the condition that $A_{\alpha+6n+1}$ does not dominate $B_{\alpha+6n+1}$, then we can find such graph for $\alpha + 6(n+1) + 1$ by adding a disjoint copy of K, identifying $B_{\alpha+6n+1}$ with the y of this copy and joining x to $A_{\lambda+6n+1}$. We set $B_{\alpha+6(n+1)+1}$ to be the x of the K copy and $A_{\alpha+6(n+1)+1} = A_{\alpha+6n+1}$. By using Lemma 4.1 it is easy to check that all properties are satisfied. Moreover, by Lemma 4.1 again, $B_{\alpha+6(n+1)+1}$ is not dominated by $A_{\alpha+6(n+1)+1}$.

To start with, the one-point graph satisfies the conditions for $\lambda = 1$. So to finish the proof we have to show that such graphs exist for all $\lambda = \alpha + 1$ where α is a countable non-zero limit ordinal. So let $\alpha \geq \omega$ be a countable non-zero limit.

By induction we may assume that such graphs exist for all ordinals $\beta < \lambda = \alpha+1$ of the form $\beta = \gamma + 6m + 1$, where γ is a limit ordinal and $m \geq 1$ is a positive integer. To obtain G_λ we take a copy of each G_β and identify all the points A_β to a single vertex, which is our new A_λ. We also add a new vertex B_λ which we join to A_λ and all the vertices B_β.

To see that B_λ has to come last in any construction ordering, suppose that $v \neq B_\lambda$ is the last vertex added. By the induction hypothesis, v has to be one of the vertices B_δ for $\delta < \lambda$. This vertex must be dominated by a neighbour of B_λ (since they are joined), or by B_λ itself. The other B_β vertices are not joined to B_δ so they cannot dominate it. Also, by the induction hypothesis we know that A_δ does not dominate B_δ, and soA_λ does not dominate B_δ either. Finally, since the neighbours of B_λ are a subset of the neighbours of A_λ, B_λ cannot dominate B_δ. This is a contradiction. So indeed B_λ must come last.

It is clear that the construction time of $G_\lambda \setminus \{B_\lambda\}$ is at least α because, when a B_β for some $\beta < \alpha$ is added, the entire G_β has to be constructed, which must take time at least β. Since B_λ comes at the very end, G_λ has construction time at least $\alpha + 1 = \lambda$.

To see that $G_\lambda \setminus \{B_\lambda\}$ has construction time at most α, we use Lemma 4.16. Indeed, let S_i be disjoint subsets of α of order type α_i. We view the union of the S_i as our (well-ordered) set of construction times, and at each time in S_i we construct the corresponding vertex of G_{α_i}. This gives a construction of time at most α. Adding B_λ after this takes one more step. We conclude that G_λ has indeed construction time at most λ,as required. The other properties are straightforward to check. $\square$

We remark that, alternatively, we could have started the induction at $\omega + 1$, using the graph in Theorem 4.15.

We are now ready to prove the following.

Theorem 4.18. *For every countable ordinal $\lambda > 0$, there exists a constructible graph with construction time λ.*

Proof. We know that for every positive integer n a finite path with n vertices, or indeed any constructible graph on n vertices, gives us construction time n, and an infinite ray gives us ω. From Lemma 4.17 we have that such graphs exist for all the ordinals of the form $\alpha + 6n + 1$ where α is a countable non-zero limit ordinal and n is a non-negative integer. Therefore we are left to show that such graphs exist for non-zero countable limit ordinals and for ordinals of the form $\alpha + 6n + i$ where $i \in \{2, 3, 4, 5, 6\}$ and α is a countable non-zero limit.

Suppose $\lambda = \alpha + 6n + i$ where α is a countable non-zero limit, n a non-negative integer and $2 \leq i \leq 6$. We take the graph $G_{\alpha+6n+1}$ constructed in Lemma 4.17 and add say a path of $i - 1$ vertices attached to the vertex $B_{\alpha+6n+1}$.

Finally, suppose λ is a countable non-zero limit ordinal. In this case we note that the graph $G_{\lambda+1} \setminus \{B_{\lambda+1}\}$ is a suitable choice. □

4.7 Open problems

The obvious open problem is to classify which graphs are weak-cop wins.

Question 4.19. *Which graphs are weak cop wins?*

There is also the question of which graphs are actual cop wins. However, as there are so many constructible graphs that are not cop wins (e.g. $\mathbb{Z}$), and as we have seen there is a graph that is a cop win and not constructible, a structural classification is very open.

There are also even weaker notions of win that we could consider: for example, we could view it as a win for the cop if he can force the robber to leave (but possibly return to) any finite set.

Question 4.20. *Which graphs have the property that the cop has a strategy that ensures that, given any finite set of vertices, the robber must leave this set at some point (although he may return to this set later), or get caught?*

Note that if there is a such a strategy for each individual finite set then, by concatenating these (necessarily finite time) strategies, we do obtain a single strategy that works for all finite sets. Obviously any locally constructible graph has this property, but we do not know whether the converse holds. The example of a graph that is locally constructible but not a weak cop win does show that this is strictly weaker notion than that of a weak cop win.

Finally, we have seen that there are graphs where the robber can avoid being trapped in one end of the graph (recall the doubly infinite chain of copies of K described at the end of Section 4.3). In particular, that graph is a weak cop win in which the robber can return to a specified vertex an arbitrarily long time after he first visited it. However, we do not know the answer to the following question.

Question 4.21. *Is there a graph G which is a weak cop win but such that the robber can guarantee to revisit his initial vertex v after an arbitrarily long time, and then guarantee to revisit v again after another arbitrarily long time?*

More precisely, for each cop starting position the robber has a starting position v such that, for every pair of positive integers m and n, the robber has a strategy that ensures that he does not get caught and either he stays in some finite set forever or he returns to v at some time $t \geq m$ and also at some time $s \geq t + n$.

5 Small Sets in Union-Closed Families

5.1 Introduction

If X is a set, a family $\mathcal{F}$ of subsets of X is said to be *union-closed* if the union of any two sets in $\mathcal{F}$ is also in $\mathcal{F}$. The union-closed conjecture (a conjecture of Péter Frankl [20]) states that if X is a finite set and $\mathcal{F}$ is a union-closed family of subsets of X (with $\mathcal{F} \neq \{\emptyset\}$), then there exists an element $x \in X$ such that x is contained in at least half of the sets in $\mathcal{F}$. Despite the efforts of many researchers over the last forty-five years, and a recent Polymath project [1] aimed at resolving it, this conjecture remains open. We mention that there has been remarkable recent progress towards this conjecture, by Gilmer [24], who showed that there exists $c > 0$ such that for any union-closed family there exists an element of the ground set contained in a proportion of at least c of the sets.

The conjecture has been proved under very strong constraints on the ground-set X or the family $\mathcal{F}$; for example, Balla, Bollobás and Eccles [5] proved it in the case where $|\mathcal{F}| \geq \frac{2}{3}2^{|X|}$; more recently, Karpas [36] proved it in the case where $|\mathcal{F}| \geq (\frac{1}{2} - c)2^{|X|}$ for a small absolute constant $c > 0$; and it is also known to hold whenever $|X| \leq 12$ or $|\mathcal{F}| \leq 50$, from work of Vučković and Živković [54] and of Roberts and Simpson [49]. For general background and a wealth of further information on the union-closed conjecture see the survey of Bruhn and Schaudt [12].

As usual, if X is a set we write $\mathcal{P}(X)$ for its power-set. If X is a finite set and $\mathcal{F} \subset \mathcal{P}(X)$ with $\mathcal{F} \neq \emptyset$, we define the *frequency* of x (with respect to $\mathcal{F}$) to be $\gamma_x = |\{A \in \mathcal{F} : \ x \in A\}|/|\mathcal{F}|$, i.e., γ_x is the proportion of members of X that contain x. If a union-closed family contains a 'small' set, what can we say about the frequencies in that set?

If a union-closed family $\mathcal{F}$ contains a singleton, then that element clearly has frequency at least $1/2$, while if it contains a set S of size 2 then, as noted by Sarvate and Renaud [50], some element of S has frequency at least $1/2$. However, they also gave an example of a union-closed family $\mathcal{F}$ whose smallest set S has size 3 and yet where each element of S has frequency below $1/2$. Generalising a construction of Poonen [48], Bruhn and Schaudt [12] gave, for each $k \geq 3$, an example of a union-closed family with (unique) smallest set of size k and with every element of that set having frequency below $1/2$.

However, in these and all other known examples, there is always some element of a minimal-size set having frequency at least $1/3$. So it is natural to ask if there is really a constant lower bound for these frequencies.

Our aim in this chapter is to show that this is not the case.

Theorem 5.1. *For any positive integer k, there exists a union-closed family in which the (unique) smallest set has size k, but where each element of this set has frequency*

$$(1 + o(1))\frac{\log k}{2k}.$$

(All logarithms in this chapter are to base 2. Also, as usual, the $o(1)$ denotes a function of k that tends to zero as k tends to infinity.) The proof of Theorem 5.1 is by an explicit construction.

Theorem 5.1 is asymptotically sharp, in view of results of Wójcik [58] and Balla [4]: Wójcik showed that if S is a set of size $k \geq 1$ in a finite union-closed family, then the average frequency of the elements in S is at least c_k, where $k \cdot c_k$ is defined to be the minimum average set-size over all union-closed families whose largest set contains k elements, and Balla showed that $c_k = (1 + o(1))\frac{\log k}{2k}$, confirming a conjecture of Wójcik from [58]. Therefore our construction is an extremal example, achieving the optimal c_k.

Remarkably, there are union-closed families containing small sets, even sets of size 3, for which we have been unable to verify the union-closed conjecture. We give some examples at the end of the chapter.

5.2 Small sets in union-closed families

For our construction, we need the following 'design-theoretic' lemma.

Lemma 5.2. *For any positive integers $k > t$ there exist infinitely many positive integers d such that t divides dk and the following holds. If X is a set of size dk/t, then there exists a family $\mathcal{A} = \{A_1, \ldots, A_k\}$ of k d-element subsets of X, such that each element of X is contained in exactly t sets in $\mathcal{A}$, and for $2 \le r \le t$, any r distinct sets in $\mathcal{A}$ have intersection of size*

$$d\frac{(t-1)(t-2)\ldots(t-r+1)}{(k-1)(k-2)\ldots(k-r+1)},$$

i.e.

$$|A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_r}| = d\frac{(t-1)(t-2)\ldots(t-r+1)}{(k-1)(k-2)\ldots(k-r+1)}$$

for any $1 \le i_1 < i_2 < \ldots < i_r \le k$.

Proof. Let q be a positive integer, and set $d = \binom{k-1}{t-1}q^t$; we will take $|X| = \binom{k}{t}q^t$. Partition $[qk]$ into k sets, $B_1, B_2, \ldots, B_k$ say, each of size q; we call these sets 'blocks'. We let X be the set of all t-element subsets of $[qk]$ that contain at most one element from each block. For each $i \in [k]$ we let A_i be the family of all sets in X that contain an element from the block B_i. Clearly, $|A_i| = \binom{k-1}{t-1}q^t = d$ for each $i \in [k]$, and each element of X appears in exactly t of the A_i. Also, for example $A_i \cap A_j$ consists of all sets in X that contain both an element from the block B_i and an element from the block B_j, so

$$|A_i \cap A_j| = \binom{k-2}{t-2}q^t = \binom{k-1}{t-1}q^t\frac{t-1}{k-1} = d\frac{t-1}{k-1}.$$

It is easy to check that the other intersections also have the claimed sizes. □

We remark that, in what follows, it is vital that the integer d in Lemma 5.2 can be taken to be arbitrarily large as a function of k and t.

Proof of Theorem 5.1. We define $n = dk/t + k$, we take $d \in \mathbb{N}$ as in the above lemma, and we let $X = [dk/t]$; the claim yields a family $\mathcal{A} = \{A_1, \ldots, A_k\}$ of k d-element subsets of $X = [dk/t]$ such that each element of $[dk/t]$ is contained in exactly t of the sets in $\mathcal{A}$, and for any $2 \le r \le t$, any r distinct sets in $\mathcal{A}$ have intersection of size

$$d\frac{(t-1)(t-2)\ldots(t-r+1)}{(k-1)(k-2)\ldots(k-r+1)}.$$

Write $m = dk/t$. We take $\mathcal{F} \subset \mathcal{P}([n])$ to be the smallest union-closed family containing the k-element set $\{m+1, \ldots, m+k\}$ and all sets of the form $\{m+i\} \cup (X \setminus \{x\})$ where $i \in [k]$ and $x \in A_i$.

For brevity, we write $S_0 = \{m+1, m+2, \ldots, m+k\}$. We will show that each element of S_0 has frequency

$$(1+o(1))\frac{\log k}{2k},$$

provided t and d are chosen to be appropriate functions of k; moreover, with these choices, S_0 will be the smallest set in $\mathcal{F}$.

Clearly, $\mathcal{F}$ contains S_0, all sets of the form $S_0 \cup (X \setminus \{x\})$ for $x \in X$, all sets of the form $R \cup X$ where R is a nonempty subset of S_0, and finally all sets of the form $R \cup (X \setminus \{x\})$, where $R = \{m+i_1, \ldots, m+i_r\}$ is a nonempty r-element subset of S_0 and $x \in A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_r}$, for $1 \le r \le t$. It is easy to see that the family $\mathcal{F}$ contains no other sets.

It follows that

$$\begin{aligned}
|\mathcal{F}| &= 1 + dk/t + (2^k - 1) + \sum_{r=1}^{t} \binom{k}{r} d \frac{(t-1)(t-2)\ldots(t-r+1)}{(k-1)(k-2)\ldots(k-r+1)} \\
&= dk/t + 2^k + \frac{dk}{t}\sum_{r=1}^{t}\binom{t}{r} \\
&= dk/t + 2^k + \frac{dk}{t}(2^t - 1) \\
&= 2^k + \frac{dk2^t}{t}.
\end{aligned}$$

On the other hand, the number of sets in $\mathcal{F}$ that contain the element $m+1$ is equal to

$$\begin{aligned}
&1 + dk/t + 2^{k-1} + \sum_{r=1}^{t}\binom{k-1}{r-1} d \frac{(t-1)(t-2)\ldots(t-r+1)}{(k-1)(k-2)\ldots(k-r+1)} \\
&= 1 + dk/t + 2^{k-1} + d\sum_{r=1}^{t}\binom{t-1}{r-1} \\
&= 1 + dk/t + 2^{k-1} + 2^{t-1}d.
\end{aligned}$$

It follows that the frequency of $m+1$ (or, by symmetry, of any other element of S_0) equals

$$\frac{1 + kd/t + 2^{k-1} + 2^{t-1}d}{2^k + dk2^t/t} = \frac{(1+2^{k-1})/d + k/t + 2^{t-1}}{2^k/d + k2^t/t}.$$

To (asymptotically) minimise this expression, we take $t = \lfloor \log k \rfloor$ and $d \to \infty$ (for fixed k); this yields a union-closed family in which the (unique) smallest set (namely S_0) has size k, and every element of that set has frequency

$$(1+o(1))\frac{\log k}{2k},$$

proving the theorem. $\square$

5.3 An open problem

We now turn to some explicit examples of union-closed families containing small sets for which we have been unable to establish the union-closed conjecture. For simplicity, we concentrate on the most striking case, when the family contains a set of size 3, and indeed is generated by sets of size 3.

Our families live on ground-set $\mathbb{Z}_n^2$, the $n \times n$ torus.

Question 5.3. *Let $n \in \mathbb{N}$ and let $R \subset \mathbb{Z}_n$ with $|R| = 3$. Does the union-closed conjecture hold for the union-closed family $\mathcal{F}$ of subsets of $\mathbb{Z}_n^2$ generated by all the translates of $R \times \{0\}$ and of $\{0\} \times R$?*

(Here, as usual, we say a union-closed family $\mathcal{F}$ is *generated by* a family $\mathcal{G}$ if it consists of all unions of sets in $\mathcal{G}$.)

Perhaps the most interesting case is when n is prime. In that case we may assume that $R = \{0, 1, r\}$ for some r, and so one feels that the verification of the union-closed conjecture should be a triviality, but it seems not to be. Note that all the families in Question 5.3 are transitive families, in the sense that all points 'look the same', so that the union-closed conjecture is equivalent to the assertion that the average size of the sets in the family is at least $n^2/2$.

We mention that the corresponding result in $\mathbb{Z}_n$ (in other words, the union-closed family on ground-set $\mathbb{Z}_n$ generated by translates of R) is known to hold: this is proved in [2].

We have verified the special case of Question 5.3 where $R = \{0, 1, 2\}$. A sketch of the proof is as follows. Assume that $n \geq 6$, and let $\mathcal{F} \subset \mathcal{P}(\mathbb{Z}_n^2)$ be the union-closed family generated by all translates of $\{0, 1, 2\} \times \{0\}$ and of $\{0\} \times \{0, 1, 2\}$ (we call these translates *3-tiles*, for brevity). Let $C = \{0, 1, 2, 3\}^2$, a 4×4 square. Consider the bipartite graph $H = (\mathcal{X}, \mathcal{Y})$ with vertex-classes $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X}$ consists of all subsets of C with size less than 8 that are intersections with C of sets in $\mathcal{F}$, $\mathcal{Y}$ consists of all subsets of C with size greater than 8 that are intersections with C of sets in $\mathcal{F}$, and we join $S \in \mathcal{X}$ to $S' \in \mathcal{Y}$ if $|S'| + |S| \geq 16$ and $S' = S \cup U$ for some union U of 3-tiles that are contained within C. It can be verified (by computer) that H has a matching $m : \mathcal{X} \to \mathcal{Y}$ of size $|\mathcal{X}| = 16520$. Such a matching m gives rise to an injection

$$f : \{S \in \mathcal{F} : \ |S \cap C| < |C|/2\} \to \{S \in \mathcal{F} : \ |S \cap C| > |C|/2\}$$

given by

$$f(S) = (S \setminus C) \cup m(S \cap C)$$

with the property that $|S \cap C| + |f(S) \cap C| \geq |C|$ for all $S \in \mathcal{F}$ with $|S \cap C| < |C|/2$. It follows that a uniformly random subset of $\mathcal{F}$ has intersection with $|C|$ of expected size at least $|C|/2$, which in turn implies that there is an element of C with frequency at least $1/2$ (and in fact, since $\mathcal{F}$ is transitive, every element has frequency at least $1/2$).

We remark that this proof does not work if one tries to replace $C = \{0,1,2,3\}^2$ by $\{0,1,2\}^2$, as the resulting bipartite graph $H' = (\mathcal{X}', \mathcal{Y}')$ does not contain a matching of size $|\mathcal{X}'|$.

We remark also that it would be nice to find a non-computer proof of the above result.

6 Bibliography

[1] *Polymath11: Frankl's union-closed conjecture*, `https://gowers.wordpress.com/2016/01/29/func1-strengthenings-variants-potential-counterexamples/`.

[2] J. Aaronson, D. Ellis, and I. Leader, *A note on transitive union-closed families*, The Electronic Journal of Combinatorics **28** (2021).

[3] I. Đanković and M.-R. Ivan, *Saturation for Small Antichains*, The Electronic Journal of Combinatorics **30** (2023).

[4] I. Balla, *Minimum density of union-closed families*, arχiv: 1106.0369.

[5] I. Balla, B. Bollobás, and T. Eccles, *Union-closed families of sets*, Journal of Combinatorial Theory, Series A **120** (2013), 531–544.

[6] P. Bastide, C. Groenland, M.-R. Ivan, and T. Johnston, *Polynomial Upper Bound for Induced Saturation Numbers*, To appear (2023).

[7] P. Bastide, C. Groenland, H. Jacob, and T. Johnston, *Exact antichain saturation numbers via a generalisation of a result of Lehman-Ron*, arχiv: 2207.07391 (2022).

[8] V. Bergelson, N. Hindman, and I. Leader, *Additive and multiplicative Ramsey theory in the reals and the rationals*, Journal of Combinatorial Theory, Series A **85** (1999), 41–68.

[9] A. Bonato, P. Golovach, G. Hahn, and J. Kratochvíl, *The capture time of a graph*, Discrete Mathematics **309** (2009), 5588–5595.

[10] A. Bonato and R. Nowakowski, *The Game of Cops and Robbers on Graphs*, American Mathematical Society (2011), ISBN-13 978-0821853474.

[11] M. Bowen and M. Sabok, *Monochromatic products and sums in the rationals*, arχiv: 2210.12290 (2022).

[12] H. Bruhn and O. Schaudt, *The journey of the union-closed sets conjecture*, Graphs and Combinatorics **31** (2015), 2043–2074.

[13] M. Chastand, F. Laviolette, and N. Polat, *On constructible graphs, infinite bridged graphs and weakly cop-win graphs*, Discrete Mathematics **224** (2000), 61–78.

[14] A. de Luca and L. Q. Zamboni, *On some variations of coloring problems for infinite words*, Journal of Combinatorial Theory, Series A **137** (2016), 166–178.

[15] W. Deuber, N. Hindman, I. Leader, and H. Lefmann, *Infinite partition regular matrices*, Combinatorica **15** (1995), 333–355.

[16] R. Diestel, *Graph Theory*, Springer-Verlag, Heidelberg Graduate Texts in Mathematics **173** (2016/17).

[17] D. Ellis, M.-R. Ivan, and I. Leader, *Small Sets in Union-Closed Families*, The Electronic Journal of Combinatorics **30** (2023).

[18] L. Evron, R. Solomon, and R. Stahl, *Dominating orders, vertex pursuit games and computability theory*, (2021), `https://www2.math.uconn.edu/~solomon/research/PursuitDraft.pdf`.

[19] M. Ferrara, B. Kay, L. Kramer, R. R. Martin, B. Reiniger, H. C. Smith, and E. T. Sullivan, *The saturation number of induced subposets of the Boolean lattice*, Discrete Mathematics **340** (2017), 2479–2487.

[20] P Frankl, *Extremal set systems*, Handbook of combinatorics **2** (1995), 1293–1329.

[21] A. Freschi, S. Piga, M. Sharifzadeh, and A. Treglown, *The induced saturation problem for posets*, arχiv: 2207.03974 (2022).

[22] D. Gerbner, B. Keszegh, N. Lemons, C. Palmer, D. Pálvölgyi, and B. Patkós, *Saturating sperner families*, Graphs and Combinatorics **29** (2013), no. 5, 1355–1364.

[23] D. Gerbner and B. Patkós, *Extremal Finite Set Theory*, CRC Press, 2018.

[24] J. Gilmer, *A constant lower bound for the union-closed sets*, arχiv: 2211.09055 (2022).

[25] G. Hahn, F. Laviolette, N. Sauer, and R. Woodrow, *On cop-win graphs*, Discrete Mathematics **258** (2002), 27–41.

[26] N. Hindman, *Finite sums from sequences within cells of a partition of* $\mathbb{N}$, Journal of Combinatorial Theory, Series A **17** (1974), 1–11.

[27] ———, *Partitions and pairwise sums and products*, Journal of Combinatorial Theory, Series A **37** (1984), 46–60.

[28] N. Hindman, I. Leader, and D. Strauss, *Extensions of infinite partition regular systems*, The Electronic Journal of Combinatorics **22** (2015).

[29] N. Hindman and D. Strauss, *Algebra in the Stone-Čech Compactification: Theory and Applications, 2nd edition*, Walter de Gruyter & Co., Berlin, 2012.

[30] V. Isler, S. Kannan, and S. Khanna, *Randomized pursuit-evasion with limited visibility*, Technical Reports (CIS) (2003).

[31] M.-R. Ivan, *Saturation for the Butterfly Poset*, Mathematika **66** (2020), 806–817.

[32] ______, *Minimal Diamond-Saturated Families*, Contemporary Mathematics **3** (2022), 81–88.

[33] M.-R. Ivan, N. Hindman, and I. Leader, *Some New Results on Monochromatic sums and Products in the Rationals*, New York Journal of Mathematics **29** (2023), 301–322.

[34] M.-R. Ivan, I. Leader, and M. Walters, *Constructible Graphs and Pursuit*, Theoretical Computer Science **930** (2022), 196–208.

[35] M.-R. Ivan, I. Leader, and L. Q. Zamboni, *A Ramsey Characterisation of Eventually Periodic Words*, Bulletin of the London Mathematical Society **54** (2022), 2437–2455.

[36] I. Karpas, *Two results on union-closed families*, arχiv: 1708.01434.

[37] A. Kechris, *Classical descriptive set theory*, Springer-Verlag, Graduate Texts in Mathematics **156** (2012).

[38] B. Keszegh, N. Lemons, R. R. Martin, D. Pálvölgyi, and B. Patkós, *Induced and non-induced poset saturation problems*, Journal of Combinatorial Theory, Series A **184** (2021).

[39] I. Leader, P. A. Russell, and M. J. Walters, *Transitive sets in Euclidean Ramsey theory*, Journal of Combinatorial Theory, Series A **119** (2012), 382–396.

[40] F. Lehner, *Pursuit evasion on infinite graphs*, Theoretical Computer Science **655** (2016), 30–40.

[41] R. R Martin, H. C. Smith, and S. Walker, *Improved bounds for induced poset saturation*, The Electronic Journal of Combinatorics **27** (2020).

[42] K. Milliken, *Ramsey's theorem with sums or unions*, Journal of Combinatorial Theory, Series A **18** (1975), 276–290.

[43] J. Moreira, *Monochromatic sums and products in* $\mathbb{N}$, Annals of Mathematics **185** (2017), 1069–1090.

[44] R. Nowakowski and P. Winkler, *Vertex-to-vertex pursuit in a graph*, Discrete Mathematics **43** (1983), 235–239.

[45] N. Polat, *Retract-collapsible graphs and invariant subgraph properties*, Journal of Graph Theory **19** (1995), 25–44.

[46] ______, *On infinite bridged graphs and strongly dismantlable graphs*, Discrete Mathematics **211** (2000), 153–166.

[47] ———, *On constructible graphs, locally Helly graphs, and convexity*, Journal of Graph Theory **43** (2003), 280–298.

[48] B. Poonen, *Union-closed families*, Journal of Combinatorial Theory, Series A **59** (1992), 253–268.

[49] I. Roberts and J. Simpson, *A note on the union-closed sets conjecture*, The Australasian Journal of Combinatorics **47** (2010), 265–267.

[50] D.G. Sarvate and J.-C. Renaud, *Improved bounds for the union-closed sets conjecture*, Ars Combinatoria **29** (1990), 181–185.

[51] M. P. Schützenberger, *Quelques problèmes combinatoires de la théorie des automates*, Cours professé à l'Institut de Programmation en 1966/67, notes by J.-F. Perrot, `http://igm.univ-mlv.fr/bers-tel/Mps/Cours/PolyRouge.pdf`.

[52] R. Stahl, *Computability and the game of cops and robbers on graphs*, Archive for Mathematical Logic (2021), 1–25.

[53] A. Taylor, *A canonical partition relation for finite subsets of ω*, Journal of Combinatorial Theory, Series A **21** (1976), 137–146.

[54] B. Vučković and M. Živković, *The 12-element case of Frankl's conjecture*, IPSI Transactions on Advanced Research (2017).

[55] C. Wojcik, *On a new conjecture about super-monochromatic factorisations and ultimate periodicity*, arχiv: 1802.08670.

[56] C. Wojcik and L. Q. Zamboni, *Coloring problems for infinite words*, Sequences, Groups and Number Theory (2018), 213–231.

[57] ———, *Monochromatic factorisations of words and periodicity*, Mathematika **64** (2018), 115–123.

[58] P. Wójcik, *Density of union-closed families*, Discrete Mathematics **105** (1992), 259–267.