Spring 4-2017

Results in Ramsey Theory and Probabilistic Combinatorics

Mikhail Lavrov
Carnegie Mellon University

Carnegie Mellon University

MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy
in Algorithms, Combinatorics and Optimization

TITLE Results in Ramsey Theory and Probabilistic Combinatorics

PRESENTED BY Mikhail Lavrov

ACCEPTED BY THE DEPARTMENT OF Mathematical Sciences

Po-Shen Loh	May 2017
MAJOR PROFESSOR	DATE
Thomas Bohman	May 2017
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca W. Doerge	May 2017
DEAN	DATE

Results in Ramsey Theory and Probabilistic Combinatorics

Mikhail Lavrov

April 2017

Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Po-Shen Loh
Boris Bukh
Kevin Milans (West Virginia University)
Ryan O'Donnell

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

For my grandfather, who taught me about algorithms

Abstract

This thesis addresses several questions in Ramsey theory and in probabilistic combinatorics.

We begin by considering several questions related to the Hales–Jewett Theorem, a central result in the study of Ramsey theory. We use bounds due to Shelah [46] to attack a geometric Ramsey problem due to Graham and Rothschild [28], improving on the the previous bound known as Graham's number (which at one point held the Guinness world record for the largest number used in a mathematical proof). Extending ideas developed in studying that question, we obtain a bound of less than 10^{11} on the Hales–Jewett number $\mathrm{HJ}(4, 2)$.

Next, we consider problems in random graphs, and especially the use of random structures in solving extremal problems. We begin by a classical random graph result, analyzing an invariant called the game chromatic number for the random 3-regular graph $\mathcal{G}_{n,3}$. Then, we extend the increasing paths problem posed by Graham and Kleitman [27] to the random setting. Finally, we consider a purely graph-theoretic problem, that of distance-uniform graphs, introduced by Alon et al. in [2]. Here, we prove a marked difference between the behavior of random graphs which are distance-uniform, and worst-case behavior of distance-uniform graphs, constructing a family of distance-uniform graphs which are separated by an exponential gap from the random example.

Acknowledgments

My Ph.D. advisor, Po-Shen Loh, played a key role in my mathematical development; without his guidance and support neither this thesis nor my mathematical career would have existed. I further owe gratitude to the many other faculty at CMU who have helped me along the path to doing independent research in mathematics. In addition, I would certainly not have made it this far without the encouragement and friendship of my fellow graduate students.

On a different note, I want to thank Canada/USA Mathcamp, a summer program without which I would not be the mathematician I am today. It was there that, as a high school student, I first saw a glimpse of the breadth and beauty of mathematics, and my liking for the subject turned into a passion for it. Teaching there in recent years has not only been a way to give back to that community, but has helped me sharpen and refine the ideas which went into the making of this thesis.

Finally, I would like to thank my family, whose patience in dealing with having a mathematician for a brother, son, or grandson has been incredible and appreciated.

Contents

Chapter 1

Introduction

1.1 Ramsey theory

As a branch of extremal combinatorics, Ramsey theory is often characterized in terms of the flavor of its results: in any sufficiently large structure, we can find a small substructure which is highly ordered in some sense. The name originates from Ramsey's theorem, one of the earliest examples of such a result. Originally proved by Ramsey in [44] as a result in formal logic, this theorem is now commonly stated a statement about edge-colorings of the complete graph K_n:

Theorem 1.1.1 (Ramsey's theorem)**.** *For any integer k, there exists a sufficiently large n so that the following holds. Whenever each edge of the complete graph K_n is assigned one of 2 colors, there must be a set S of k vertices such that the complete subgraph induced by S is monochromatic: all $\binom{k}{2}$ of its edges have been assigned the same color.*

We are often interested in quantifying the "sufficiently large n" of such a result. In the example of Theorem 1.1.1, as in many others, once the desired statement holds for $n = n_0$, it holds for all $n > n_0$ as well (since K_n will contain K_{n_0} as a subgraph). It's therefore meaningful to ask: what is the least value of n, as a function of k, for which the statement holds? In the case of Ramsey's theorem, this is known as the (diagonal) Ramsey number $R(k)$. A sufficiently concrete proof of Theorem 1.1.1 will imply an upper bound on $R(k)$, and we may continue work

on the problem in the hope of finding better upper bounds.

If we declare Ramsey theory to be the study of combinatorial results of this flavor, it is not at first obvious that what we get is anything more than a collection of theorems. In fact, the theorems are connected by an intricate network of reductions that allow us to deduce upper or lower bounds on one Ramsey-type result from bounds on another. It is then natural to ask: can we find some universal Ramsey-type theorem from which all or most others will follow?

One of the first attempts in this direction was the Graham–Rothschild parameter sets theorem [28]. This is a very similar statement to that of Theorem 1.1.1 in spirit, replacing the notions of "edge" and "complete subgraph" by a more general notion of "k-parameter set". The full statement of the theorem is complex, and is given as Theorem 2.1.1 in Chapter 2.

In practice, it was found that the Hales–Jewett theorem, which (like many other results) is a special case of the parameter sets theorem, captures enough of its complexity to be equally useful in deducing other results. First proven in [30], the Hales–Jewett theorem is a statement about coloring points of the grid $[t]^n$ (where $[t]$, from here on, will be shorthand for the set $\{1, 2, \ldots, t\}$). Imprecisely speaking, it states that an r-coloring of this grid will contain a line t collinear points of the same color, provided that the dimension n is sufficiently large: at least the Hales–Jewett number $\mathrm{HJ}(t, r)$. The exact statement restricts the set of lines we consider. It is described in terms of the parameter sets theorem in Chapter 2, and stated independently in Chapter 3.

Somewhat facetiously, Ramsey theory has also been called "the branch of combinatorics consisting of iterated use of the pigeonhole principle". This observation actually has an important consequence: very often, pigeonhole-type arguments lead to extremely fast-growing upper bounds, which differ greatly from the best lower bounds known.

For example, the lower bounds to the dimension n required for the Hales–Jewett theorem to hold are merely exponential in r and t. On the other hand, the upper bounds require defining new quickly-growing functions (such as those defined in Section 2.5) to state.

Chapters 2 and 3 of this thesis focus on problems in Ramsey theory. Specifically, they provide partial progress on one fundamental question: can we prove upper bounds on quantities like the Hales–Jewett numbers *without* heavy use of the pigeonhole principle? This produces significant improvements in all cases where it has succeeded. For example, in Chapter 2, a lemma proved via SAT solver finds an exact answer of 6 to a problem for which the classical upper bound is on the order of $2 \uparrow\uparrow 18$ (in the notation of Section 2.5). In Chapter 3, the main result improves an upper bound on $\mathrm{HJ}(4, 2)$ between $2 \uparrow\uparrow 7$ and $2 \uparrow\uparrow 8$ to one which is "merely" an 11-digit number.

1.2 Random graphs

Whereas upper bounds for problems in Ramsey theory must come from universal arguments—ones that find order in every possible example of a structure—lower bounds have a very different nature. To prove a lower bound for the Ramsey number $R(k)$, for example, it's natural to try to provide an edge-coloring of K_n, for some large n, which does not contain a monochromatic complete subgraph of size k.

Finding an explicit rule for such an edge-coloring is challenging, more or less because any rule which is simple enough for us to verify its correctness is too structured to avoid large monochromatic complete subgraphs. Instead, the best known lower bound for $R(k)$ comes from coloring the edges of K_n randomly: an idea originally due to Paul Erdős [17].

Random structures, including multiple models for choosing random graphs, have now become both a mainstay source of counterexamples in combinatorics and an object of study in their own right. One particularly fundamental model is the Erdős–Rényi graph $\mathcal{G}_{n,p}$, in which each edge between n vertices in total is independently present with probability p. Often, to get a handle on the behavior of a graph parameter such as the diameter or the chromatic number in a "typical' case, we study its distribution for a randomly graph in this model or in others. In many situations, one can prove that interesting properties hold with probability $1 - o(1)$, in which case

the property is said to hold *asymptotically almost surely*, or a.a.s. for short.

For example, in Chapter 4, we consider a graph invariant called the game chromatic number. We would like to understand when this invariant differs from the ordinary chromatic number of the graph, and by how much. In Chapter 4, we show that these two invariants a.a.s. differ for the random 3-regular graph $\mathcal{G}_{n,3}$ (chosen uniformly at random from all n-vertex graphs in which all vertices have degree 3).

Similarly, when faced with a problem in extremal combinatorics, we can often shed light on it by considering its random analog. Chapter 5 of this thesis deals with such a situation. Given a total ordering of the edges of the complete graph K_n, which can be specified by a bijection $\binom{f:E(K_n)\to\{1,2,\ldots,\{n}{2\}\}}{}$, an increasing path is one whose edges $(e_1, e_2, \ldots, e_k)$ satisfy $f(e_1) < f(e_2) < \cdots < f(e_k)$.

It has recently been shown by Milans [42] that for any such f, we can find an increasing path of length at least $O(n/\log n)^{2/3}$, improving a previous bound of $O(\sqrt{n})$ due to Graham and Kleitman [27]. However, edge-orderings with no long increasing path seem difficult to find: several constructions with a linear upper bound are known, but none that improve on $(\frac{1}{2}+o(1))n$. In Chapter 5, we give a partial explanation of this gap. In fact, if the bijection f is chosen uniformly at random, then the longest increasing path is very long: we show that not only is there a path of the maximum length $n-1$ with probability at least $\frac{1}{e}$, but also that even a path chosen by the greedy algorithm attains linear length a.a.s. Thus, edge-orderings with no long increasing paths must be rare, if they exist at all.

However, imagining that all cases of a problem behave like the typical, randomly chosen case can be misleading. An instance of this is explored in Chapter 6, which considers distance-uniform graphs: graphs with the surprising property, that, for some constant d, every vertex is at distance exactly d from all but a small fraction of the other vertices. Though it is not initially obvious that nontrivial examples of distance-uniform graphs exist at all, it is a consequence of the work of Bollobás on the diameter of random graphs [10] that, for many choices of p, the

Erdős–Rényi random graph $\mathcal{G}_{n,p}$ is distance-uniform a.a.s.

A distance-uniform graph obtained in this way has a diameter logarithmic in n (and therefore the constant d is also at most logarithmic in n). It is reasonable to conjecture that all distance-uniform graphs resemble $\mathcal{G}_{n,p}$, and also satisfy $d = O(\log n)$. However, we show in Chapter 6 that a class of distance-uniform graphs exists with much larger diameter, satisfying $d = 2^{\Omega(\frac{\log n}{\log \log n})}$.

1.3 Acknowledgment

The chapters of this thesis are based on research papers which are in various stages of publication. Specifically, Chapter 2 corresponds to [37] (co-authored with John Mackey and Mitchell Lee), Chapter 3 to [36], Chapter 4 to a section of [22] (co-authored with Alan Frieze and Simcha Haber), Chapter 5 to [38] (co-authored with Po-Shen Loh), and Chapter 6 to [39] (also co-authored with Po-Shen Loh).

Chapter 2

Improved Upper and Lower Bounds on a Geometric Ramsey Problem

2.1 Background

2.1.1 The parameter sets theorem

Let A be a finite set with at least 2 elements, and A^n the set of n-tuples of elements of A. Fix a finite group G which acts on A. We define a k-parameter subset of A^n to be the image of an injection $f : A^k \to A^n$ such that for all $1 \leq i \leq n$, f_i has one of two forms:

1. either there is some $a \in A$ such that for all $(x_1, \ldots, x_k) \in A^k$, $f_i(x_1, \ldots, x_k) = a$,
2. or there are some $1 \leq j \leq k$ and some $\sigma \in G$ such that for all $(x_1, \ldots, x_k) \in A^k$, $f_i(x_1, \ldots, x_k) = \sigma(x_j)$.

The requirement that f is an injection is equivalent to asking that for all $1 \leq j \leq k$, there exist i, σ such that $f_i(x_1, \ldots, x_k) = \sigma(x_j)$. Notably, if f defines a t-parameter subset of A^n, and g defines a k-parameter subset of A^t, then $f \circ g$ defines a k-parameter subset of A^n.

The n-parameter sets were introduced by Graham and Rothschild [28], who proved the following result:

Theorem 2.1.1 (Graham–Rothschild Parameter Sets Theorem)**.** *Pick an alphabet A, a group G acting on A, and integers $0 \leq k \leq t$ and $r \geq 2$. Then there exists a N such that for all $n \geq N$, if the k-parameter subsets of A^n are colored one of r colors, then some t-parameter subset of A^n can be found, all of whose k-parameter subsets receive the same color.*

For the remainder of this chapter, we will only consider 2-colorings ($r = 2$) and, when necessary, we will call these two colors "red" and "blue".

There are several special cases of Theorem 2.1.1 which are of interest.

Graham's number

Take $A = \{\pm 1\}$, and let G be the group of both permutations of A: $\{x \mapsto x, x \mapsto -x\}$. Then any two points in $\{\pm 1\}^n$ form a 1-parameter set. More generally, a d-parameter set consists of 2^d points that lie on a d-dimensional affine subspace of $\mathbb{R}^n$ (if we include $\{\pm 1\}^n \subset \mathbb{R}^n$ in the natural way). We will also call this a d-dimensional subcube of $\{\pm 1\}^n$.

An edge-coloring of $\{\pm 1\}^n$ is a 2-coloring of the edges of the complete graph on the 2^n points of $\{\pm 1\}^n$: a coloring of the 1-parameter subsets of $\{\pm 1\}^n$. Let $\mathrm{Graham}(d)$ be the smallest dimension n such that every edge-coloring of the n-dimensional cube contains a monochromatic d-dimensional subcube. Then Theorem 2.1.1 implies that $\mathrm{Graham}(d)$ exists and is finite for all d.

In particular, $\mathrm{Graham}(2)$ is the smallest dimension n such that every edge-coloring of $\{\pm 1\}^n$ contains a monochromatic planar K_4: a set of 4 coplanar points in $\{\pm 1\}^n$ such that all 6 edges between them are the same color. An incredibly large upper bound on $\mathrm{Graham}(2)$ was popularized as "Graham's number" by Martin Gardner [24], and appeared in the 1980 Guinness Book of World Records as the highest number ever used in a mathematical proof.

The Hales–Jewett number

Take $A = [t] := \{1, 2, \ldots, t\}$, let $G = \{e\}$, and color the 0-parameter sets of $[t]^n$, which is equivalent to coloring elements of $[t]^n$. A 1-parameter set in $[t]^n$ is called a combinatorial line, and a d-parameter set is called a d-dimensional combinatorial space.

The Hales–Jewett number $\mathrm{HJ}(t, r)$ be the least dimension n such that every r-coloring of $[t]^n$ contains a monochromatic combinatorial line. More generally, $\mathrm{HJ}(t, r, d)$ is the least dimension n such that every r-coloring of $[t]^n$ contains a monochromatic d-dimensional combinatorial space.

The Hales–Jewett number is probably the most well-studied of the three problems. In [46], Shelah proved a primitive recursive upper bound on $\mathrm{HJ}(t, r, d)$.

The tic-tac-toe number

Again consider coloring the elements of $[t]^n$, but this time allow a wider variety of d-parameter subsets: let $G = \{e, \pi\}$, which acts on $[t]$ by $e(x) = x$ and $\pi(x) = t + 1 - x$. A 1-parameter subset using this G is called a tic-tac-toe line, and a d-parameter subset a d-dimensional tic-tac-toe space. We define the tic-tac-toe numbers $\mathrm{TTT}(t, r)$ and $\mathrm{TTT}(t, r, d)$ analogously to the Hales–Jewett numbers.

Since the same set of points is colored, but more subsets are acceptable, it is clear that $\mathrm{TTT}(t, r, d) \leq \mathrm{HJ}(t, r, d)$. Furthermore, it is easily shown that for all t, r, and d, $\mathrm{HJ}(\lceil t/2 \rceil, r, d) \leq \mathrm{TTT}(t, r, d)$, so the overall behavior of the tic-tac-toe numbers and Hales–Jewett numbers are similar. However, for small values of t (and we will only consider the case $t = 4$) the behavior of these two bounds is potentially quite different, and it is therefore worthwhile to state our results in terms of the tic-tac-toe number instead.

2.1.2 Previous results on $\mathrm{Graham}(d)$

In [28], Graham and Rothschild observed that $\mathrm{Graham}(2) \geq 6$, and conjectured that $\mathrm{Graham}(2) < 10$. This conjecture has been proven false, but not by much: the lower bound was later improved to 11 by Exoo [19] and then to 13 by by Barkley [3].

The upper bound is a more complicated story. Most sources list the bound from [28] as the best upper bound known. However, by combining several known results in a straightforward way, we see that there is a primitive recursive upper bound on $\mathrm{Graham}(d)$.

First, by using Theorem 5.1 in [12], we can reduce $\mathrm{Graham}(d)$ to an application of Theorem 2.1.1 in which $G = \{e\}$, at the cost of increasing k and t: we now look for $2d$-parameter sets with monochromatic 2-parameter subsets. Theorem 2.2 in [46] now applies, and gives the primitive recursive bound on $\mathrm{Graham}(d)$ which follows.

Let m be the diagonal Ramsey number $R(2d)$ (e.g., for $d = 2$, $m = 18$). Then $\mathrm{Graham}(d) \leq k_m$, where $k_0, \ldots, k_m$ are defined recursively by

$$k_0 = 0, \qquad k_{i+1} = k_i + \mathrm{HJ}(k_i + 2, 2^{(2d+2)^{k_i+m-i-1}}).$$

2.1.3 New results

Our primary results are the following improved bounds on $\mathrm{Graham}(d)$ and especially $\mathrm{Graham}(2)$:

Theorem 2.1.2. $\mathrm{TTT}(4, 2, d) + 1 \leq \mathrm{Graham}(d + 1)$.

Theorem 2.1.3. $\mathrm{Graham}(2) \leq \mathrm{TTT}(4, 2, 6) + 1$.

To convert this to a numerical upper bound, we resort to Knuth's up-arrow notation, defined in Section 2.5. In particular, if we bound $\mathrm{TTT}(4, 2, 6)$ by $\mathrm{HJ}(4, 2, 6)$, then by Lemma 2.5.2, which analyzes the growth rate of Shelah's bound on $\mathrm{HJ}(t, k, d)$, we have $\mathrm{Graham}(2) \leq 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 9)) < 2 \uparrow\uparrow\uparrow 6$. This is a significant improvement on all previously known bounds.

The proof of Theorem 2.1.3, however, is not completely satisfactory. For Lemma 2.2.2, which states that under some strong simplifying assumptions a monochromatic K_4 exists in dimension

$n = 6$ (a tight bound), we only have a computer-aided proof. However, a weaker version of this lemma can be easily shown, yielding:

Theorem 2.1.4. *There exists a positive integer d, e.g. $d = 2 \uparrow\uparrow 18$, such that* $\mathrm{Graham}(2) \leq \mathrm{TTT}(4, 2, d) + 1$.

This is still easily strong enough to yield $2 \uparrow\uparrow\uparrow 6$ as an upper bound.

We also consider a simpler problem: given an edge-coloring $\{\pm 1\}^n$, to find a monochromatic planar rectangle. The points defining a rectangle are still a 2-parameter set; however, rather than requiring that all 6 edges between them are monochromatic, we only consider the 4 edges between "adjacent" points.

This simplified problem has a much smaller upper bound:

Theorem 2.1.5. *An edge-coloring of $\{\pm 1\}^{78}$ necessarily contains a monochromatic planar square whose sides have Hamming length 2.*

2.2 Bounds on $\mathrm{Graham}(d)$

2.2.1 Setup

Let Q be the cube $\{\pm 1\}^{n+1}$ with coordinates numbered $0, \ldots, n$. Let $Q^- = \{x \in Q : x_0 = -1\}$ and $Q^+ = \{x \in Q : x_0 = +1\}$.

Let $\phi : \{\pm 1\}^2 \to [4]$ be given by $\phi(-1, -1) = 1$, $\phi(-1, +1) = 2$, $\phi(+1, -1) = 3$, $\phi(+1, +1) = 4$. We use ϕ to define a bijection Φ from the edges $Q^- \times Q^+$ to $[4]^n$:

$$\Phi(x, y) = (\phi(x_1, y_1), \ldots, \phi(x_n, y_n)).$$

For a $(d + 1)$-dimensional subcube of Q, there are three possibilities: either it is contained entirely in Q^-, or entirely in Q^+, or half of its vertices are in Q^- and half are in Q^+. In the third case, we call the subgraph formed by the edges of the subcube going from Q^- to Q^+ a *d-dimensional hyperbowtie* (the name is formed by analogy with the case $d = 1$, in which case the four edges make a bowtie shape).

Lemma 2.2.1. *If $S \subseteq Q^- \times Q^+$ is a set of edges of Q, then S is a d-dimensional hyperbowtie if and only if $\Phi(S)$ is a d-dimensional tic-tac-toe subspace of $[4]^n$.*

Proof. Let $f : \{\pm 1\}^{d+1} \to \{\pm 1\}^{n+1}$ be a function whose image is a $(d+1)$-dimensional subcube half contained in Q^- and half in Q^+. Then f_0 cannot be constant; because so far, all coordinates of $\{\pm 1\}^{d+1}$ are symmetric, we may assume that $f_0(x_0, \ldots, x_d) = x_0$. Define $g : [4]^d \to [4]^n$ as follows: if $(z_1, \ldots, z_d) \in [4]^d$, let $(x_i, y_i) = \phi^{-1}(z_i)$ for $1 \leq i \leq d$, and let $g(z_1, \ldots, z_d) = \Phi(f(-1, x_1, \ldots, x_d), f(+1, y_1, \ldots, y_d))$. As z varies, the edge from $(-1, x)$ to $(+1, y)$ varies over all edges in S, the d-dimensional hyperbowtie corresponding to the image of f. Therefore the image of g is $\Phi(S)$.

For each coordinate $1 \leq i \leq d$, we consider all possibilities for f_i, and check what form g_i then has:

- If $f_i(x) = a$ for a constant $a \in \{\pm 1\}$, then $f_i(-1, x_1, \ldots, x_d) = a$ and $f_i(+1, y_1, \ldots, y_d) = a$, and so $g_i(z) = \phi(a, a)$ which is a constant: either $\phi(-1, -1) = 1$ or $\phi(1, 1) = 4$.
- If $f_i(x) = ax_0$ for $a \in \{\pm 1\}$, then $f_i(-1, x_1, \ldots, x_d) = -a$ and $f_i(+1, y_1, \ldots, y_d) = a$, so $g_i(z) = \phi(-a, a)$ which is a constant: either $\phi(-1, 1) = 2$ or $\phi(1, -1) = 3$.
- If $f_i(x) = x_j$, for $j \geq 1$, then $g_i(z) = \phi(x_j, y_j) = z_j$.
- If $f_i(x) = -x_j$, for $j \geq 1$, then $g_i(z) = \phi(-x_j, -y_j)$. It can be checked that $\phi(-x, -y) = 5 - \phi(x, y)$, and so $g_i(z) = 5 - \phi(x_j, y_j) = 5 - z_j$.

Therefore g has the correct form for the image of g to be a d-dimensional tic-tac-toe subspace. Moreover, every possibility for g_i can be obtained by some choice of f_i, and so every d-dimensional tic-tac-toe subspace can be obtained in this way: as the image under Φ of a d-dimensional hyperbowtie. □

2.2.2 The lower bound

Proof of Theorem 2.1.2. Let $n = \mathrm{Graham}(d+1) - 1$ and let Q be the cube $\{\pm 1\}^{n+1}$.

Pick an arbitrary 2-coloring of $[4]^n$. The map Φ is a bijection between $[4]^n$ and those edges of Q which change the first coordinate, so we use this bijection to assign those edges a color. To color the remaining edges, we assign the edge from $(x_0, x_1, \ldots, x_n)$ to $(y_0, y_1, \ldots, y_n)$, where $x_0 = y_0$, the same color as the edge from $(-1, x_1, \ldots, x_n)$ to $(+1, y_1, \ldots, y_n)$.

Because $n + 1 = \mathrm{Graham}(d + 1)$, a $(d + 1)$-dimensional subcube of Q is monochromatic. Suppose this subcube is half contained in Q^+ and half in Q^-. Then the edges of the subcube contained in $Q^+ \times Q^-$ form a monochromatic d-dimensional hyperbowtie, and by Lemma 2.2.1, Φ maps it to a monochromatic d-dimensional tic-tac-toe space in $[4]^n$.

Now consider the other possibility: the subcube is entirely contained in Q^+ or Q^-. Let i be the first coordinate which is not constant on this subcube. We restrict our attention to the 4^d edges in the subcube which change coordinate i: edges from $(x_0, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n)$ to $(y_0, \ldots, y_{i-1}, +1, y_{i+1}, \ldots, y_n)$, where $x_0 = y_0$, $x_1 = y_1$, ..., $x_{i-1} = y_{i-1}$. Alter each edge by replacing x_0 with -1 and y_0 with $+1$. By construction, the new edge has the same color, so the edges we obtain will also be monochromatic. But now the edges we get form a d-dimensional hyperbowtie, and we use Lemma 2.2.1 again to obtain a monochromatic d-dimensional tic-tac-toe space in $[4]^n$.

Therefore we have shown that $[4]^n$ always contains a monochromatic d-dimensional tic-tac-toe space, and so $n \geq \mathrm{TTT}(4, 2, d)$. □

2.2.3 A special case

To prove the upper bound on $\mathrm{Graham}(2)$, we will first state a lemma about a special case of this problem.

Lemma 2.2.2. *Suppose the cube $\{\pm 1\}^6$ is 2-colored so that all parallel edges receive the same color. Then the cube contains a monochromatic planar K_4.*

Unfortunately, we do not have a proof of this lemma. However, with the parallel edge assumption, we can write a Boolean formula with 364 variables which is satisfied if and only if a

coloring with no monochromatic planar K_4 exists. The Boolean satisfiability problem is tractable on instances of this size, and a computerized search showed that no solutions exist.

It is possible, however, to prove a weaker version of the lemma.

Lemma 2.2.3. *Let $n = 2 \uparrow\uparrow 18$, and suppose the cube $\{\pm 1\}^n$ is 2-colored so that all parallel edges receive the same color. Then the cube contains a monochromatic planar K_4.*

Proof. An equivalence class of parallel edges in the cube $\{\pm 1\}^n$ can be described by a *direction* $a \in \{-1, 0, 1\}^n$, corresponding to all possible edges of the form $(x - a, x + a)$ for some x; a and $-a$ represent the same direction.

We define addition and subtraction of directions componentwise. In order for $a + b$ and $a - b$ to also be directions, we require that a and b have disjoint support: for all $1 \leq i \leq n$, at most one of a_i and b_i are nonzero. (Otherwise, we risk that $a_i \pm b_i \notin \{-1, 0, 1\}$.)

Suppose that, for two directions a and b with disjoint support, the four directions a, b, $a + b$, and $a - b$ are all the same color. Choose $x \in \{-1, 0, 1\}^n$ such that for all $1 \leq i \leq n$, exactly one of a_i, b_i, and x_i is nonzero, and let $P = x - a - b$, $Q = x - a + b$, $R = x + a - b$, $S = x + a + b$. Then $PQRS$ is a monochromatic planar K_4: edges PR and QS have direction a, PQ and RS have direction b, PS has direction $a + b$, and QR has direction $a - b$. Therefore it suffices to find a monochromatic set of directions $a, b, a + b, a - b$.

We appeal to the finite unions theorem also proven in [28] by Graham and Rothschild:

Theorem 2.2.1 (Finite Unions Theorem (Corollary 3 in [28]))**.** *Given integers ℓ and r, there exists an integer $N = N(\ell, r)$ such that if the subsets of $[N]$ are r-colored, there exist ℓ disjoint nonempty subsets $S_1, \ldots, S_\ell$ of $[N]$ such that all $2^\ell - 1$ unions $\bigcup_{j \in J} S_j$ with $J \subseteq [\ell]$, $J \neq \emptyset$, have one color.*

We take $\ell = 4$ and $r = 2$ in this result, and will later verify that $N(4, 2)$ is no greater than $n = 2 \uparrow\uparrow 18$. By the usual correspondence between subsets of $[n]$ and elements of $\{0, 1\}^n$, the finite unions theorem lets us choose four directions $a, b, c, d \in \{0, 1\}^n$ with the following properties:

- a, b, c, and d have disjoint support.
- The 15 directions $a, \ldots, d, a+b, a+c, \ldots, c+d, a+b+c, \ldots, b+c+d, a+b+c+d$ are the same color (say, red).

If $a-b$ is red, then $\{a, b, a+b, a-b\}$ determine a red planar K_4, so assume $a-b$ is blue. Similarly, if $c-d$ is red, then $\{c, d, c+d, c-d\}$ determine a red planar K_4, so assume $c-d$ is blue.

If $a-b+c-d$ is red, then $\{a+c, b+d, a+b+c+d, a-b+c-d\}$ determine a red planar K_4, so assume $a-b+c-d$ is blue. Similarly, if $a-b-c+d$ is red, then $\{a+d, b+c, a+b+c+d, a-b-c+d\}$ determine a red planar K_4, so assume $a-b-c+d$ is blue.

But now $\{a-b, c-d, a-b+c-d, a-b-c+d\}$ determine a blue planar K_4, and we have what we wanted.

It remains to check that the dimension $N(4,2)$ in the finite unions theorem is not too large. We rely on the second proof outlined in [29], p. 83.

Let $n(k)$ be the dimension needed to obtain k directions $a^1, \ldots, a^k$ with the following properties: for each nonempty $I \subseteq [k]$, the color $\sum_{i \in I} a^i$ is determined only by $\max\{I\}$. As a base case, $n(1) = 1$, since then any direction suffices.

To go from $n(k)$ to $n(k+1)$, let $n = \mathrm{HJ}(2, 2, n(k))$ and choose a monochromatic $n(k)$-dimensional combinatorial subspace of $\{0,1\}^n$. This can be described by directions $b^0, \ldots, b^{n(k)} \in \{0,1\}^n$ (with disjoint support) such that for all $I \subseteq [n(k)]$, $b^0 + \sum_{i \in I} b^I$ is the same color (say, red).

The set of all possible sums of $b^1, \ldots, b^{n(k)}$ is isomorphic to $\{0,1\}^{n(k)}$, so we can find k directions $a^1, \ldots, a^k$, which are sums of some of the b^i and have the property we want. Furthermore, let $a^{k+1} = b^0$. Then for all nonempty $I \subseteq [k+1]$, the sum $\sum_{i \in I} a^i$ is determined by $\max\{I\}$: this is true by the inductive hypothesis if $\max\{I\} \leq k$, and if $\max\{I\} = k+1$, the sum lies in the combinatorial subspace we found, and is red. Therefore $n(k+1) \leq n$.

By the bound in [46], $\mathrm{HJ}(2,2,d) \leq 2^{2^{2d}}$, so $n(k+1) \leq 2^{2^{2n(k)}} \leq 2^{2^{2^{n(k)}}}$. Since $n(1) = 1 =$

$2 \uparrow\uparrow 0$, $n(7) \leq 2 \uparrow\uparrow 18$.

Finally, if we take $n = n(7)$, we can find seven directions $a^1, \ldots, a^7$ as above. Choose four of these that are the same color; then because $\sum_{i \in I} a^i$ has the color of $a^{\max\{I\}}$, all their sums will share that color, and we can use them above to obtain a monochromatic planar K_4. □

2.2.4 The upper bound

Proof of Theorem 2.1.3. Let $n = \mathrm{TTT}(4, 2, d)$, where d is either 6 or $2 \uparrow\uparrow 18$, depending on whether Lemma 2.2.2 or Lemma 2.2.3 is used. Let Q be the cube $\{\pm 1\}^{n+1}$. Given a 2-coloring of the edges of Q, we consider just the edges from Q^- to Q^+, and apply Φ to them to get a coloring of $[4]^n$. This coloring must contain a monochromatic d-dimensional tic-tac-toe space; by Lemma 2.2.1, its preimage in Q is a d-dimensional monochromatic hyperbowtie.

From now on, we will look only at the $(d+1)$-dimensional subcube containing this hyperbowtie. What we know about this cube is that all edges which change the first coordinate (which we will call the "middle" of the cube) are colored the same color, which may as well be red. The remaining edges are contained in one of two d-dimensional cubes: the "top" and "bottom".

We reduce the problem of finding a monochromatic planar K_4 in this subcube to Lemma 2.2.2. We color the edges of $\{\pm 1\}^d$ as follows:

(1) An equivalence class of parallel edges is colored blue, if the corresponding edges on the top are all colored blue.

(2) An equivalence class of parallel edges is colored red, if the corresponding edges on the bottom are all colored blue, and (1) does not hold.

(3) If neither of these occurs, then there is a pair of parallel edges, one on the top and one on the bottom, which are colored red. Together with four edges in the middle, which are also red, they form a monochromatic planar K_4.

We are done if (3) holds for some equivalence class of parallel edges. Otherwise, by Lemma 2.2.2 or Lemma 2.2.3, the coloring we obtain contains a monochromatic planar K_4. If it is blue, then

the corresponding K_4 on the top is monochromatic blue. If it is red, then the corresponding K_4 on the bottom is monochromatic blue. □

2.3 Monochromatic planar squares

For a vertex v of the n-dimensional cube and $1 \leq i \leq n$, let $v \oplus i$ denote the vertex obtained by flipping the i-th coordinate of v. Whenever we refer to length or distance between two vertices, it will be Hamming distance: the number of coordinates in which the two coordinates differ.

Lemma 2.3.1. *For $n \geq 4$, in any edge-coloring of the n-dimensional cube, at least $\frac{1}{2} - \frac{1}{2\lfloor \frac{n}{2} \rfloor - 2}$ of all right angles formed by edges of length 2 are monochromatic.*

Proof. Choose a vertex v of the n-dimensional cube, and a permutation π of $\{1, \ldots, n\}$.

Let $k = \lfloor \frac{n}{2} \rfloor$; for $1 \leq i \leq k$, let $w_i = v \oplus \pi(2i-1) \oplus \pi(2i)$. Then the edges $(v, w_1), \ldots, (v, w_k)$ are mutually perpendicular and have length 2. There are $\binom{k}{2}$ pairs of edges in this set; the number of monochromatic pairs is minimized if $\frac{k}{2}$ of the edges are red, and $\frac{k}{2}$ are blue, for a total of $2\binom{k/2}{2}$ monochromatic pairs.

When $k \geq 2$, the ratio of these is $\frac{1}{2} - \frac{1}{2(k-1)}$, which is the proportion of monochromatic pairs among these edges. By averaging over all choices of v and π, we obtain the same proportion over the entire cube. □

Lemma 2.3.2. *For $n \geq 4$, in any edge-coloring of the n-dimensional cube, at least $\frac{1}{2} - \frac{1}{2(n-3)}$ of all pairs of parallel edges of length 2, which are also at distance 2 from each other, are monochromatic.*

Proof. Choose a vertex v of the n-dimensional cube, and a permutation π of $\{1, \ldots, n\}$.

Let $k = n - 2$; for $1 \leq i \leq k$, let $v_i = v \oplus \pi(i)$ and $w_i = v_i \oplus \pi(n-1) \oplus \pi(n)$. Then the edges $(v_1, w_1), \ldots, (v_k, w_k)$ are all parallel, have length 2, and are at distance 2 from each other. There are $\binom{k}{2}$ pairs of edges in this set; the number of monochromatic pairs is minimized if $\frac{k}{2}$ of the edges are red, and $\frac{k}{2}$ are blue, for a total of $2\binom{k/2}{2}$ monochromatic pairs.

When $k \geq 2$, the ratio of these is $\frac{1}{2} - \frac{1}{2(k-1)}$, which is the proportion of monochromatic pairs among these edges. By averaging over all choices of v and π, we obtain the same proportion over the entire cube. □

Lemma 2.3.3. *For $n \geq 5$, in any edge-coloring of the n-dimensional cube, at most $\frac{14}{15}$ of all 2×2 squares have an odd number of red edges.*

Proof. Choose a vertex v of the n-dimensional cube, and a permutation π of $\{1, \ldots, n\}$.

Let $v_1, \ldots, v_{10}$ be $v \oplus \pi(i) \oplus \pi(j)$, for $1 \leq i < j \leq 5$. Using only edges of length 2 between these vertices, 15 squares can be formed, which together use each edge exactly twice.

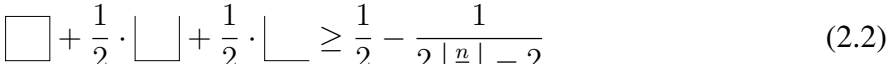
Assume for the sake of contradiction that these edges are colored so that all 15 squares have an odd number of red edges. Represent the two colors, red and blue, by 1 and 0, and let the sum of a square be the sum of the colors of its edges. The sums of all 15 squares must be odd, so adding up all 15 sums, we also get an odd number. However, each edge is used twice and therefore contributes an even number to this total; a contradiction.

Therefore at most $\frac{14}{15}$ of these squares have an odd number of red edges. By averaging over all choices of v and π, we obtain the same proportion over the entire cube. □

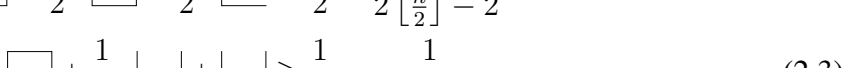
Proof of Theorem 2.1.5. There are four types of colorings of 2×2 squares, up to symmetry and interchanging the two colors:

For $n \geq 5$, fix an edge-coloring of the n-dimensional cube. We will use the four symbols above to denote the proportions of 2×2 squares of each type. By applying the lemmas, we can write the following system of inequalities:

$$= 1 \tag{2.1}$$



$$\geq \frac{1}{2} - \frac{1}{2\left\lfloor \frac{n}{2} \right\rfloor - 2} \tag{2.2}$$



$$\geq \frac{1}{2} - \frac{1}{2(n-3)} \tag{2.3}$$

$$\square \leq \frac{14}{15}. \tag{2.4}$$

Solving for $\square$ by taking $2 \cdot (2) + (3) - (1) - \frac{1}{2} \cdot (4)$, we get

$$2 \cdot \square \geq \frac{1}{30} - \frac{1}{\lfloor \frac{n}{2} \rfloor - 1} - \frac{1}{2(n-3)}.$$

For $n \geq 78$, the left-hand side is positive, and therefore a monochromatic 2×2 square exists. □

2.4 Computer proof of Lemma 2.2.2

The following Mathematica code, which was used to verify Lemma 2.2.2, is also available at `http://www.math.cmu.edu/~mlavrov/other/Graham.nb`. The online version includes additional code which, for $n \leq 5$, will draw an edge coloring containing no monochromatic planar K_4.

2.4.1 Subroutines

Here we define several short subroutines that will be used in the next section to construct the 6-dimensional cube.

1. `CanonicalForm`: A class of parallel edges will be represented by an n-element list of elements of $\{-1, 0, 1\}$ indicating the edges' direction in each coordinate. Each list indicates the same class as its negation; `CanonicalForm` picks a canonical sign for each such list.

   ```
   CanonicalForm[edge_] := Last[Sort[{edge, -edge}]]
   ```

2. `DisjointSupportQ`: Returns `True` if two directions have disjoint support (in each coordinate, at most one is nonzero); edges in these directions are perpendicular.

   ```
   DisjointSupportQ[{edge1_, edge2_}] :=
   Module[{support1, support2},
   ```

```
    support1 = Flatten[Position[edge1, 1 | -1]];
       (* which elements of edge1 are nonzero *)
    support2 = Flatten[Position[edge2, 1 | -1]];
       (* which elements of edge2 are nonzero *)
    Return[Intersection[support1, support2] == {}];
];
```

3. `MakeK4FromRectangle`: Given two directions (which are assumed to have disjoint support, as above), completes them to a set of four directions spanning a planar K_4.

```
MakeK4FromRectangle[{a_, b_}] :=
   {a, b, CanonicalForm[a+b], CanonicalForm[a-b]};
```

4. `NotMonochromatic`: Given four directions, defines a Boolean formula on the corresponding variables which is satisfied if and only if the variables are not all equal: in the coloring interpretation, if and only if the corresponding planar K_4 is not monochromatic.

```
NotMonochromatic[{a_, b_, c_, d_}] :=
   (x[a]~Or~x[b]~Or~x[c]~Or~x[d])
      ~And~
   Not[x[a]~And~x[b]~And~x[c]~And~x[d]]
```

2.4.2 Structure

Next, we initialize variables that store the structure of the cube: the parallel edge classes and the planar K_4's they form.

1. `n`: the dimension of the cube.

```
n=6;
```

2. `edges`: all possible parallel edge classes, defined by whether they increase (+1), decrease (1) or do not change (0) in each coordinate. We exclude the trivial direction which does

not change any coordinate, and put each direction in canonical form, removing duplicates.

```
edges = Tuples[{-1, 0, 1}, n];
edges = Cases[edges, Except[{0 ..}]];
edges = DeleteDuplicates[Map[CanonicalForm, edges]];
```

3. `rectangles`: Two directions form the sides of a rectangle if they change disjoint sets of coordinates.

   ```
   rectangles = Select[Subsets[edges, {2}], DisjointSupportQ];
   ```

4. `k4s`: All planar K_4's are formed by the sides and diagonals of a rectangle.

   ```
   k4s = Map[MakeK4FromRectangle, rectangles];
   ```

2.4.3 Satisfiability

Finally, we express the coloring problem as a SAT instance, which allows us to use Mathematica's built-in commands to check that no solution exists.

For each direction (stored in `edges`), we have a binary variable, `True` or `False` depending on the color of the edge. The constraints are simply that no planar K_4 (stored in `k4s`) is monochromatic.

```
variables = Map[x, edges];
constraints = Map[NotMonochromatic, k4s];
formula = Apply[And, constraints];
```

The `SatisfiabiliyInstances` command attempts to assign values to the variables (colors to the edges) satisfying the constraints, returning $\{\}$ if this is impossible. Expect this command to take several minutes to execute.

```
output = SatisfiabilityInstances[formula, variables]
```

2.5 Rate of growth of Shelah's Hales–Jewett bounds

We will make use of Knuth's up-arrow notation. Let $a \uparrow b := a^b$; then define

$$a \uparrow\uparrow b = \underbrace{a \uparrow (a \uparrow (a \uparrow \cdots \uparrow a))}_{b\ a's}.$$

Finally, define

$$a \uparrow\uparrow\uparrow b = \underbrace{a \uparrow\uparrow (a \uparrow\uparrow (a \uparrow\uparrow \cdots \uparrow\uparrow a))}_{b\ a's}.$$

We will use the following rules to rewrite expressions written in this notation:

- An expression of the form $a \uparrow a \uparrow \cdots \uparrow a$ with arbitrarily-inserted parentheses is always maximized when the parentheses are placed as in the definition of $a \uparrow\uparrow b$.
- Therefore $(a \uparrow\uparrow b) \uparrow\uparrow c \leq a \uparrow\uparrow (b \cdot c)$, by expanding and rearranging the parentheses.
- Since $(2 \uparrow\uparrow k)^2 \leq 2 \uparrow\uparrow (k+1)$, it is also true that $a \cdot (2 \uparrow\uparrow k) < 2 \uparrow\uparrow (k+1)$ for any $a < 2 \uparrow\uparrow k$.
- Finally, $a + (2 \uparrow\uparrow k) < 2 \cdot (2 \uparrow\uparrow k) \leq 2 \uparrow\uparrow (k+1)$, for any $a < 2 \uparrow\uparrow k$.

Lemma 2.5.1. *If* $\mathrm{HJ}(t-1, 2, d) \leq 2 \uparrow\uparrow m$*, then* $\mathrm{HJ}(t, 2, d) \leq 2 \uparrow\uparrow (2 \uparrow\uparrow (m+3))$.

Proof. From [46], we know the following: suppose $\mathrm{HJ}(t-1, r, d) = n$. Then $\mathrm{HJ}(t, r, d) \leq nf(n, r^{t^n})$, where $f(\ell, k)$ is defined recursively by $f(1, k) = k+1$ and $f(\ell+1, k) = k^{f(\ell,k)^{2\ell}} + 1$.

We begin by bounding $f(\ell, k)$ in up-arrow notation. Whenever we will need to find $f(\ell, k)$, we will have $k > 2\ell$. Thus, we can write $f(\ell, k) < k^{f(\ell-1,k)^k}$; iterating this bound, and rearranging the parentheses, we get

$$f(\ell, k) < \underbrace{k^{k^{k^{\cdot^{\cdot^{\cdot^{k}}}}}}}_{2\ell} = k \uparrow\uparrow 2\ell.$$

Suppose $\mathrm{HJ}(t-1, 2, d) \leq 2 \uparrow\uparrow m$. It is easy to check that $\mathrm{HJ}(t-1, 2, d) \geq t-1$, and therefore $t \leq 1 + 2 \uparrow\uparrow m \leq 2 \uparrow\uparrow (m+1)$. So we have $2^{t^n} \leq 2 \uparrow\uparrow (2m+2)$. Therefore

$$\begin{aligned} nf(n, 2^{t^n}) &\leq (2 \uparrow\uparrow m) \cdot f(2 \uparrow\uparrow m, 2 \uparrow\uparrow (2m+2)) \\ &\leq (2 \uparrow\uparrow m) \cdot [(2 \uparrow\uparrow (2m+2)) \uparrow\uparrow (2 \cdot 2 \uparrow\uparrow m)] \end{aligned}$$

$$\leq (2 \uparrow\uparrow m) \cdot [(2 \uparrow\uparrow (2m+2)) \uparrow\uparrow (2 \uparrow\uparrow (m+1))]$$
$$\leq (2 \uparrow\uparrow m) \cdot [2 \uparrow\uparrow ((2m+2) \cdot 2 \uparrow\uparrow (m+1))]$$
$$\leq (2 \uparrow\uparrow m) \cdot [2 \uparrow\uparrow (2 \uparrow\uparrow (m+2))]$$
$$\leq 2 \uparrow\uparrow (1 + 2 \uparrow\uparrow (m+2))$$
$$\leq 2 \uparrow\uparrow (2 \uparrow\uparrow (m+3)). \qquad \square$$

Lemma 2.5.2. $\mathrm{HJ}(4,2,6) < 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 9)) < 2 \uparrow\uparrow\uparrow 6$.

Proof. The doubly exponential bound on $\mathrm{HJ}(2,2,6)$ from [46] yields $\mathrm{HJ}(2,2,6) \leq 2^{2^{12}} < 2 \uparrow\uparrow 5$. Applying Lemma 2.5.1, we get $\mathrm{HJ}(3,2,6) < 2 \uparrow\uparrow (2 \uparrow\uparrow 8)$, and

$$\begin{aligned}
\mathrm{HJ}(4,2,6) &< 2 \uparrow\uparrow (2 \uparrow\uparrow (3 + 2 \uparrow\uparrow 8)) \\
&< 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 9)) \\
&< 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 65536)) \\
&= 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 2)))) \\
&= 2 \uparrow\uparrow\uparrow 6. \qquad \square
\end{aligned}$$

With $d = 2 \uparrow\uparrow 18$ in place of 6, we have $\mathrm{HJ}(2,2,d) \leq 2^{2^{2 \cdot (2 \uparrow\uparrow 18)}} < 2 \uparrow\uparrow 21$ instead. Applying Lemma 2.5.1, we get $\mathrm{HJ}(3,2,d) < 2 \uparrow\uparrow (2 \uparrow\uparrow 24)$, and

$$\begin{aligned}
\mathrm{HJ}(4,2,d) &< 2 \uparrow\uparrow (2 \uparrow\uparrow (3 + 2 \uparrow\uparrow 24)) \\
&< 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 25)).
\end{aligned}$$

Chapter 3

An Upper Bound for the Hales–Jewett Number $\mathrm{HJ}(4, 2)$

3.1 Background

Consider r-colorings of the n-dimensional grid $[t]^n = \{1, 2, \ldots, t\}^n$. We define a *combinatorial line* in $[t]^n$ to be an injective function $\ell : [t] \to [t]^n$ such that for each coordinate $1 \leq i \leq n$, ℓ_i is either constant or the identity function on $[t]$. An example of such a line in $[4]^5$ is the function

$$\ell(x) = (3, x, 1, x, 4)$$

whose image is the set of four points $\{(3, 1, 1, 1, 4), (3, 2, 1, 2, 4), (3, 3, 1, 3, 4), (3, 4, 1, 4, 4)\}$. A classic result in Ramsey theory, the Hales–Jewett Theorem [30], asserts that for all values of the parameters t and r, there exists a sufficiently large n such that any r-coloring of $[t]^n$ will contain a monochromatic combinatorial line (that is, a line such that the points in its image are all assigned the same color). The Hales–Jewett number $\mathrm{HJ}(t, r)$ is defined to be the least n which suffices.

A pigeonhole argument is enough to show that $\mathrm{HJ}(2, r) = r$ for all r. Moreover, Hindman and Tressler [31] have shown that $\mathrm{HJ}(3, 2) = 4$. For more difficult cases, no exact values are known, and the best upper bounds were shown by Shelah [46]. For $\mathrm{HJ}(4, 2)$, this upper bound

is already enormous. Shelah proves that if $\mathrm{HJ}(t-1,r) \le n$, then $\mathrm{HJ}(t,r) \le nf(n, r^{t^n})$, where $f(\ell,k)$ is an upper bound for a related problem that satisfies $k \uparrow\uparrow \ell \le f(\ell,k) \le k \uparrow\uparrow (2\ell)$. Starting from $\mathrm{HJ}(3,2) = 4$, we get a bound for $\mathrm{HJ}(4,2)$ between $2 \uparrow\uparrow 7$ and $2 \uparrow\uparrow 8$.

On the other hand, the best lower bounds for the Hales–Jewett numbers, which can be obtained from the van der Waerden theorem (as in [46]), are very far away from these upper bounds. For instance, the integers $\{1, 2, \ldots, 34\}$ can be 2-colored in such a way that no 4-term arithmetic progression is monochromatic [13]. This coloring can be used to define a coloring of $[4]^{11}$ with no monochromatic combinatorial line, by coloring a point $(x_1, x_2, \ldots, x_{11})$ with the color of $x_1 + x_2 + \cdots + x_{11} - 10$, which shows that $\mathrm{HJ}(4,2) > 11$. In general, this argument yields merely exponential lower bounds on the Hales–Jewett numbers: Berlekamp [5] showed that for prime p, we can 2-color $p \cdot 2^p$ consecutive integers with no monochromatic p-term arithmetic progression. This opens up the possibility that the true values of the Hales–Jewett numbers may be much smaller.

We find a much smaller upper bound on $\mathrm{HJ}(4,2)$, proving the following result:

Theorem 3.1.1. *Whenever the 10^{11}-dimensional grid $[4]^{10^{11}}$ is 2-colored, there exists a monochromatic combinatorial line. That is, $\mathrm{HJ}(4,2) \le 10^{11}$.*

3.2 Setup

Given a combinatorial line $\ell : [4] \to [4]^n$, we define its *length* $|\ell|$ to be the number of coordinates $1 \le i \le n$ for which $\ell_i(x)$ varies with x. For example, the line given by $\ell(x) = (3, x, 1, x, 4)$ has length 2. (To justify this terminology, note that $|\ell|$ is the Hamming distance between the two endpoints $\ell(1)$ and $\ell(4)$, or indeed between any two points of ℓ.)

Fix a 2-coloring of $[4]^n$. For each length k, we classify the combinatorial lines of length k into three types, and count their densities:

- $p_2(k)$ is the fraction of lines of length k which have 2 points of each color.

- $p_3(k)$ is the fraction of lines of length k which have 3 points of one color, and 1 of the other.
- $p_4(k)$ is the fraction of monochromatic lines of length k.

Since this classification is exhaustive, we have $p_2(k) + p_3(k) + p_4(k) = 1$.

We are also interested in the fraction of pairs of collinear points (on lines of length k) assigned the same color. On each line, there are 6 pairs of points. For lines counted by p_2, 2 pairs are monochromatic; for lines counted by p_3, 3 pairs are monochromatic; for lines counted by p_4, all 6 pairs are monochromatic. Therefore

$$q(k) := \frac{1}{3}p_2(k) + \frac{1}{2}p_3(k) + p_4(k) \tag{3.1}$$

counts the fraction of monochromatic pairs of points on lines of length k.

The grid $[4]^n$ contains large "cliques": sets of points in which any two are collinear. In such a clique, the number of monochromatic pairs is least when the colors are balanced, and even then is close to $\frac{1}{2}$. Thus, we expect that for at least some lengths k, $q(k) > \frac{1}{2} - \epsilon$ for some ϵ that goes to 0 with n. This intuition is correct in a way that we will make more precise.

If we could show the stronger statement that $q(k) > \frac{1}{2}$ for some k, the proof would be complete: $p_4(k)$ occurs in (3.1) with a coefficient greater than $\frac{1}{2}$, so we would know that $p_4(k) > 0$, which means that a monochromatic line exists.

Even if $q(k) < \frac{1}{2}$, a large value of $q(k)$ gives partial information: either $p_4(k) > 0$, or else $p_3(k)$ is close to 1. Therefore we can also prove that $p_4(k) > 0$ by showing that for some k, $q(k)$ is close to $\frac{1}{2}$, but $p_3(k)$ is bounded away from 1. More formally, we can solve (3.1) for $p_4(k)$ (substituting $p_2(k) = 1 - p_3(k) - p_4(k)$) to get

$$p_4(k) = \frac{3}{2}\left(q(k) - \frac{1}{6}p_3(k) - \frac{1}{3}\right). \tag{3.2}$$

We will prove that a monochromatic line exists by showing that the right-hand side of (3.2) is positive for some k.

3.3 Showing that $q(k)$ is close to $\frac{1}{2}$

It is hopeless to show that $q(k)$ approaches $\frac{1}{2}$ for any individual k. For example, the "checkerboard" coloring, which colors a point by the sum of its coordinates modulo 2, has $p_2(k) = 1$, and therefore $q(k) = \frac{1}{3}$, for all odd k. Instead, we prove an inequality for a weighted sum of the first s values of $q(k)$, where s is a parameter to be determined later.

3.3.1 A bound for n-dimensional hypercubes

We begin by considering collinear pairs in the hypercube $[2]^n$. Here, lines consist of only 2 points, and therefore $q(k)$, defined as before, simply counts monochromatic lines of length k.

Lemma 3.3.1. *For every $s \leq n$, whenever $[2]^n$ is 2-colored,*

$$\sum_{k=1}^{s}(s-k+1)q(k) \geq \frac{s^2-1}{4}\left(1 - s\sqrt{\frac{2}{\pi n}}\right). \tag{3.3}$$

Proof. The hypercube $[2]^n$ is the union of $n!$ chains of length $n+1$: maximal sequences of points $C_0, \ldots, C_n$ such that any two points C_i and C_j are collinear. For each permutation σ of $[n]$, we obtain such a chain by letting $C_i(\sigma)$ be the point with 2 in the coordinates $\sigma(1), \sigma(2), \ldots, \sigma(i)$ and 1 in all others. The line through $C_i(\sigma)$ and $C_j(\sigma)$, for $i < j$, has length $j - i$. Any permutation σ' such that $\{\sigma(1), \ldots, \sigma(i)\} = \{\sigma'(1), \ldots, \sigma'(i)\}$ and $\{\sigma(i+1), \ldots, \sigma(j)\} = \{\sigma'(i+1), \ldots, \sigma'(j)\}$ will satisfy $C_i(\sigma) = C_i(\sigma')$ and $C_j(\sigma) = C_j(\sigma')$; therefore $C_i(\sigma)$ and $C_j(\sigma)$ occur together in $i!(j-i)!(n-j)!$ chains.

Fix any 2-coloring of $[2]^n$. Let $Q_{i,j}(\sigma) = 1$ if $C_i(\sigma)$ and $C_j(\sigma)$ are given the same color, and 0 otherwise. There are $\binom{n}{k}2^{n-k}$ total lines of length k; if a $q(k)$ fraction of them are monochromatic, the number of monochromatic lines is $\binom{n}{k}2^{n-k}q(k)$. We can express this quantity in terms of the indicators $Q_{i,j}(\sigma)$ as:

$$\begin{aligned}\binom{n}{k}2^{n-k}q(k) &= \sum_{\sigma}\sum_{i=0}^{n-k}\frac{Q_{i,i+k}(\sigma)}{i!k!(n-i-k)!} \\ &= \frac{1}{n!}\binom{n}{k}\sum_{\sigma}\sum_{i=0}^{n-k}\binom{n-k}{i}Q_{i,i+k}(\sigma).\end{aligned}$$

Therefore

$$q(k) = \frac{1}{n!} \sum_{\sigma} \left(\sum_{i=0}^{n-k} \frac{\binom{n-k}{i}}{2^{n-k}} Q_{i,i+k}(\sigma) \right). \tag{3.4}$$

Let

$$q(k,\sigma) = \sum_{i=0}^{n-k} \frac{\binom{n-k}{i}}{2^{n-k}} Q_{i,i+k}(\sigma).$$

Then equation (3.4) shows that $q(k)$ is the average of $q(k,\sigma)$ over all σ. To complete the proof of Lemma 3.3.1, it suffices to show that for each permutation σ, inequality (3.3) holds with $q(k,\sigma)$ in place of $q(k)$.

Define $w_{i,j} := \frac{\binom{n-(j-i)}{i}}{2^{n-(j-i)}}$, a shorthand for the coefficient of $Q_{i,j}(\sigma)$ in $q(j-i,\sigma)$. We have

$$\sum_{k=1}^{s} (s-k+1) q(k,\sigma) \geq \sum_{h=0}^{n-s} \left(\sum_{\substack{i,j \in [h,h+s] \\ i<j}} w_{i,j} Q_{i,j}(\sigma) \right) \tag{3.5}$$

because each pair i, j is contained in at most $s-(j-i)+1$ intervals $[h, h+s]$ (fewer if $j < s$ or $i > n-s$), and so each term $w_{i,j}Q_{i,j}(\sigma)$ occurs in the right-hand side of (3.5) for at most $s-(j-i)+1$ values of h.

Each sum

$$\sum_{\substack{i,j \in [h,h+s] \\ i<j}} w_{i,j} Q_{i,j}(\sigma)$$

counts (with varying weights) the number of monochromatic pairs among the $s+1$ points $C_h(\sigma)$, $C_{h+1}(\sigma), \ldots, C_{h+s}(\sigma)$, any two of which are collinear. There must be at least $2\binom{(s+1)/2}{2} = \frac{s^2-1}{4}$ such pairs; their number is minimized if half the points receive one color and half receive the other. We do not know which weights correspond to those pairs. However, at the very least, we have the lower bound

$$\sum_{\substack{i,j \in [h,h+s] \\ i<j}} w_{i,j} Q_{i,j}(\sigma) \geq \frac{s^2-1}{4} w_h^*,$$

where w_h^* is the least of all weights $w_{i,j}$ for $i, j \in [h, h+s]$ with $i \leq j$. (We allow $i = j$ to simplify calculations later, though such a weight does not occur in the sum.) Substituting this

lower bound into inequality (3.5), we get

$$\sum_{k=1}^{s}(s-k+1)q(k,\sigma) \geq \frac{s^2-1}{4}\sum_{h=0}^{n-s} w_h^*.$$

It remains to find a lower bound for the sum of the w_h^*.

From Pascal's identity it follows that $w_{i,j} = \frac{1}{2}(w_{i-1,j} + w_{i,j+1})$. Therefore each coefficient $w_{i,j}$ is a weighted average of some of the coefficients

$$w_{h,h}, w_{h,h+1}, \ldots, w_{h,h+s}, w_{h+1,h+s}, \ldots, w_{h+s,h+s}.$$

Since all of these coefficients are included in the minimum defining w_h^*, we know that w_h^* must be one of these. Furthermore, this sequence is unimodal, so $w_h^* = \min\{w_{h,h}, w_{h+s,h+s}\}$.

The sequence $w_{0,0}, w_{1,1}, \ldots, w_{n,n}$ is just the sequence $\frac{\binom{n}{0}}{2^n}, \frac{\binom{n}{1}}{2^n}, \ldots, \frac{\binom{n}{n}}{2^n}$, and is also unimodal. So the sum $\sum_{h=0}^{n-s} w_h^*$ will begin by summing $w_{h,h}$ and eventually switch to summing $w_{h+s,h+s}$, skipping some s terms. Therefore

$$\sum_{h=0}^{n-s} w_h^* \geq \sum_{h=0}^{n} w_{h,h} - s\max_{0\leq h\leq n} w_{h,h} = 1 - s\frac{\binom{n}{\lfloor n/2\rfloor}}{2^n} \geq 1 - s\sqrt{\frac{2}{\pi n}}.$$

It follows that

$$\sum_{k=1}^{s}(s-k+1)q(k,\sigma) \geq \frac{s^2-1}{4}\left(1 - s\sqrt{\frac{2}{\pi n}}\right)$$

and by averaging this inequality over all permutations σ and applying equation (3.5), we obtain the desired inequality (3.3). □

3.3.2 Extending the bound to the grid $[4]^n$

By giving away another error term, we can extend Lemma 3.3.1 to all collinear pairs in $[4]^n$.

Lemma 3.3.2. *For every $s \leq \frac{n}{4}$, whenever $[4]^n$ is 2-colored,*

$$\sum_{k=1}^{s}(s-k+1)q(k) \geq \frac{s^2-1}{4}\left(1 - e^{-(n-s)/8} - 3s\sqrt{\frac{2}{\pi(n-s)}}\right).$$

Proof. We say that a collinear pair of points $\ell(a), \ell(b)$ for some line ℓ and some $a, b \in [4]$ has *type* m if there are m coordinates in total in which either point is equal to a or b; in other words, $\ell_i(x)$ is the constant a or b for $m - |\ell|$ values of i. We define $q(k, m)$ to be the fraction of collinear pairs of type m and on lines of length k which are monochromatic.

The type of a collinear pair matters because a collinear pair of type m is contained in the m-dimensional subcube of $[4]^n$ obtained by letting all m coordinates of either point which are equal to a or b vary freely between the two values. In this m-dimensional subcube, two points are collinear if and only if the corresponding points of $\{a, b\}^m$ (obtained by dropping all coordinates not equal to a or b) are collinear, so it has the structure of the hypercube $[2]^m$. The fraction of collinear pairs in this subcube which are monochromatic satisfies Lemma 3.3.1. By averaging over all m-dimensional subcubes, which cover each collinear pair of type m exactly once, we obtain

$$\sum_{k=1}^{s}(s-k+1)q(k,m) \geq \frac{s^2-1}{4}\left(1 - s\sqrt{\frac{2}{\pi m}}\right). \tag{3.6}$$

There are $6\binom{n}{k}4^{n-k}$ collinear pairs on lines of length k; of them, $\binom{m}{k}2^{m-k}$ are in each m-dimensional subcube, and there are $6\binom{n}{m}2^{n-m}$ such subcubes. So the fraction of lines of length k which have type m is

$$\frac{\binom{m}{k}2^{m-k}\binom{n}{m}2^{n-m}}{\binom{n}{k}4^{n-k}} = \frac{\binom{n-k}{m-k}}{2^{n-k}}.$$

Therefore we may express $q(k)$ as a weighted average of all the $q(k, m)$ by

$$q(k) = \sum_{m=k}^{n} \frac{\binom{n-k}{m-k}}{2^{n-k}} q(k,m). \tag{3.7}$$

Unfortunately, the weight of $q(k, m)$ in this average depends on k as well as m, which prevents us from simply averaging inequality (3.6) over all m. To fix this problem, we replace the weights in (3.7) by lower bounds independent of k, which will result in an inequality relating $q(k)$ to $q(k, m)$. (We will assume that $1 \leq k \leq s \leq \frac{n}{4}$.)

For $m \leq \frac{n}{4}$, our lower bound will be 0: we drop all terms where m is too low, because the statement of inequality (3.6) is too weak in such cases. Otherwise, we want to replace the weight by the minimum of $\binom{n-k}{m-k}2^{-(n-k)}$ over all $k \leq s$.

From Pascal's identity, we have $\binom{n}{r}2^{-n} = \frac{1}{2}\left(\binom{n-1}{r-1}2^{-(n-1)} + \binom{n-1}{r}2^{-(n-1)}\right)$. Applying this iteratively, we can express each $\binom{n-k}{m-k}2^{-(n-k)}$ as a weighted average of some of

$$\frac{\binom{n-s}{m-s}}{2^{n-s}}, \frac{\binom{n-s}{m-s+1}}{2^{n-s}}, \ldots, \frac{\binom{n-s}{m}}{2^{n-s}}.$$

This sequence is unimodal, so the minimum is achieved at one of the endpoints, and we may replace equation (3.7) by

$$q(k) \geq \sum_{m=n/4}^{n} \frac{\min\left\{\binom{n-s}{m-s}, \binom{n-s}{m}\right\}}{2^{n-s}} q(k,m). \tag{3.8}$$

Sum the inequality (3.6) over all $m \geq \frac{n}{4}$ with weights as in inequality (3.8). The right-hand side of (3.8) will be smallest when $m = \frac{n}{4}$, so we may use that value for all m. We obtain

$$\sum_{k=1}^{s}(s-k+1)q(k) \geq \frac{s^2-1}{4}\left(1 - s\sqrt{\frac{2}{\pi n/4}}\right)\sum_{m=n/4}^{n}\frac{\min\left\{\binom{n-s}{m-s}, \binom{n-s}{m}\right\}}{2^{n-s}}. \tag{3.9}$$

It remains to simplify the right-hand side.

The omission of the first $n/4$ terms of the sum in (3.9) results in an error of $\sum_{m<n/4}\binom{n-s}{m-s}2^{-(n-s)}$, which is simply the binomial probability $\Pr[\mathrm{Bin}(n-s, \frac{1}{2}) < \frac{n}{4} - s]$. By the Chernoff bound (see, e.g., [1]),

$$\Pr\left[\mathrm{Bin}(n-s,\tfrac{1}{2}) < \frac{n}{4} - s\right] < \Pr\left[\mathrm{Bin}(n-s,\tfrac{1}{2}) < \frac{n-s}{4}\right] \leq \exp\left(-\frac{n-s}{8}\right).$$

With these initial terms, the sum in (3.9) would be equal to 1, except for skipping s terms near the middle, which occurs when the minimum switches from selecting $\binom{n-s}{m-s}$ to selecting $\binom{n-s}{m}$. Each of these terms is at most $\binom{n-s}{(n-s)/2}2^{-(n-s)} \leq \sqrt{\frac{2}{\pi(n-s)}}$, so we lose at most s times this quantity. Therefore the sum in inequality (3.9) satisfies

$$\sum_{m=n/4}^{n}\frac{\min\left\{\binom{n-s}{m-s}, \binom{n-s}{m}\right\}}{2^{n-s}} \geq 1 - e^{-(n-s)/8} - s\sqrt{\frac{2}{\pi(n-s)}}.$$

Combining the two error terms, we complete the proof. □

3.4 Showing that $p_3(k)$ cannot be arbitrarily close to 1

In this section, we say that a combinatorial line in a 2-colored grid $[4]^n$ is *odd* if it has an odd number of points of each color. That is, an odd line has 3 points of one color and 1 point of the other, so it is exactly the type of line counted by $p_3(k)$.

To bound $p_3(k)$ away from 1, we first find a set of lines in $[4]^4$ which cannot all be odd:

Lemma 3.4.1. *Whenever $[4]^4$ is 2-colored, the 15 lines*

$$\begin{array}{lll}
\ell^1(x) = (x,2,3,4) & \ell^6(x) = (x,2,x,4) & \ell^{11}(x) = (x,x,x,4) \\
\ell^2(x) = (1,x,3,4) & \ell^7(x) = (x,2,3,x) & \ell^{12}(x) = (x,x,3,x) \\
\ell^3(x) = (1,2,x,4) & \ell^8(x) = (1,x,x,4) & \ell^{13}(x) = (x,2,x,x) \\
\ell^4(x) = (1,2,3,x) & \ell^9(x) = (1,x,3,x) & \ell^{14}(x) = (1,x,x,x) \\
\ell^5(x) = (x,x,3,4) & \ell^{10}(x) = (1,2,x,x) & \ell^{15}(x) = (x,x,x,x)
\end{array}$$

cannot all be odd.

Proof. A key observation is that each point of $[4]^4$ lies on an even number of these lines. The point $(1,2,3,4)$ lies on the 4 lines of length 1, and no other. Take any other point (x_1,x_2,x_3,x_4) expressible as $\ell^j(x)$ for some index j and some $x \in \{1,2,3,4\}$. Any coordinate i where $x_i \neq i$ must be a variable coordinate of ℓ^j; any coordinate i where $x_i = i \neq x$ must be a constant coordinate of ℓ^j. There is always exactly one coordinate where $x_i = i = x$, so there are 2 choices for j, depending on whether that coordinate is variable or constant.

If $[4]^4$ is 2-colored, choose either of the colors, and add up the number of points of that color on each of the fifteen lines. This total must be even, because each point is counted an even number of times. However, 15 odd numbers cannot add up to an even total, so one of the lines must contribute an even number. Therefore not all 15 lines can be odd. □

Structures isomorphic to the set of lines $\ell^1,\ldots,\ell^{15}$ occur many times in $[4]^n$, and in each such structure at most $\frac{14}{15}$ of the lines are odd. So our next step is to show that by (more or less)

averaging over all such structures, we get an upper bound of $\frac{14}{15}$ for the overall densities of odd lines of certain lengths, up to an error term.

Lemma 3.4.2. *For every $k \leq \frac{n}{4}$, whenever $[4]^n$ is 2-colored,*

$$\left(1 - \frac{16k^2}{n}\right)\left(\frac{4}{15}p_3(k) + \frac{6}{15}p_3(2k) + \frac{4}{15}p_3(3k) + \frac{1}{15}p_3(4k)\right) \leq \frac{14}{15}. \tag{3.10}$$

Proof. Fix a 2-coloring of $[4]^n$ and some $k \leq \frac{n}{4}$. Let a *k-embedding* of $[4]^4$ into $[4]^n$ be a function $L : [4]^4 \to [4]^n$ such that for each coordinate $1 \leq i \leq n$, L_i is either constant or given by $L_i(x_1, x_2, x_3, x_4) = x_j$ for some $j \in \{1, 2, 3, 4\}$. Moreover, we require that for each j, there are exactly k coordinates in which L_i varies with x_j. Let $\mathcal{L}_k$ be the set of all k-embeddings $[4]^4 \to [4]^n$.

Each $L \in \mathcal{L}_k$ induces a 2-coloring of $[4]^4$, by taking the preimage under L of the coloring of $[4]^n$. Moreover, a line $\ell : [4] \to [4]^4$ corresponds to a line $L \circ \ell : [4] \to [4]^n$, with $|L \circ \ell| = k|\ell|$, which is odd if and only if ℓ is odd in the induced coloring.

Count the number of odd lines $L \circ \ell^j$, where $L \in \mathcal{L}_k$ and ℓ^j is one of the 15 lines of Lemma 3.4.1. For a fixed L, at most 14 of the lines $L \circ \ell^j$ are odd; therefore we count at most $14|\mathcal{L}_k|$ odd lines total.

Let $P_3(k)$ be the number of odd lines of length k in $[4]^n$, related to the density $p_3(k)$ by $P_3(k) = \binom{n}{k}4^{n-k}p_3(k)$. If each line of length k could be expressed $M(k)$ times as $L \circ \ell^j$, we would have the inequality

$$M(k)P_3(k) + M(2k)P_3(2k) + M(3k)P_3(3k) + M(4k)P_3(4k) \leq 14|\mathcal{L}_k|. \tag{3.11}$$

Unfortunately, the number of ways to express a line $\ell : [4] \to [4]^n$ as $L \circ \ell^j$ depends on ℓ; specifically, on the number of coordinates of ℓ with each constant value. Inequality (3.11) still holds, however, if we instead define $M(k)$ to be the minimum multiplicity of any line of length k.

We compute the minimum multiplicity for each of the four possible lengths. Below, let n_1, n_2, n_3, and n_4 denote the number of coordinates of ℓ with constant value 1, 2, 3, and 4, respectively.

- If $|\ell| = k$, then ℓ can be expressed as $L(x, 2, 3, 4)$ in $\binom{n_2}{k}\binom{n_3}{k}\binom{n_4}{k}$ ways; we also get the corresponding counts for expressions of the form $L(1, x, 3, 4)$, $L(1, 2, x, 4)$, and $L(1, 2, 3, x)$. In total the line is counted with multiplicity $\binom{n_2}{k}\binom{n_3}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_3}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_2}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_2}{k}\binom{n_3}{k}$. This sum is minimized when n_1, n_2, n_3, n_4 are as equal as possible, so

$$M(k) \geq 4\binom{\frac{n-k}{4}}{k}^3.$$

- If $|\ell| = 2k$, then ℓ can be expressed as $L(x, x, 3, 4)$ in $\binom{2k}{k}\binom{n_3}{k}\binom{n_4}{k}$ ways; we also get the corresponding counts for expressions of the form $L(x, 2, x, 4)$ and so on, for a total of $\binom{2k}{k}\left(\binom{n_3}{k}\binom{n_4}{k} + \binom{n_2}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_4}{k} + \binom{n_2}{k}\binom{n_3}{k} + \binom{n_1}{k}\binom{n_3}{k} + \binom{n_1}{k}\binom{n_2}{k}\right)$. This is, once again, minimized when n_1, n_2, n_3, n_4 are as equal as possible, so

$$M(2k) \geq 6\binom{2k}{k}\binom{\frac{n-2k}{4}}{k}^2.$$

- If $|\ell| = 3k$, then ℓ can be expressed as $L(1, x, x, x)$ or $L(x, 2, x, x)$ or $L(x, x, 3, x)$ or $L(x, x, x, 4)$ in a total of $\binom{3k}{k,k,k}\left(\binom{n_1}{k} + \binom{n_2}{k} + \binom{n_3}{k} + \binom{n_4}{k}\right)$ ways, so

$$M(3k) \geq 4\binom{3k}{k, k, k}\binom{\frac{n-3k}{4}}{k}.$$

- Finally, if $|\ell| = 4k$, then ℓ can only be expressed as $L(x, x, x, x)$, which can be done in

$$M(4k) = \binom{4k}{k, k, k, k}$$

ways.

We can further replace $|\mathcal{L}_k|$ by $\binom{n}{k,k,k,k,n-4k}4^{n-4k}$. This allows us to rewrite inequality (3.11) as

$$\begin{aligned} 4\binom{\frac{n-k}{4}}{k}^3\binom{n}{k}4^{n-k}p_3(k) + 6\binom{2k}{k}\binom{\frac{n-2k}{4}}{k}^2\binom{n}{2k}4^{n-2k}p_3(2k) + \\ + 4\binom{3k}{k, k, k}\binom{\frac{n-3k}{4}}{k}\binom{n}{3k}4^{n-3k}p_3(3k) + \binom{4k}{k, k, k, k}\binom{n}{4k}4^{n-4k}p_4(4k) \leq \\ \leq \binom{n}{k, k, k, k, n-4k}4^{n-4k}. \end{aligned}$$

This inequality can be simplified by factoring out $\frac{4^{n-4k}}{k!^4}$ from each term. If we also replace falling powers of the form $x(x-1)(x-2)(\cdots)(x-r+1)$ by x^r on the right-hand side (as an upper bound) and by $(x-r)^r$ on the left-hand side (as a lower bound), we obtain

$$4(n-k)^k(n-5k)^{3k}p_3(k) + 6(n-2k)^{2k}(n-6k)^{2k}p_3(2k)+ \\ + 4(n-3k)^{3k}(n-7k)^k p_3(3k) + (n-4k)^{4k}p_3(4k) \leq 14n^{4k}.$$

Finally, dividing through by n^{4k} yields factors such as $\left(1-\frac{k}{n}\right)^k$. By iteratively applying the inequality $(1-x)(1-y) \geq 1-x-y$ for $x, y \geq 0$, we bound the first such factor:

$$\left(1-\frac{k}{n}\right)^k\left(1-\frac{5k}{n}\right)^{3k} \geq \left(1-\frac{k^2}{n}\right)\left(1-\frac{15k^2}{n}\right) \geq 1-\frac{16k^2}{n}.$$

Similarly, the factors of $\left(1-\frac{2k}{n}\right)^{2k}\left(1-\frac{6k}{n}\right)^{2k}$, $\left(1-\frac{3k}{n}\right)^{3k}\left(1-\frac{7k}{n}\right)^k$, and $\left(1-\frac{4k}{n}\right)^{4k}$ are each at most $1-\frac{16k^2}{n}$. After pulling out this factor, we obtain the inequality (3.10). □

3.5 Completing the proof of Theorem 5.1.2

To simplify notation, let $p_3^+(k) := \frac{4}{15}p_3(k) + \frac{6}{15}p_3(2k) + \frac{4}{15}p_3(3k) + \frac{1}{15}p_3(4k)$ (the quantity bounded by Lemma 3.4.2) and let $q^+(k) := \frac{4}{15}q(k) + \frac{6}{15}q(2k) + \frac{4}{15}q(3k) + \frac{1}{15}q(4k)$. We noted previously that if $q(k) - \frac{1}{6}p_3(k) > \frac{1}{3}$ for some k, then equation (3.2) implies that $p_4(k) > 0$, so a monochromatic line exists. Similarly, showing that $q^+(k) - \frac{1}{6}p_3^+(k) > \frac{1}{3}$ is positive suffices: this is a weighted average, so $q(ik) - \frac{1}{6}p_3(ik) > \frac{1}{3}$ will hold for some $1 \leq i \leq 4$.

We express as much of the left-hand side of Lemma 3.3.2 as possible in terms of q^+. We assume that the still-undetermined parameter s is a multiple of 4 for simplicity. In the sum

$$\sum_{k=1}^{s/4}(s+1-2k)q^+(k) \tag{3.12}$$

the coefficient of each $q(k)$ is 0 for $k > s$, and otherwise maximized if k is divisible by both 3 and 4, in which case it is at most

$$\frac{4}{15}(s+1-2\cdot k)+\frac{6}{15}\left(s+1-2\cdot\frac{k}{2}\right)+\frac{4}{15}\left(s+1-2\cdot\frac{k}{3}\right)+\frac{1}{15}\left(s+1-2\cdot\frac{k}{4}\right) = s+1-\frac{103}{90}k,$$

so it is always less than $s+1-k$. This means pulling out the sum (3.12) from the left-hand side of Lemma 3.3.2 leaves each $q(k)$ with a positive coefficient: we may write

$$\sum_{k=1}^{s}(s+1-k)q(k) = \sum_{k=1}^{s/4}(s+1-2k)q^+(k) + \sum_{k=1}^{s} R_k q(k) \tag{3.13}$$

where $R_1, \ldots, R_s$ are all positive. Furthermore, though each R_k is tedious to calculate, since equation (3.13) is valid for all values of $q(1), \ldots, q(s)$, it remains valid if we set each of them to 1, and therefore

$$\sum_{k=1}^{s} R_k = \sum_{k=1}^{s}(s+1-k) - \sum_{k=1}^{s/4}(s+1-2k) = \frac{s^2+s}{2} - \frac{3s^2}{16} = \frac{5s^2+8s}{16}.$$

If $q(k) > \frac{1}{2}$ for any k, then from equation (3.1) we can conclude that $p_4(k) > 0$ and a monochromatic line exists. So assume the contrary: that $q(k) \leq \frac{1}{2}$ for all k. Then equation (3.13) implies that

$$\sum_{k=1}^{s}(s+1-k)q(k) \leq \sum_{k=1}^{s/4}(s+1-2k)q^+(k) + \frac{5s^2+8s}{16}\cdot\frac{1}{2}.$$

Therefore, by applying Lemma 3.3.2,

$$\sum_{k=1}^{s/4}(s+1-2k)q^+(k) \geq \frac{s^2-1}{4}(1-\epsilon(n,s)) - \frac{5s^2+8s}{32},$$

where $\epsilon(n,s)$ is the relative error term

$$\epsilon(n,s) := e^{-(n-s)/8} + 3s\sqrt{\frac{2}{\pi(n-s)}}.$$

Dividing by $\frac{3s^2}{16}$ to obtain a weighted average and simplifying, we are left with

$$\frac{16}{3s^2}\sum_{k=1}^{s/4}(s+1-2k)q^+(k) \geq \frac{1}{2} - \frac{4}{3s} - \frac{4}{3}\epsilon(n,s) + \frac{4(1-\epsilon(n,s))}{3s^2}.$$

The last term is positive and may be dropped. Therefore there is some length $k^* \leq s/4$ for which $q^+(k^*) \geq \frac{1}{2} - \frac{4}{3s} - \frac{4}{3}\epsilon(n,s)$.

On the other hand, Lemma 3.4.2 tells us that, as long as $4k^* < \sqrt{n}$, $p_3^+(k^*) \leq \frac{14}{15}\left(1 - \frac{16(k^*)^2}{n}\right)^{-1}$, which is at most $\frac{14}{15}\left(1-\frac{s^2}{n}\right)^{-1}$. Therefore a lower bound on $q^+(k^*) - \frac{1}{6}p_3^+(k^*) - \frac{1}{3}$ is

$$\frac{1}{6} - \frac{4}{3s} - \frac{4}{3}\epsilon(n,s) - \frac{7}{45}\left(1-\frac{s^2}{n}\right)^{-1}, \tag{3.14}$$

which is valid for any $s < \sqrt{n}$.

As $n \to \infty$ and $s \to \infty$, provided that $\frac{s^2}{n} \to 0$, (3.14) approaches $\frac{1}{90}$, and so a monochromatic line must exist. In particular, (3.14) is already positive for $n = 10^{11}$ and $s = 368$, completing the proof. (More precisely, $n = 19\,012\,590\,257$ and $s = 240$ are enough; this is the least value of n which makes (3.14) positive, though this bound could almost certainly be improved by tweaking the overall argument.)

Chapter 4

The Game Chromatic Number of Random 3-Regular Graphs

4.1 Background

We now switch gears and consider more graph-theoretic results. This chapter, which solves an easy problem related to graph coloring, may serve as a warm-up to the rest of this thesis.

Let $G = (V, E)$ be a graph and let k be a positive integer. We consider a game in which players Alice and Bob take turns in coloring the vertices of G with k colors. Each move consists of choosing an uncolored vertex of the graph and assigning to it a color from $\{1, \ldots, k\}$ so that the resulting coloring is *proper*, i.e., adjacent vertices get different colors.

Alice wins if all the vertices of G are eventually colored. Bob wins if, at some point in the game, the current partial coloring cannot be extended to a complete coloring of G. This impossibility may be initially subtle, but once Bob wins, we may as well continue the game until his victory is obvious: there is an uncolored vertex such that each of the k colors appears at least once in its neighborhood.

We assume that Alice goes first (our results will not be sensitive to this choice). The *game chromatic number* $\chi_g(G)$ is the least integer k for which Alice has a winning strategy. Note

that (as Winkler points out in [51, 52]) it does not immediately follow that Alice has a winning strategy for each $k > \chi_g(G)$; this seems intuitively clear, but is still open.

This parameter is well defined, since it is easy to see that Alice always wins if the number of colors is larger than the maximum degree of G. Clearly, $\chi_g(G)$ is at least as large as the ordinary chromatic number $\chi(G)$, but it can be considerably more.

The game was first considered by Brams about 25 years ago in the context of coloring planar graphs and was described in Martin Gardner's column [23] in Scientific American in 1981. The game remained unnoticed by the graph-theoretic community until Bodlaender [6] re-invented it. For a survey see Bartnicki, Grytczuk, Kierstead and Zhu [4].

Some results are known for the game chromatic number of random graphs. Bohman, Frieze, and Sudakov [7] consider the game chromatic number of the Erdős–Rényi random graph $\mathcal{G}_{n,p}$ in the relatively dense setting: $p \geq \frac{(\log n)^c}{n}$ for large c. In this range, we know that the ordinary chromatic number satisfies $\chi(\mathcal{G}_{n,p}) \sim \frac{n}{2\log_b np}$, where $b = \frac{1}{1-p}$. To contrast with this, Bohman, Frieze, and Sudakov prove the following:

1. There exists $K > 0$ such that for $\epsilon > 0$ and $p \geq (\log n)^{K\epsilon^{-3}}/n$ we have that a.a.s.

$$\chi_g(\mathcal{G}_{n,p}) \geq (1-\epsilon)\frac{n}{\log_b np}.$$

2. If $\alpha > 2$ is a constant, $K = \max\{\frac{2\alpha}{\alpha-1}, \frac{\alpha}{\alpha-2}\}$, and $p \geq (\log n)^K/n$, then a.a.s.

$$\chi_g(\mathcal{G}_{n,p}) \leq \alpha\frac{n}{\log_b np}.$$

Thus, for p sufficiently large depending on $\epsilon > 0$, $(2-\epsilon)\chi(\mathcal{G}_{n,p}) < \chi_g(\mathcal{G}_{n,p}) < (4+\epsilon)\chi(\mathcal{G}_{n,p})$. In [22], weaker results are shown for smaller p, still bounding $\chi_g(\mathcal{G}_{n,p})$ within a multiplicative factor of $\chi(\mathcal{G}_{n,p})$.

In this chapter, we consider the game chromatic number of the uniformly chosen random 3-regular graph $\mathcal{G}_{n,3}$. Here, it follows from Brooks's theorem that the chromatic number of $\mathcal{G}_{n,3}$ is 3 a.a.s. We prove that, on the other hand:

Theorem 4.1.1. $\chi_g(\mathcal{G}_{n,3}) = 4$ *a.a.s.*

It is worth discussing how the random 3-regular graph $\mathcal{G}_{n,3}$ is chosen. The usual approach is via the configuration model due to Bollobás [9], which produces each 3-regular graph on n vertices with the same probability. However, Robinson and Wormald [45] prove that a.a.s., a graph chosen in this way will be Hamiltonian, which suggests a different method for sampling a 3-regular graph: starting with a cycle C on n vertices, a random perfect matching is added.

This process does not yield the same distribution. In particular, 3-regular graphs which are not Hamiltonian are sampled with probability 0; moreover, if the perfect matching shares edges with C, the result is actually a 3-regular multigraph. However, the deviation is known to be sufficiently slight that a result which holds a.a.s. in one model will also hold a.a.s. in the other; see [54]. We may therefore prove Theorem 4.1.1 in this model instead.

It is always possible to complete a partial coloring of a 3-regular graph to a 4-coloring; therefor $\chi_g(\mathcal{G}_{n,3}) \leq 4$. We show that $\chi_g(\mathcal{G}_{n,3}) > 3$ a.a.s. by proving that, if the coloring game is played on $\mathcal{G}_{n,3}$ with $k = 3$ colors, the Bob wins a.a.s.

This proof will proceed in two steps. We first describe a kind of subgraph H such that, if $\mathcal{G}_{n,3}$ contains a copy of H, then Bob can win the coloring game with $k = 3$ colors. Second, we show that a.a.s., $\mathcal{G}_{n,3}$ contains such a subgraph H (and, moreover, contains enough such subgraphs that we can choose one sufficiently far from the vertex where Alice makes her first move).

4.2 The winning strategy

We will say that two vertices are *close* if they are connected by a path of length two or less, and that a path is *short* if some vertex on it is close to both endpoints. (This is not the same as being of length at most four). Vertices that are not close are *far apart* and a path that is not short is *long*. The motivation for this terminology is that coloring a vertex can only have an effect on vertices that are close to it; we will make this precise later on.

We first assume the existence of a subgraph H with the following properties (see Figure 4.1):

1. H consists of two vertices, v and w, together with three (internally disjoint) paths from

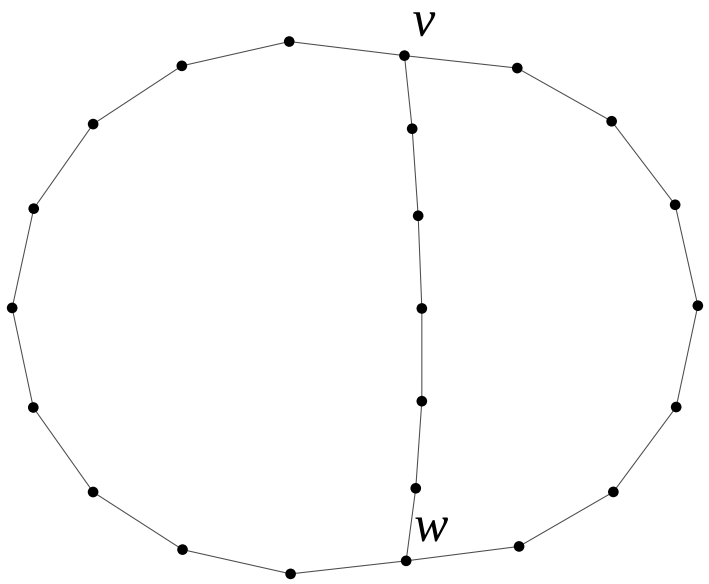


Figure 4.1: The subgraph H required for Bob's strategy.

one to the other.

2. Each of the paths consists of an even number of edges.
3. No two vertices in H are connected by a short path outside of H (in particular, H is induced).
4. The three paths themselves are all long.

In addition,

Property F: if Alice goes first, then the vertex colored by Alice on her first move is far from H.

Bob first plays on the vertex v. Provided Alice's next move is not on the vertex w, or on the neighbors of v or w, it is close to at most one of the three paths which make up H (this follows from Properties 3 and 4). The remaining two paths form a cycle containing v, with no other already colored vertices close to the cycle; by Property 2, the cycle is even. Call the vertices around the cycle $(v, v_1, v_2, \ldots, v_{2k-1})$.

Starting from this even cycle, Bob proceeds as follows. He colors v_2 a different color from v; this creates the threat that on his next move, he will color the third neighbor of v_1 the remaining

color, leaving no way to color v_1 and winning. We will call such a move by B a *forcing move at* v_1. A can counter this threat in several ways:

- By coloring v_1 the only remaining viable color.
- By coloring v_1's third neighbor the same color as either v or v_2.
- By coloring that neighbor's other neighbors in the color different from both v and v_2.

In all cases, Alice must color some vertex close to v_1, that does not lie on the cycle.

Bob continues by making a forcing move at v_3: coloring v_4 a different color from v_2. Continuing to play on the even vertices v_{2i}, Bob makes forcing moves at each odd v_{2i-1}. By Property 3 of H, the set of vertices Alice must play on to counter each threat are disjoint; thus, Alice's response to each forcing move does not affect the rest of the strategy. By Property F, Alice's first play does not affect the strategy either.

When Bob colors v_{2k-2}, this is a forcing move both at v_{2k-3} and at v_{2k-1} (provided Bob chooses a color different both from v_{2k-4} and v). Alice cannot counter both threats, therefore Bob wins.

We now account for the remaining few cases. If Alice colors a neighbor of v or w on her second move, this vertex will be close to all three possible even cycles. However, we know that all three paths in H have even length. Therefore we can still apply this strategy to the even cycle not containing the vertex Alice colored. Even though it will be close to v or w, we will never need to force at v or at w, because we only force at odd numbered vertices along the path.

Finally, if Alice colors w itself, then there is no path we can choose that will avoid the vertex. Instead, Bob picks any of the paths from v to w, and makes forcing moves down that path. Provided that the path is sufficiently long to do so (which follows from Property 4), the final move will be a forcing move in two ways, winning the game for Bob once again.

4.3 Proof of existence

It remains to show that the subgraph H exists a.a.s. (even allowing for Alice's first move). In the following, let c be a constant; we will later see that we need c to be less than 1 for the proof to hold.

We begin by counting *good* segments of length $m = \lfloor c\sqrt{n} \rfloor$ on C, by which we mean those with no internal chords. First of all let X be twice the number of chords that intercept segments of length m or less – these are the only chords that could possibly be internal to a segment of the desired length. X can be written as the sum $X_1 + X_2 + \cdots X_n$, where X_i is the 0-1 indicator for the i-th vertex (call it v_i) to be the endpoint of such a chord. Also, let Y_i denote the length of the smaller of the two segments defined by v_i (this segment stretches from v_i to its partner). Thus

$$\Pr(Y_i = t) = \begin{cases} \frac{2}{n-1} & 2 \le t \le \lfloor (n-1)/2 \rfloor \\ \frac{1}{n-1} & t = n/2, \ n \text{ even} \end{cases}$$

Clearly $X_i = 1$ if and only if $Y_i \le m$, and so

$$\mathrm{E}(X_i) = \frac{2m}{n-1} \text{ and } \mathrm{E}(X) = \frac{2mn}{n-1} \approx 2c\sqrt{n}.$$

In addition, $\mathrm{Var}(X_i) \le \mathrm{E}(X_i)$, and so

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \ne j} \mathrm{Cov}(X_i, X_j) \le \mathrm{E}(X) + \sum_{i \ne j} \mathrm{Cov}(X_i, X_j).$$

Now

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= -\frac{4m^2}{(n-1)^2} + \sum_{t=2}^{m} \Pr(X_i = 1 \mid Y_j = t) \Pr(Y_j = t) \\ &\le -\frac{4m^2}{(n-1)^2} + \frac{2m}{n-3} \cdot \frac{2m}{n-1} \\ &= \frac{8m^2}{(n-1)^2(n-3)}. \end{aligned}$$

Thus,

$$\mathrm{Var}(X) \le \mathrm{E}(X) + \frac{8m^2 n}{(n-1)(n-3)} \approx \mathrm{E}(X).$$

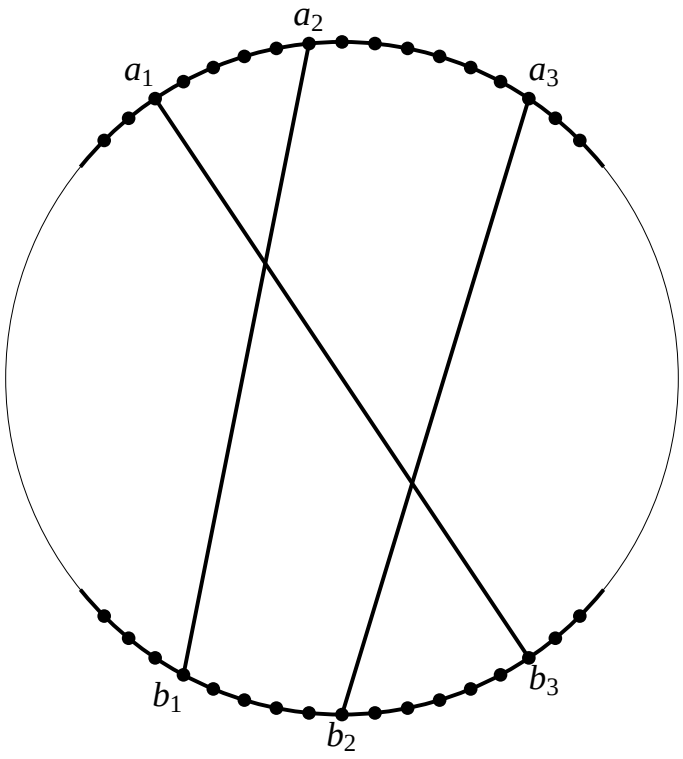


Figure 4.2: A typical example of the subgraph H found in the Hamiltonian cycle model.

By Chebyshev's inequality,

$$\Pr(|X - \mathrm{E}(X)| \le \lambda \,\mathrm{E}(X)) \le \frac{\mathrm{Var}(X)}{\lambda^2 \,\mathrm{E}(X)^2} \le \frac{2}{\lambda^2 c\sqrt{n}}.$$

Putting $\lambda = n^{-1/5}$ we see that a.a.s. $X \sim 2c\sqrt{n}$.

Consider the n different segments of length m on C. Each chord counted by X eliminates at most m of these segments as being good, which leaves $(1 - c^2)n$ segments remaining. We will want non-overlapping good segments; each good segment overlaps at most $2m$ other good segments and so we can assume that we can find $2n_1 \sim (c^{-1} - c)\sqrt{n}/2$ non-overlapping good segments a.a.s. Here the segments are $\sigma_1, \sigma_2, \ldots, \sigma_{2n_1}$ are in clockwise order around C. We pair them together $P_i = (\sigma_i, \sigma_{n_1+i}), i = 1, 2, \ldots, n_1$.

Pick any pair P_j. If there are exactly 3 chords from one segment to the other, as in Figure 4.2, then we will construct H as follows (assuming a_i and b_i are the endpoints of the chords, as labeled in Figure 4.2):

- Set v and w to be a_2 and b_2, respectively.
- The first path from v to w is $(a_2, \ldots, a_1, b_3, \ldots, b_2)$, where the vertices in the ellipses are chosen along C.
- The second path from v to w is $(a_2, b_1, \ldots, b_2)$.
- The third path from v to w is $(a_2, \ldots, a_3, b_2)$.

The paths given above require that the three chords are (a_1, b_3), (a_2, b_1), and (a_3, b_2). In order for H to satisfy Properties 2 and 4, we impose conditions on the lengths of the paths $(a_1, \ldots, a_2)$, $(a_2, \ldots, a_3)$, $(b_1, \ldots, b_2)$, and $(b_2, \ldots, b_3)$: they must not be too small, and must have the right parity so that the three paths from v to w have even length. However, these conditions (and the condition that (a_2, b_2) must not be a chord) eliminate only a constant fraction of the possible chords; therefore there are $\Omega(m^6)$ ways to choose the chords.

The probability, then, that a subgraph H can be found between two given good segments, is at least

$$\Omega\left(\frac{m^6}{n^3}\right) \cdot \left(1 - \frac{m}{n-2m}\right)^{2m}$$

where the last factor bounds the probability that there are no extra chords between the two segments. This tends to a constant ζ_1 that does not depend on n. Thus the expected number of j for which P_j has Properties 1,2 and 4 are satisfied is at least $\zeta_1 n_1$.

We now consider the number of pairs of good segments in which we can hope to find this structure. In order to ensure that, should a subgraph H be found, it satisfies Property F, we eliminate all pairs which contain a vertex close to the vertex Alice chooses on her first move – a constant number of pairs.

To ensure Property 3 we eliminate all pairs P_j in which two vertices have chords whose other endpoints are 1 or 2 edges apart. This happens with probability $O(1/n)$ for any two vertices, regardless of the disposition of the other chords incident with the segments in P_j. The pair P_js contains $\binom{2m}{2} \leq 2c^2 n$ pairs of vertices. Therefore with probability at least $(1 - O(1/n))^{2c^2 n}$, which tends to a constant, ζ_2 say, a pair P_j satisfies Property 3.

Thus the expected number of j for which the pair P_j give rise to a copy of H satisfying all required properties is at least ζn_1 where $\zeta = \zeta_1\zeta_2$. To prove concentration for the number of j we can simply use the Chebyshev inequality. This will work, because exposing the chords incident with a particular pair P_j will only have a small effect on the probability that any other P_j' has the required properties.

Chapter 5

Increasing Hamiltonian Paths in Random Edge Orderings

5.1 Background

The classical result of Erdős and Szekeres [18] states that any permutation of $\{1, 2, \ldots, n^2 + 1\}$ contains a monotonic subsequence of length $n + 1$. Many extensions have been found for this theorem: see, e.g., any of [21, 34, 43, 47, 48]. In this chapter, we consider the direction started by Chvátal and Komlós [14]. They posed the natural analogue of the problem for walks in a graph, which may be considered an extension of the Erdős–Szekeres result in a similar spirit to how Ramsey's theorem is an extension of the pigeonhole principle. Rather than ordering the integers $\{1, 2, \ldots, n\}$, order the edges of K_n by setting a bijection $f : E(K_n) \to \{1, 2, \ldots, \binom{n}{2}\}$. A walk in K_n whose edges are $(e_1, e_2, \ldots, e_k)$ is called f-increasing if the labels $f(e_1), f(e_2), \ldots, f(e_k)$ form an increasing sequence. (In this setting, we can assume that the monotone sequence of labels is increasing without loss of generality, since a decreasing walk is just an increasing walk traversed backwards.) As in the Erdős–Szekeres result, the objective is to prove a worst-case lower bound on the length of the longest increasing walk. Here, a walk is permitted to visit the same vertex multiple times.

This question was resolved by Graham and Kleitman [27]. In [52], Winkler communicates an elegant formulation of their solution, which is due to Friedgut: a pedestrian stands at every vertex of K_n, and the edges are called out in increasing order; whenever an edge is called out, the pedestrians on its endpoints switch places. After all edges have been called out, the n pedestrians have taken a total of $n(n-1)$ steps, and therefore at least one must have taken at least $n-1$ steps, producing an increasing walk with length (number of edges) at least $n-1$.

This is easily seen to be tight for even n, for which K_n can be partitioned into $n-1$ perfect matchings, and edges within each individual matching can receive consecutive labels. For odd n, a partition into n maximal matchings only gives an upper bound of n; a more complicated argument in [27] shows that $n-1$ is still correct for all n except $n=3$ and $n=5$, where n is the right answer.

Chvátal and Komlós also posed the corresponding problem for self-avoiding walks, or paths, which are not permitted to revisit any vertex. Self-avoiding walks are generally much harder to analyze, and indeed, in this setting, even determining the answer asymptotically is still an open question. Calderbank, Chung, and Sturtevant [11] construct an ordering of K_n for which no increasing path is longer than $\left(\frac{1}{2}+o(1)\right)n$. For a long time, the best known lower bound was the first bound proven by Graham and Kleitman in [27], where they show that there must always be an increasing path of length $\sqrt{n-1}$. A recent result of Milans [42] improves this bound to length $O(n/\log n)^{2/3}$.

Many extremal questions for combinatorial structures have also been studied in the random setting (see, e.g., either of the books [8, 33] on random graphs), which is in a sense equivalent to asking about the average-case rather than the worst-case behavior of some property. For example, the random analogue of the Erdős–Szekeres result considers the length I_n of the longest increasing subsequence in a random permutation of $\{1, 2, \ldots, n\}$, and this is a well-studied topic: it is known [40, 50] that $I_n \sim 2\sqrt{n}$ a.a.s. (Here and in the remainder, we write $X \sim Y$ to denote $\lim_{n\to\infty} \frac{X}{Y} = 1$. To avoid two separate limits as $n \to \infty$, we define $X \sim Y$ a.a.s. to mean that

for all $\epsilon > 0$, $\lim_{n\to\infty} \Pr[\left|\frac{X}{Y} - 1\right| < \epsilon] = 1$.)

In this chapter, we consider the random version of the increasing path problem. Suppose the ordering f is chosen uniformly at random. What can we say about the length of the longest f-increasing path? It is natural to begin by considering the performance of the greedy algorithm on a randomly ordered graph, since all walks traced by pedestrians in the above argument are greedy in the following sense: every walk exits each vertex along the minimally-labeled edge which maintains the increasing property.

Proposition 5.1.1. *Let v_0 be an arbitrary vertex in K_n. Given an edge ordering f of the edges of K_n, let the greedy f-increasing path from v_0 be the path $v_0v_1v_2\dots v_t$ with the following properties: (i) v_0v_1 is the lowest-labeled edge incident to v_0, (ii) for each $1 \le i \le t-1$, v_{i+1} is the vertex x which minimizes the label of v_ix over all $x \notin \{v_0, v_1, \dots, v_i\}$ with v_ix exceeding the label of $v_{i-1}v_i$, and (iii) every vertex $x \notin \{v_0, \dots, v_t\}$ has v_tx labeled below $v_{t-1}v_t$. Then, if f is chosen uniformly at random, the length of the greedy f-increasing path from v_0 is $(1-\frac{1}{e}+o(1))n$ a.a.s.*

Since the analysis in this result is tight, one must consider more complex algorithms in order to find longer paths in the random setting. The main challenge in analyzing more sophisticated algorithms arises from the fact that randomness is revealed during the algorithm's execution. We introduce a novel extension of the greedy algorithm which adds some foresight, but which is formulated in a way that is amenable to analysis. At each step, this *k-greedy algorithm* greedily finds a tree of k potential edges that can extend the increasing path, before choosing the one that has the best short-term prospects. (The ordinary greedy algorithm is the $k = 1$ case of this algorithm.) A detailed description of this algorithm appears in Section 5.3.1.

The performance of the k-greedy algorithm is related to statistics which arise in the Chinese Restaurant Process, or equivalently, to the random variable L_k which measures the length of the longest cycle in a uniformly random permutation of $\{1, \dots, k\}$. Let $\alpha_k = \mathrm{E}[\frac{1}{L_k} + \frac{1}{L_k+1} + \cdots + \frac{1}{k}]$.

Theorem 5.1.1. *If an edge ordering f of K_n is chosen uniformly at random, then the k-greedy*

algorithm finds an f-increasing path of length $(1 - e^{-1/\alpha_k} + o(1))n$ a.a.s.

Remark. The constant α_k is monotone decreasing in k and explicitly computable. The particular choice $k = 100$ produces an increasing path of length $0.85n$ a.a.s. As $k \to \infty$, this monotonicity implies that α_k converges to $\alpha = \lim_{k\to\infty} \mathrm{E}\left[-\log \frac{L_k}{k}\right]$, a constant related to the Golomb–Dickman constant $\lim_{k\to\infty} \mathrm{E}\left[\frac{L_k}{k}\right] \approx 0.6243$. Numerically, we estimate $\alpha \approx 0.5219$, and $1 - e^{-1/0.5219} \approx 0.853$, so $k = 100$ appears to be a near-optimal choice.

The first two results establish successively stronger linear lower bounds on the increasing path length in a random edge ordering. There is a trivial upper bound of $n - 1$: the length of a Hamiltonian path, and at first glance, one may assume that an increasing Hamiltonian path would be too much to hope for. Indeed, when one calculates the expected number of f-increasing Hamiltonian paths, the total number of paths, which is $n!$, is almost exactly canceled by the probability that each is increasing, which is $\frac{1}{(n-1)!}$. Thus, the expected number of increasing Hamiltonian paths is only n, which grows to infinity relatively slowly. In comparison, in the Erdős–Rényi random graph model $G_{n,p}$, where each edge appears independently with probability p, the expected number of Hamiltonian paths is about n when $p \sim \frac{e}{n}$, but Hamiltonian paths don't appear until $p \sim \frac{\log n}{n}$. Furthermore, for Hamiltonian cycles in the random graph process, at the moment Hamiltonicity is achieved, the number of Hamiltonian cycles jumps from 0 to $\left[(1+o(1))\frac{\log n}{e}\right]^n$, as shown recently by Glebov and Krivelevich [25] (improving an earlier result of Cooper and Frieze [15]). It may therefore come as a surprise that random edge orderings often have Hamiltonian paths, despite the extremely low expected value.

Theorem 5.1.2. *If an edge ordering f of K_n is chosen uniformly at random, then an f-increasing Hamiltonian path exists in K_n with probability at least $\frac{1}{e} + o(1)$.*

The proof of Theorem 5.1.2 uses the second moment method. Let H_n be the number of increasing Hamiltonian paths. As mentioned above, it is easy to see that $\mathrm{E}[H_n] = n! \cdot \frac{1}{(n-1)!} = n$. The main step of our proof is to upper-bound the second moment $\mathrm{E}[H_n^2]$. We actually go one step further, and asymptotically determine $\mathrm{E}[H_n^2] = (1 + o(1))en^2$, from which the result follows.

Throughout this chapter, we make use of several well-known probabilistic results, which all appear, e.g., in [1]. In Section 5.2, we use the following corollary of Chebyshev's inequality: for a random variable X with $\mathrm{Var}[X] = o(\mathrm{E}[X]^2)$, $X \sim \mathrm{E}[X]$ a.a.s. In Section 5.3.1, we rely on the Azuma–Hoeffding inequality for martingales, which can be stated as the bound

$$\Pr[|X_n| > \lambda] < 2\exp\left(-\frac{\lambda^2}{2L^2 n}\right)$$

for a martingale (X_t) with $X_0 = 0$ and $|X_{i+1} - X_i| \leq L$ for all $0 \leq i < t$. Moreover, if τ is a stopping time for (X_t) that always satisfies $\tau \leq n$, $\Pr[|X_\tau| > \lambda]$ satisfies the same inequality. Finally, Theorem 5.1.2 will follow from the calculation in Section 5.4 by use of the second moment inequality $\Pr[X > 0] \geq \frac{\mathrm{E}[X]^2}{\mathrm{E}[X^2]}$, valid for a nonnegative random variable X.

5.1.1 Recent work

Theorems 5.1.1 and 5.1.2 complement each other, as they establish a.a.s. almost-Hamiltonicity and almost-a.a.s. Hamiltonicity. Numerical simulations led us to conjecture a stronger result:

Conjecture 5.1.1. *If an edge ordering f of K_n is chosen uniformly at random, then an f-increasing Hamiltonian path exists in K_n a.a.s.*

This conjecture was recently proven by Martinsson in [41] in a quantitative form, by extending our second moment calculations in Section 5.4 to a third moment analysis of H_n (the number of increasing Hamiltonian paths) and related quantities.

Theorem 5.1.3 (Theorem 1.2 in [41])**.** *If an edge ordering f of K_n is chosen uniformly at random, then a.a.s. as $n \to \infty$ we have $H_n > 0$. Moreover, for any $x > 0$,*

$$\limsup_{n\to\infty} \Pr[H_n \leq x\, n] \leq e \cdot x.$$

5.2 The length of the greedy increasing path

As a warm-up for the k-greedy algorithm, we begin by proving Proposition 5.1.1, which establishes that the greedy algorithm produces an increasing path of linear length a.a.s. In this section,

it is convenient to introduce a different (but equivalent) model for generating the edge labels, which features more independence. In order to sample a uniform permutation of $\{1, 2, \ldots, \binom{n}{2}\}$ for the labels, we choose a labeling $f : E(K_n) \to [0, 1]$, where the labels $f(e)$ are i.i.d. $\mathrm{Uniform}(0, 1)$ random variables. Since with probability 1 no two labels will be equal, this induces a total ordering on the edges, and by symmetry, all orderings occur with the same probability.

Let $(e_1, e_2, \ldots, e_k)$ be the edges of any path, not necessarily f-increasing. We define the *jumps* $X_1, \ldots, X_k$ along this path by $X_1 = f(e_1)$, and

$$X_k = (f(e_k) - f(e_{k-1})) \bmod 1 = \begin{cases} f(e_k) - f(e_{k-1}), & \text{if } f(e_k) > f(e_{k-1}) \\ 1 + f(e_k) - f(e_{k-1}), & \text{if } f(e_k) \leq f(e_{k-1}). \end{cases}$$

The sum $X_1 + X_2 + \cdots + X_k$ telescopes to $f(e_k) + p$, where p is the number of points at which the path fails to be f-increasing. Therefore the path is f-increasing if and only if $X_1 + X_2 + \cdots + X_k \leq 1$.

Choose a Hamiltonian path $(e_1, e_2, \ldots, e_{n-1})$ by the following rule: starting from an arbitrary vertex, always take the edge with the smallest jump. The result coincides with the greedy path for the entire length of the greedy path. However, when the greedy path would stop, this rule merely makes a step that isn't f-increasing. This allows us to keep going until $n - 1$ edges are chosen, no matter what. The length of the greedy increasing path will therefore be the largest k for which the initial segment $(e_1, e_2, \ldots, e_k)$ forms an f-increasing path; equivalently, the largest k such that $X_1 + X_2 + \cdots + X_k \leq 1$.

When constructing this path, we only expose the labels of the edges as we encounter them. Specifically, if we already have the partial Hamiltonian path $(e_1, \ldots, e_{k-1})$, then the next edge e_k will be one of the $n - k$ edges from the last vertex to a new vertex not on the current path. Call those possible edges $e_k^1, \ldots, e_k^{n-k}$, and expose $f(e_k^1), \ldots, f(e_k^{n-k})$. The jump X_k is given by

$$X_k = \min\left\{(f(e_k^1) - f(e_{k-1})) \bmod 1, \ldots, (f(e_k^{n-k}) - f(e_{k-1})) \bmod 1\right\}.$$

The values $(f(e_k^j) - f(e_{k-1})) \bmod 1$ are also uniformly distributed on [0,1], and exposing them is equivalent to exposing $f(e_k^1), \ldots, f(e_k^{n-k})$. This means that X_k is the minimum of $n-k$ uniform random variables, which are independent from each other and from all previously exposed values.

Since $\Pr[X_k \geq x] = (1-x)^{n-k}$, the probability density function is $(n-k)(1-x)^{n-k-1}$, so:

$$\mathrm{E}[X_k] = \int_0^1 x(n-k)(1-x)^{n-k-1}\,dx = \frac{1}{n-k+1},$$
$$\mathrm{E}[X_k^2] = \int_0^1 x^2(n-k)(1-x)^{n-k-1}\,dx = \frac{2}{(n-k+1)(n-k+2)}.$$

Suppose $t = \tau n$ for some constant $0 < \tau < 1$, and let $S_t = X_1 + X_2 + \cdots + X_t$. Then $\mathrm{E}[S_t] = \frac{1}{n} + \cdots + \frac{1}{n-t+1}$, and therefore $\mathrm{E}[S_t] = \log \frac{n}{n-t} + O(n^{-1}) = \log \frac{1}{1-\tau} + O(n^{-1})$. Furthermore, we have $\mathrm{Var}[X_k] = O(n^{-2})$ for $1 \leq k \leq t$, so (by the independence of the X_k) $\mathrm{Var}[S_t] = O(n^{-1})$, which means that $S_t = (1+o(1))\,\mathrm{E}[S_t]$ a.a.s. When $\tau < 1 - \frac{1}{e}$, $\mathrm{E}[S_t] < 1$, so $S_t < 1$ a.a.s., and the greedy path is f-increasing through the first t steps. On the other hand, when $\tau > 1 - \frac{1}{e}$, $\mathrm{E}[S_t] > 1$, so $S_t > 1$ a.a.s., and the greedy path is not f-increasing after the first t steps. This completes the proof of Proposition 5.1.1. □

5.3 The k-greedy algorithm

Throughout this section, let k be a constant. The k-greedy algorithm extends the greedy algorithm by adding some limited look-ahead to the choice of each edge. We analyze it using the same model as in the previous section: each edge receives an independent random label from $\mathrm{Uniform}(0,1)$. For the purposes of intuition, we think of the label $f(e)$ of an edge e as the time at which e appears in the graph. We first describe how the algorithm would run, given a (deterministic) full labeling of the edges of K_n with real numbers from $[0,1]$.

The challenge with any complex algorithm is dependency between iterations. Our main innovation is to distill the algorithm and pose it in a way that is particularly clean and amenable to analysis.

5.3.1 Algorithm k-greedy

1. Initialize the path P to be a single (arbitrary) vertex v_1. Initialize the rooted tree T to be the 1-vertex tree with v_1 as the root. Initialize the time t to be 0.
2. While the rooted tree T has fewer than k edges, do:
   (a) Let S be the set of all edges with one endpoint in T, the other endpoint not in $P \cup T$, and label at least t.
   (b) If S is empty, then set $t = 1$, and terminate the algorithm.
   (c) Identify the edge of S with minimum label, add it to T, and set t to be its label.
3. The rooted tree T now has exactly k edges. Among the children of the root, identify the child x whose subtree (rooted at x) is the largest. Extend P by one edge to x, and set T to be the subtree rooted at x. This may substantially reduce the size of T, as it deletes all of the other subtrees, as well as the root.
4. Go back to step (2).

Lemma 5.3.1. *The path P produced by the k-greedy algorithm is always a simple increasing path.*

Proof. Consider any moment at which an edge e is added to T. Suppose that $e = xy$, where x was previously in T, and y is a new leaf of T. By construction, the label of e is at least t, but the label of every other edge in $P \cup T$ is at most t (and a.s. not equal to t). So, by induction, at all times during the algorithm, all paths from the first vertex $v_1 \in P$ to any leaf of T are increasing paths. They are all simple paths because we only extend T by edges to vertices not currently in $P \cup T$. □

5.3.2 Managing revelation of randomness

We now take a closer look at what information needs to be revealed at each step in order to run the algorithm. We find that Step (2b) requires a yes/no answer, and Step (2c) requires the identification of a single edge, together with its label. Therefore, if we have access to an oracle which reports this information upon request, we will be able to run the complete algorithm.

The information revelation in Step (2b) is a minor issue, which we can easily sidestep. If S is nonempty, the revelation of a single step from S in Step (2c) will screen off that information anyway. We will not have to worry about the after-effects of the revelation that S is empty, because at that point the algorithm halts.

We will carefully describe how we manage the exposures in Step (2c). At the beginning, the labels on all edges are independent, and each is uniformly distributed in $[0, 1]$. Consider the exposure the first time Step (2c) is encountered. The oracle reports an edge v_1x and its label, and so at this point, the label of v_1x is certainly determined. Let us refer to it as t_1. We also learn some information about all other edges v_1y, with $y \notin \{v_1, x\}$: their labels are not in the range $[0, t_1)$. We do not learn any restrictions on any the labels of any other edges. Importantly, the labels on all edges are still independent, and uniformly distributed over their (possibly-restricted) ranges. They are just not identically distributed.

Consider the second time the algorithm encounters Step (2c). Now, the oracle reports another edge, say xz, and suppose its label is t_2. Then, we know that all edges v_1y and xy with $y \notin \{v_1, x, z\}$ have labels outside of the interval $(t_1, t_2]$. We already learned that some of those edges had labels outside $[0, t_1)$; for those, we now know that their labels avoid $[0, t_2)$. Again, all labels are still independent and uniformly distributed over their ranges.

So, at every intermediate time t during the course of the algorithm, some edges will have their labels determined, but independence between non-determined labels is preserved throughout. Each non-determined label is still uniformly distributed over some range of the form $[0, 1] \setminus (I_1 \cup I_2 \cup \cdots \cup I_t)$, where the I_i are disjoint sub-intervals. Note that all ranges still completely include

$[t, 1]$, and so this phenomenon can only increase the likelihood that we can still use the edge: Step (2a) queries only edges with labels at least t. Since all non-determined edges have label ranges including $[t, 1]$, they satisfy this property with probability exactly $(1 - t)/\mu$, where μ is the measure of their current range. This is worst when $\mu = 1$, which corresponds to the fully unrestricted $[0, 1]$ range.

5.3.3 Intuitive calculation

Now that we have a clean model which definitively indicates how much randomness is surrendered at each step, we conduct a rough analysis which captures the main structure of the argument. This will also derive the constant in Theorem 5.1.1. The key statistic to estimate is the typical *waiting time*, which we define to be the difference between the labels of successive edges added to T by Step (2c) of the algorithm.

Suppose that at some stage of the algorithm, the path has length ℓ, and the tree T has j vertices. If all of the $j(n - \ell - j)$ edges between T and vertices outside $P \cup T$ still had labels which were uniformly distributed over $[0, 1]$, then the waiting time would be the minimum of that many random variables $\mathrm{Uniform}(0, 1)$. The waiting time would then be exactly $\frac{1}{j(n-\ell-j)+1}$ in expectation. In our situation, this is not exactly true. We still have independence, but the labels are distributed uniformly over sub-ranges of $[0, 1]$. Also, some of those $j(n - \ell - j)$ edges could potentially already have their labels determined, if they had previously been added as edges to T in an earlier stage of the algorithm, but were discarded by some Step (3). As mentioned at the end of Section 5.3.2, the first issue is in our favor, because it only reduces the waiting time. The second issue works against us, because it reduces the number of independent random labels that are competing in Step (2c), but we will show in Section 5.3.5 that in fact both of these effects are negligible. So, we first analyze the (fictitious) ideal case.

For now, let us proceed using $\frac{1}{j(n-\ell-j)+1} \sim \frac{1}{n-\ell} \cdot \frac{1}{j}$ as the expected waiting time. Then, the

expected time until the search tree fills up from $j-1$ to k edges is asymptotically

$$\frac{1}{n-\ell} \cdot \left(\frac{1}{j} + \frac{1}{j+1} + \cdots + \frac{1}{k} \right) \sim \frac{1}{n-\ell} \log \frac{k}{j}.$$

At this point, the increasing path is extended by 1, and the search tree shrinks to the largest subtree determined by a child of the root. From the formula above, we see that only the size of this subtree affects the next waiting time.

In our ideal setting, when a potential edge is added to a search tree with j vertices, its endpoint in the search tree is randomly distributed uniformly over all j vertices currently in the tree. From the point of view of subtree sizes among children of the root, it therefore starts a new subtree with probability $\frac{1}{j}$ (if its tree-endpoint is the root itself), or is added to one of the existing children's subtrees with probability proportional to current size (depending on which subtree its tree-endpoint is in). This is equivalent to the Chinese restaurant process, which generates the cycle decomposition of a uniformly random permutation: if π is a uniformly random permutation of $\{1, \ldots, j-1\}$, then we can transform π into a uniformly random permutation of $\{1, \ldots, j\}$ by making j a fixed point with probability $\frac{1}{j}$, and otherwise inserting j in a uniformly chosen point in any cycle (which means a cycle of length i is chosen with probability $\frac{i}{j}$). Therefore the number of vertices in the largest subtree has the same distribution as a well-studied random variable: the length L_k of the longest cycle in a uniformly random permutation of $\{1, \ldots, k\}$.

Define the random variable $A_k = \frac{1}{L_k} + \frac{1}{L_k+1} + \cdots + \frac{1}{k} \sim -\log \frac{L_k}{k}$, and define the constant $\alpha_k = \mathrm{E}[A_k]$. Let X_ℓ be the waiting time for the increasing path to grow from length $\ell - 1$ to length ℓ. From what we have shown, $\mathrm{E}[X_\ell] \sim \frac{\alpha_k}{n-\ell}$. As in the analysis of the greedy algorithm, we expect it to be true that a.a.s., the current time $t = X_1 + \cdots + X_\ell$ satisfies

$$\sum_{i=1}^{\ell} X_i \sim \mathrm{E}\left[\sum_{i=1}^{\ell} X_i \right].$$

We determine the length of the path by finding the point at which this expected value reaches 1:

$$1 = \sum_{i=1}^{\ell} \mathrm{E}[X_i] \sim \alpha_k \left(\frac{1}{n} + \cdots + \frac{1}{n-\ell} \right) \sim \alpha_k \log \frac{n}{n-\ell}.$$

Therefore the algorithm typically achieves a length ℓ such that $\frac{\ell}{n} \sim 1 - e^{-1/\alpha_k}$.

To make this argument rigorous, we will have to deal with two main issues. First, we must show that even when we deviate from the ideal setting, the expectations $\mathrm{E}[X_i]$ remain mostly the same. Second, we need to show that the sum $\sum_{i=1}^{\ell} X_i$ does not deviate too much from its expected value.

5.3.4 Determination of constant

In order to determine the numerical bounds in Theorem 5.1.1, we must understand α_k. In [26], a recurrence relation is given for the number of permutations of $\{1, \ldots, n\}$ with greatest cycle length s. Using our notation, we present a modified version of this recurrence: if L_n is the length of the longest cycle in a random permutation of $\{1, 2, \ldots, n\}$, then for $1 \leq s \leq n$,

$$\Pr[L_n = s] = \sum_{j=1}^{\lfloor n/s \rfloor} \frac{1}{j!\, s^j} \Pr[L_{n-sj} \leq s - 1],$$

where L_0 is the constant 0 whenever it occurs. This recurrence allows for an exact numerical computation of $\alpha_k = \mathrm{E}[A_k]$ for any k. Several seconds of computation are enough to confirm that $\alpha_{100} < 0.523$, which implies that the 100-greedy algorithm typically finds an increasing path of length at least cn, where $c > 1 - e^{-1/0.523} > 0.85$.

It is natural to wonder whether a particular finite choice of k would be optimal for the k-greedy algorithm. Using a careful coupling argument, we can show that α_k is monotone decreasing with respect to k: if we consider L_k and L_{k+1} as stages in the same Chinese restaurant process, we have

$$\mathrm{E}[A_{k+1} - A_k \mid L_k] \leq \frac{1}{k+1} - \frac{1}{L_k} \cdot \frac{L_k}{k+1} = 0$$

since with probability at least $\frac{L_k}{k+1}$, the longest cycle increases in length. Therefore as $k \to \infty$, α_k approaches some constant α.

Therefore, no finite k is optimal. Since the Golomb–Dickman constant $\lim_{k\to\infty} \mathrm{E}[\frac{L_k}{k}] \approx 0.6243$ has no closed form, we expect the same to be true for $\alpha = \lim_{k\to\infty} \mathrm{E}[-\log \frac{L_k}{k}]$. Our

numerical methods estimate $\alpha \approx 0.5219$, so our choice of $k = 100$ already achieves a bound which is close to optimal for k-greedy algorithms.

5.3.5 Rigorous analysis

There are two obstacles in the way of the uniformity we assumed for this analysis. On one hand, we may expose potential edges, add them to the search tree, but fail to use them (deleting them from the search tree as we pass to a subtree), and then encounter these edges again, which increases the waiting time because these edges can't be added to the search tree. On the other hand, when a minimal edge is found, we learn that all other edges we consider are outside some range $[t_1, t_2]$. We describe this as gaining a *partial exposure* of $t_2 - t_1$ for those edges. When edges with partial exposure are considered a second time, the waiting time decreases. In this section, we show that both of these obstacles are asymptotically irrelevant. In our discussion, an *exposed* edge is one whose label has been completely determined. Even if an edge has gained partial exposure, we still call it *unexposed*.

At any point in the k-greedy algorithm's execution, we say that the algorithm is *well-behaved* if it satisfies the following two conditions:

1. No vertex of the graph has entered the search tree more than $C_1 = \log n$ times.
2. In each iteration of Step (2c) (including the final one, if the algorithm has terminated), t increased by at most $C_2 = \frac{20 \log n}{n}$.

Lemma 5.3.2. *While the algorithm is well-behaved, the following additional conditions also hold:*

3. *The partial exposure on each edge is at most $2k^2 C_1 C_2 = o(1)$.*
4. *No vertex is incident to more than $k^2 C_1 = o(n)$ exposed edges.*

Proof. An edge vw gains partial exposure only when v or w is in the search tree; by condition (1), this happens at most $2C_1$ times total for either v or w.

Suppose v has entered the search tree. After at most k iterations of Step (2c), the search tree reaches k edges, and is then replaced by a subtree. This means v either leaves the search tree or is now one step closer to the new root. Initially, v is at most k steps away from the root; thus, after at most k^2 iterations of Step (2c), v must leave the search tree. The same is true for w.

Therefore there are at most $2k^2C_1$ iterations of Step (2c) when v or w are in the search tree. On each of these, the edge vw can gain at most C_2 partial exposure, by condition (2). Therefore vw cannot have accumulated partial exposure greater than $2k^2C_1C_2$.

Furthermore, edges incident to a vertex v become exposed only while v is in the search tree, or as it enters the search tree. By the same argument as above, there are at most k^2 iterations of Step (2c) for every time v enters the search tree, and so v can gain at most k^2 exposed edges each time, for a total of k^2C_1. □

Lemma 5.3.3. *If the algorithm has been well-behaved up until an iteration of Step (2c), each unexposed edge with one endpoint in the search tree T and one endpoint outside $P \cup T$ has, up to a factor of $1 + o(1)$, an equal chance of being added to T.*

Proof. Let e_1 and e_2 be two such edges. Taking partial exposure into account, suppose that for $i \in \{1, 2\}$ the label of e_i is uniformly distributed on $S_i \subseteq [0, 1]$, with $[t, 1] \subseteq S_i$; let A be the event that the labels of e_1 and e_2 are both contained in $S_1 \cap S_2$. Conditioned on A, the labels of e_1 and e_2 are identically distributed, so $\Pr[e_1 \text{ is added to } T \mid A] = \Pr[e_2 \text{ is added to } T \mid A]$.

However, by condition (3), the partial exposures on both edges are each $o(1)$ so A holds a.a.s. Therefore without conditioning on A the probabilities are also equal up to a factor of $1+o(1)$. □

We expect the k-greedy algorithm to remain a.a.s. well-behaved until it terminates, but we will not be able to prove this until the end of this section. For now, we prove that the algorithm is well-behaved until it terminates, or until the path becomes sufficiently long (which we expect not to happen except with probability $o(1)$).

Lemma 5.3.4. *The algorithm is a.a.s. well-behaved for as long as $\ell = |P| < \frac{9}{10}n$ (or until it terminates).*

Proof. We begin by defining an auxiliary process which runs in parallel to the k-greedy algorithm. This process maintains all the same data as the k-greedy algorithm, and updates it by the following rule:

- If the algorithm was still well-behaved before the current iteration of Step (2c), the auxiliary process copies the k-greedy algorithm.
- Otherwise, the auxiliary process does nothing (in particular, the time t does not change).

The auxiliary process is well-behaved (that is, it satisfies conditions (1) and (2)) if and only if the original algorithm is, because the two only deviate after one of the conditions fails. So it suffices to show that while $\ell < \frac{9}{10}n$, the auxiliary process is a.a.s. well-behaved.

Fix a vertex v. We show that while $\ell < \frac{9}{10}n$, on each iteration of Step (2c), v enters the search tree with probability at most $\frac{20}{n}$, regardless of its previous history. Suppose the process was well-behaved on previous iterations, and v is not in the search tree; if these do not both hold, then the probability that v enters the search tree is 0, at least for the auxiliary process.

By Lemma 5.3.3, each of the unexposed edges from T to v is chosen asymptotically uniformly. Each vertex of T has at most one unexposed edge to v, but (by condition 4) at least $n - \ell - |T| - o(n) > (1 + o(1))\frac{n}{10}$ unexposed edges to vertices outside $P \cup T$, so the probability that an edge to v is chosen is at most $(1 + o(1))\frac{10}{n} < \frac{20}{n}$.

There are at most kn iterations (actually, at most $\frac{9}{10}kn$) on which v could enter the search tree. The probability that v enters the search tree on some C_1 of them is at most

$$\binom{nk}{C_1}\left(\frac{20}{n}\right)^{C_1} \leq \left(\frac{20ek}{C_1}\right)^{C_1}.$$

(Though the probabilities are not independent, the value $\frac{20}{n}$ is an upper bound when conditioned on any previous history, so the multiplication is still valid.)

Because $C_1 = \log n$, for large n we have $\frac{20ek}{C_1} < e^{-2}$, and so this probability is less than n^{-2}. By a union bound over all vertices v, the probability is $o(1)$ that the auxiliary process ever violates condition (1).

Second, we prove that while $\ell < \frac{9}{10}n$, on each iteration of Step (2c), the auxiliary process increases t by at most C_2. It's enough to show that on a single iteration, this holds with probability $1 - o(n^{-1})$, because there are at most $O(n)$ iterations.

Fix a single iteration; we may assume the process was well-behaved up until now, because otherwise t is not going to increase at all. Choose a vertex $v \in T$; by condition (4), v has $n - \ell - |T| - o(n) > (1 + o(1))\frac{n}{10}$ unexposed edges to vertices outside $P \cup T$. This is also a lower bound for the total number of edges from T to the complement of $P \cup T$, which may be larger if there are multiple vertices in T. If t increases to a new value t', whether it's because a new edge with label t' has been found, or because the algorithm terminated (and $t' = 1$), we learn that none of these edges have a label in the interval $[t, t')$. In particular, if $t' - t > C_2$, none of these edges can have labels in the range $[t, t + C_2]$. The probability that this happens for a single edge is at most $1 - C_2$; this only decreases if the edge has any partial exposure. Therefore t increases by more than C_2 with probability at most

$$(1 - C_2)^{(1+o(1))\frac{n}{10}} \leq \exp\left(-(1 + o(1))\frac{n}{10} \cdot \frac{20 \log n}{n}\right) = n^{-2+o(1)}.$$

By the union bound, the probability is $o(1)$ that t increases by more than C_2 on any of the $O(n)$ iterations, and therefore the auxiliary process satisfies condition (2) a.a.s. □

Let $\mathcal{F}_\ell$ be the σ-algebra generated by the information revealed at the time the path reaches length ℓ, and recall that X_ℓ is the waiting time for the increasing path to grow from length $\ell - 1$ to ℓ, so that X_ℓ is $\mathcal{F}_\ell$-measurable. Define the stopping time τ to be the lesser of $\frac{9}{10}n$, or the first ℓ for which the algorithm is no longer well-behaved, or the length of the path when the algorithm terminates. Our first goal is to show that for all $\epsilon > 0$, a.a.s.

$$\left|\sum_{\ell=1}^{\tau} X_\ell - \alpha_k \log \frac{n}{n - \tau}\right| \leq \epsilon.$$

Define the martingale (Z_ℓ) as follows. Let $Z_0 = 0$, and for each $\ell < \tau$, let $Z_{\ell+1} = Z_\ell + X_{\ell+1} - \mathrm{E}[X_{\ell+1} \mid \mathcal{F}_\ell]$. For each $\ell \geq \tau$, let $Z_{\ell+1} = Z_\ell$.

We next study the successive martingale differences $Z_{\ell+1} - Z_\ell = X_{\ell+1} - \mathrm{E}[X_{\ell+1} \mid \mathcal{F}_\ell]$. For this, it is helpful to identify that the most critical information from $\mathcal{F}_\ell$ is the number of vertices in the search tree at the time the path reaches length ℓ.

Lemma 5.3.5. *Suppose that $\ell < \tau$ (our stopping time). Let $A_{\ell,s}$ be the event that at the time the path reaches length ℓ, the search tree contains s vertices. Then*

$$\mathrm{E}[X_{\ell+1} \mid \mathcal{F}_\ell, A_{\ell,s}] \sim \frac{1}{n-\ell} \sum_{i=s}^{k} \frac{1}{i}.$$

Proof. It suffices to show that if the search tree currently contains i vertices, then the expected waiting time until the search tree contains $i+1$ vertices is $(1+o(1))\frac{1}{i(n-\ell)}$. To this end, recall that in the ideal case, there are $i(n-\ell-i)$ edges to choose from, each associated with a $\mathrm{Uniform}(0,1)$ waiting time, and the minimum of $i(n-\ell-i)$ waiting times has expected value $\frac{1}{i(n-\ell-i)+1}$. This is $(1+o(1))\frac{1}{i(n-\ell)}$ since $\ell < \tau$ implies that $n-\ell$ is still linear in n, while $i \le k$ is constant.

In reality, the edge labels are not $\mathrm{Uniform}(0,1)$. However, since $\ell < \tau$, by condition (3), any edge we look at has $o(1)$ total partial exposure. Therefore its associated waiting time is uniform on an interval of length $1 - o(1)$; furthermore, by condition (4), the vertices of the search tree are incident to $n - \ell - i - o(n)$ unexposed edges. As in the ideal case, the minimum of $i(n-\ell-i-o(n))$ such random variables has expected value $(1+o(1))\frac{1}{i(n-\ell)}$, and summing over i as the search tree grows from s to $k+1$ vertices, we establish the lemma. □

For $\ell < \tau$, condition (2) bounds the time to add an edge to the search tree by $\frac{20\log n}{n}$. Since X_ℓ is the sum of at most k such waiting times, $X_\ell < \frac{20k\log n}{n}$, and therefore $Z_{\ell+1} - Z_\ell$ is bounded from above by $L = \frac{20k\log n}{n}$. But $Z_{\ell+1} - Z_\ell$ is at least $-\mathrm{E}[X_{\ell+1} \mid \mathcal{F}_\ell]$, which is always $\Theta(\frac{1}{n})$ due to Lemma 5.3.5. Therefore (Z_ℓ) is Lipschitz with $|Z_{\ell+1} - Z_\ell| \le L$ for all ℓ. Hence by the Azuma–Hoeffding inequality, we have $|Z_\tau| > L\sqrt{2n\log n}$ with probability at most $\frac{2}{n}$. Otherwise, $|Z_\tau| \le L\sqrt{2n\log n} = O(\frac{(\log n)^{3/2}}{\sqrt{n}})$, which is $o(1)$. Therefore $|Z_\tau| < \frac{\epsilon}{4}$ a.a.s. By unraveling the construction of (Z_ℓ), we see that a.a.s.,

$$\left| \sum_{\ell=1}^{\tau} X_\ell - \sum_{\ell=1}^{\tau} \mathrm{E}[X_\ell \mid \mathcal{F}_{\ell-1}] \right| \le \frac{\epsilon}{4} \tag{5.1}$$

where the sum of conditional expectations is itself another random variable, which we must control.

Let S_ℓ, for $\ell - 1 \leq \tau$, be the initial size of the search tree T as the path is growing from length $\ell - 1$ to length ℓ; since this is determined by the shape of the search tree just as the path reaches length $\ell - 1$, S_ℓ is $\mathcal{F}_{\ell-1}$-measurable. If $\ell - 1 > \tau$, let it be an independent random variable, distributed as the length of the longest cycle in a uniformly random permutation of $\{1, \ldots, k\}$. Define the random variables

$$Y_\ell = \frac{1}{n-\ell} \sum_{i=S_\ell}^{k} \frac{1}{i}.$$

Y_ℓ is an $\mathcal{F}_{\ell-1}$-measurable estimate of the waiting time X_ℓ.

By Lemma 5.3.5, if $\ell \leq \tau$, then $\mathrm{E}[X_\ell \mid \mathcal{F}_{\ell-1}] \sim Y_\ell$, and so $\sum_{\ell=1}^{\tau} \mathrm{E}[X_\ell \mid \mathcal{F}_{\ell-1}] = (1 + o(1)) \sum_{\ell=1}^{\tau} Y_\ell$. Since $Y_\ell = O(\frac{1}{n})$, it follows that a.a.s.,

$$\left| \sum_{\ell=1}^{\tau} \mathrm{E}[X_\ell \mid \mathcal{F}_{\ell-1}] - \sum_{\ell=1}^{\tau} Y_\ell \right| \leq \frac{\epsilon}{4}. \tag{5.2}$$

Therefore, it now remains to control $\sum_{\ell=1}^{\tau} Y_\ell$.

We have an adapted process in which each Y_ℓ is $\mathcal{F}_{\ell-1}$-measurable, but it helps to study the Y_ℓ (or equivalently, S_ℓ) with respect to the coarsest possible filtration. Specifically, to observe S_ℓ, we now only need to watch the evolution of the search tree, and crucially, we may proceed by revealing only the number of vertices in the subtree of each child of the root. If we reveal these numbers at every step when an edge is added to the search tree, then in the ideal case, each subtree receives the edge with probability proportional to its size, and we have exactly the Chinese Restaurant Process. When we reach k edges and pass to the largest branch of the root, we reveal the next level of subtree size information. In the ideal case, conditioned on the previous partition and the size of the new search tree, when we reveal the new partition of subtree sizes, the distribution is precisely a new and independent Chinese Restaurant Process. It turns out that reality is not far off. Let $\mathcal{G}_\ell$ be the σ-algebra generated by $\{S_1, \ldots, S_\ell\}$, i.e., the natural filtration.

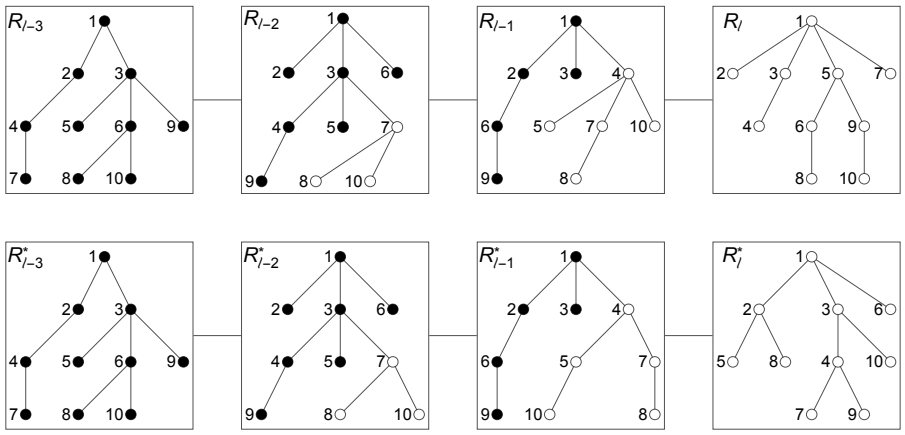


Figure 5.1: Two histories paired by the bijection replacing R_ℓ with R^*_ℓ. White nodes in the top history eventually become nodes of R_ℓ, and must be replaced in the bottom history by the corresponding nodes of R^*_ℓ.

Lemma 5.3.6. *For all $\ell \le \tau$, $\Pr[S_{\ell+1} = s \mid \mathcal{G}_\ell] = (1 + o(1)) \Pr[L_k = s]$, where L_k is the length of the longest cycle in a uniformly random permutation of $\{1, \dots, k\}$.*

Proof. If $\ell \le \tau$, then each vertex in the search tree T is incident to somewhere between $n - |P \cup T|$ and $n - |P \cup T| - o(n)$ unexposed edges whose other endpoint is outside $P \cup T$, by condition 4. Since $\ell \le \frac{9}{10}n$, this number is $(1 + o(1))(n - \ell)$ for each vertex in T. Furthermore, by Lemma 5.3.3, each of these edges has the same probability, up to a factor of $1 + o(1)$, of being added to T. Therefore each vertex in the search tree has the same probability, up to a factor of $1 + o(1)$, of acquiring a child as any other vertex.

The search tree at the moment when the path reaches length ℓ can be described as a recursive tree on k edges: if the vertices of the tree are labeled by the order in which they enter the search tree, then every child receives a label greater than its parent. In the ideal case when new leaves are added to uniformly random vertices, all $k!$ recursive trees would be equally likely. We will show that a similar statement holds here.

Define the *history up to length* ℓ to be the sequence of $(R_1, \dots, R_\ell)$ of recursive trees obtained

at lengths $1, 2, \ldots, \ell$. Not every sequence of recursive trees is a valid history: the trees must be consistent, since R_{i+1} must be a suitably relabeled extension of the largest branch of the root of R_i. Nevertheless, if we partition all valid histories by the value of R_ℓ, there is a natural bijective correspondence between any two parts. To change the final tree in a history $(R_1, \ldots, R_\ell)$ to R^*_ℓ, we make the same substitution in subtrees of $R_{\ell-1}$, $R_{\ell-2}$, and so on, going back to $R_{\ell-k+1}$ at worst. If the first i nodes of R_ℓ are a subtree of $R_{\ell-j}$, we replace them by the first i nodes of R^*_ℓ. An example of two histories paired by this correspondence is given in Figure 5.1.

Two histories corresponding in this way have the same value of $S_1, S_2, \ldots, S_\ell$: while the shapes of the subtrees measured by these random variables may be different, their sizes are the same. Moreover, since the two histories agree on all but the last k trees, they only disagree in at most k^2 steps of the algorithm, so the probability of obtaining them differs by a factor of $(1 + o(1))^{k^2} = 1 + o(1)$. It follows that $(R_\ell \mid \mathcal{G}_\ell)$ is asymptotically uniformly distributed.

Since R_ℓ can take on exactly $k!$ values, we must have $\Pr[R_\ell = R \mid \mathcal{G}_\ell] = (1 + o(1))\frac{1}{k!}$ for any recursive tree R. It follows that any event determined by the shape of R_ℓ has asymptotically the same probability of occurring, conditioned on $\mathcal{G}_\ell$, as it would if $(R_\ell \mid \mathcal{G}_\ell)$ were a uniformly random recursive tree.

In a uniformly chosen recursive tree, the size of the largest branch of the root has the same distribution as L_k. Therefore $S_{\ell+1}$, the size of the largest branch of the root of R_ℓ, satisfies $\Pr[S_{\ell+1} = s \mid \mathcal{G}_\ell] = (1 + o(1)) \Pr[L_k = s]$, as desired. □

Since k is a constant, each $\Pr[L_k = s]$ is a constant, for each fixed $s \in \{1, \ldots, k\}$. So, by Lemma 5.3.6, the conditional distribution of Y_ℓ given $\mathcal{G}_{\ell-1}$ is also supported on k values in the range $\Theta(\frac{1}{n})$, with all probabilities bounded away from 0 and 1 by at least some constant depending on k. Crucially, regardless of the particular $\mathcal{G}_{\ell-1}$, the distribution of Y_ℓ is always asymptotically $\frac{1}{n-\ell}A_k$, where $A_k = \frac{1}{L_k} + \cdots + \frac{1}{k}$ is the random variable defined with respect to the longest cycle in the Chinese Restaurant Process at the end of Section 5.3.3.

To complete the analysis, define the martingale

$$W_\ell = (Y_1 - \mathrm{E}[Y_1]) + (Y_2 - \mathrm{E}[Y_2 \mid \mathcal{G}_1]) + \cdots + (Y_\ell - \mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}]).$$

It is Lipschitz with successive variations bounded by $O(\frac{1}{n})$ because k is a constant, and so the Azuma–Hoeffding inequality applied to (W_ℓ) implies that a.a.s.,

$$|W_\tau| \le \frac{\log n}{\sqrt{n}} \le \frac{\epsilon}{4},$$

or equivalently, that a.a.s.,

$$\left| \sum_{\ell=1}^{\tau} Y_\ell - \sum_{\ell=1}^{\tau} \mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}] \right| \le \frac{\epsilon}{4}. \tag{5.3}$$

Since the distribution of $(Y_\ell \mid \mathcal{G}_{\ell-1})$ is asymptotically $\frac{1}{n-\ell} A_k$, the random variable $\mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}]$ is asymptotically constant: $\mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}] = (1 + o(1)) \frac{1}{n-\ell} \mathrm{E}[A_k]$, which is $(1 + o(1)) \frac{\alpha_k}{n-\ell}$, by definition of α_k. Therefore a.a.s.,

$$\sum_{\ell=1}^{\tau} \mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}] = (1 + o(1)) \alpha_k \log \frac{n}{n-\tau}.$$

Since $\tau \le \frac{9}{10} n$, the right-hand side is bounded by a constant, so a.a.s.,

$$\left| \sum_{\ell=1}^{\tau} \mathrm{E}[Y_\ell \mid \mathcal{G}_{\ell-1}] - \alpha_k \log \frac{n}{n-\tau} \right| \le \frac{\epsilon}{4}. \tag{5.4}$$

Combining inequalities (5.1), (5.2), (5.3), and (5.4), we obtain that a.a.s.,

$$\left| \sum_{\ell=1}^{\tau} X_\ell - \alpha_k \log \frac{n}{n-\tau} \right| \le \epsilon,$$

as desired. Since $\sum_{\ell=1}^{\tau} X_\ell \in [0, 1]$, we have $\alpha_k \log \frac{n}{n-\tau} \le 1 + \epsilon$ and $\tau \le (1 - \exp(-\frac{1+\epsilon}{\alpha_k}))n$ a.a.s.

In particular, $\tau < \frac{9}{10} n$ a.a.s., and by Lemma 5.3.4, the algorithm is a.a.s. well-behaved for $\ell < \frac{9}{10} n$; therefore a.a.s. the algorithm is still well-behaved at length τ, and the stopping rule triggered because the algorithm terminated. In that case, by condition (2), the algorithm's time t never increased by more than C_2, yet t reached 1 when the algorithm terminated. Therefore when the path first reached length τ, which was at most k steps before the algorithm terminated,

the time $t = \sum_{\ell=1}^{\tau} X_\ell$ must have been at least $1 - kC_2 = 1 - o(1)$ a.a.s. Therefore a.a.s., $\alpha_k \log \frac{n}{n-\tau} \geq 1 - 2\epsilon$ and $\tau \geq (1 - \exp(-\frac{1-2\epsilon}{\alpha_k}))n$.

Since ϵ can be made arbitrarily small, and a.a.s. the algorithm produces a path of length τ, the a.a.s. bounds

$$\left(1 - \exp\left(-\frac{1-2\epsilon}{\alpha_k}\right)\right) n \leq \tau \leq \left(1 - \exp\left(-\frac{1+\epsilon}{\alpha_k}\right)\right) n$$

complete the proof of Theorem 5.1.1. □

5.4 Computing the second moment of H_n

The core of our proof of Theorem 5.1.2 is the second moment calculation for H_n, the random variable which tracks the number of increasing Hamiltonian paths in a uniformly random edge ordering. This second moment, H_n^2, counts the number of ordered pairs of increasing Hamiltonian paths, which can be expressed as a sum of indicator variables: $H_n^2 = \sum_A \sum_B I_{A,B}$, where A and B range over all Hamiltonian paths, and $I_{A,B} = 1$ if both paths are increasing when f is chosen, and 0 otherwise. Note that although we are working with undirected graphs, we consider Hamiltonian paths with direction, and therefore, when we speak of a Hamiltonian path in this section, we are referring to a permutation of the n vertices. In particular, each undirected n-vertex path will correspond to two such permutations, and will appear twice in our indexing, once in each direction.

We begin by grouping pairs of paths into equivalence classes which we call *intersection profiles*. Two pairs of paths (A, B) and (A', B') are in the same intersection profile if there is a permutation of the edges of K_n (without necessarily preserving all pairwise incidence relations between edges) that maps A to A' and B to B'. We can represent such a profile as a graph by choosing any representative (A, B), and separating the vertices of A and B that are not endpoints of a common edge. Figure 5.2 shows an example of such a graph. Note that when the two paths share a single common edge, it's possible for them to visit its endpoints in two different orders,

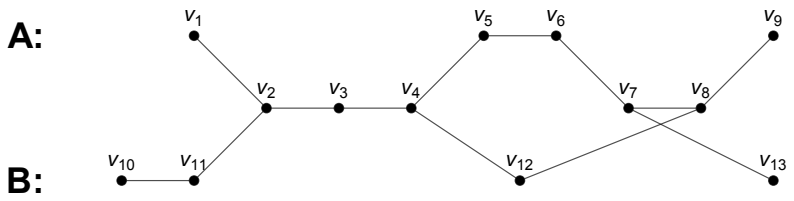


Figure 5.2: A graph representation of an intersection profile for two paths: $A = (v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9)$ and $B = (v_{10}, v_{11}, v_2, v_3, v_4, v_{12}, v_8, v_7, v_{13})$.

as with edge (v_7, v_8) in Figure 5.2.

Vertices not incident to a common edge of A and B appear twice: in Figure 5.2, vertices $\{v_1, v_5, v_6, v_9\}$ are the same as $\{v_{10}, v_{11}, v_{12}, v_{13}\}$. However, changing the exact correspondence between these sets of vertices would not change the intersection profile, as long as no new common edges are created.

We keep track of three parameters of a pair of paths (A, B); these are preserved by the permutations that map (A, B) to a different pair of paths (A', B') in the same intersection profile, and so we may treat them as properties of the profile P containing (A, B).

1. $c(P)$ is the number of common edges shared by the two paths A and B. In Figure 5.2, $c(P) = 3$, as the two paths share edges (v_2, v_3), (v_3, v_4), and (v_7, v_8).
2. $k(P)$ is the number of common segments shared by the two paths: this satisfies $k(P) \leq c(P)$ because each common segment contains at least one edge, and $k(P) \leq n - c(P)$ because each path contains at least one edge between two common segments. In Figure 5.2, the two common segments are (v_2, v_3, v_4) and (v_7, v_8), so $k(P) = 2$.
3. $\ell(P)$ is the number of common segments which consist of exactly one edge: since $c(P) \geq \ell(P) + 2[k(P) - \ell(P)]$, this satisfies $\ell(P) \geq 2k(P) - c(P)$. In Figure 5.2, $\ell(P) = 1$, counting the segment (v_7, v_8).

We define $\mathcal{P}(c, k, \ell)$ as the set of profiles P such that $c(P) = c$, $k(P) = k$, and $\ell(P) = \ell$.

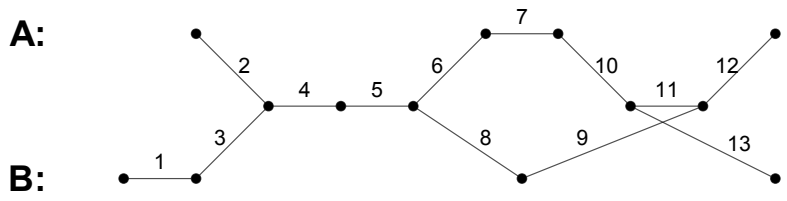


Figure 5.3: A graph representation of a labeled profile for two paths of length 8, whose underlying profile is the one given in Figure 5.2.

Given a profile P represented by the pair of paths (A, B), there are many ways to choose a relative order for the edges of A and B so that both A and B are increasing. Furthermore, the number of ways to do so does not depend on the choice of representative (A, B). If a relative order of the edges of A and B is chosen such that A and B are increasing, and (A', B') is another pair that fits P, a permutation of the edges of K_n mapping A to A' and B to B' will induce a unique relative order of the edges of A' and B' such that A' and B' are increasing. Therefore we may speak of a labeling of P, by which we mean a choice of relative order for the edges of some representative (A, B) that makes A and B increasing.

We define a *labeled profile* L as a profile, together with such a labeling. We say that two paths A and B fit L if they fit (are a representative of) the unlabeled profile corresponding to L. An example of a labeled profile is given in Figure 5.3. We let $\mathcal{L}(c, k, \ell)$ denote the set of labeled profiles whose underlying profile is an element of $\mathcal{P}(c, k, \ell)$.

Given a labeled profile $L \in \mathcal{L}(c, k, \ell)$, the total number of edges in two paths that fit L is $2(n-1) - c$. For any pair of paths A and B that fits L, if an edge ordering f is chosen randomly, the ordering of A and B will match the labeling of L with probability $\frac{1}{(2n-c-2)!}$. Therefore

$$\mathrm{E}[H_n^2] = \sum_{c,k,\ell} \sum_{L \in \mathcal{L}(c,k,\ell)} \frac{|L|}{(2n-c-2)!} \tag{5.5}$$

where $|L|$ is the number of pairs of paths (A, B) that fit L.

We split the sum (5.5) into several parts:

1. S_1, the sum over $c \leq \log n$ (most other notions of "small" c would also be sufficient here);
2. S_2, the sum over $\log n < c \leq \frac{9}{10}n$; and
3. S_3, the sum over $c > \frac{9}{10}n$.

It will be S_1 that contributes the most to $\mathrm{E}[H_n^2]$. We state two lemmas that provide asymptotically exact estimates (with multiplicative error tending to zero as $n \to \infty$) for S_1 while only serving as rough upper bounds for S_2 and S_3. In the following, $\binom{n}{a,b,c}$ denotes the multinomial coefficient $\frac{n!}{a!b!c!}$.

Lemma 5.4.1. *For all c, k, ℓ,*

$$|\mathcal{L}(c,k,\ell)| \leq 2^\ell \binom{k}{\ell}\binom{c-k-1}{k-\ell-1}\binom{2(n-c-1)+k}{n-c-1,n-c-1,k}$$

and $|\mathcal{L}(c,k,\ell)|$ is asymptotically equal to the right hand side when $c \leq \log n$. Furthermore,

$$\sum_\ell |\mathcal{L}(c,k,\ell)| \leq 2^k \binom{c-1}{k-1}\binom{2(n-c-1)+k}{n-c-1,n-c-1,k}$$

Lemma 5.4.2. *For all $L \in \mathcal{L}(c,k,\ell)$, $|L| \leq n!(n-c-k)!$; furthermore, if $c \leq \log n$, then $|L| \sim e^{-2}n!(n-c-k)!$.*

By using these lemmas, we can write out algebraic expressions for S_1, S_2, and S_3:

$$S_1 \sim e^{-2}\sum_{c=0}^{\log n}\sum_{k,\ell} 2^\ell \binom{k}{\ell}\binom{c-k-1}{k-\ell-1}\binom{2(n-c-1)+k}{n-c-1,n-c-1,k}\frac{n!(n-c-k)!}{(2n-c-2)!} \tag{5.6}$$

$$S_2 \leq \sum_{c=\log n}^{9n/10}\sum_{k} 2^k \binom{c-1}{k-1}\binom{2(n-c-1)+k}{n-c-1,n-c-1,k}\frac{n!(n-c-k)!}{(2n-c-2)!} \tag{5.7}$$

$$S_3 \leq \sum_{c=9n/10}^{n-1}\sum_{k} 2^k \binom{c-1}{k-1}\binom{2(n-c-1)+k}{n-c-1,n-c-1,k}\frac{n!(n-c-k)!}{(2n-c-2)!}. \tag{5.8}$$

Therefore, $\mathrm{E}[H_n^2] \sim en^2$ will follow from:

Lemma 5.4.3. *The right hand side of the expression* (5.6) *is asymptotic to en^2, and both of the right hand sides of* (5.7) *and* (5.8) *simplify to $o(n^2)$.*

5.4.1 Asymptotics for $|\mathcal{L}(c,k,\ell)|$ and for $|L|$ when $L \in \mathcal{L}(c,k,\ell)$

Proof of Lemma 5.4.1. Let c, k, and ℓ be given. For the remainder of the proof, let $m = n-c-1$ stand for the number of edges that belong to A but not B (equivalently, to B but not A), when paths A and B fit a profile from $\mathcal{L}(c,k,\ell)$. Consider the following two-stage method for selecting a labeled profile from $\mathcal{L}(c,k,\ell)$. All such labeled profiles will be reachable in this way.

1. Choose the sequence of lengths for the common segments, and their relative orientations within paths A and B. The segments of length 1 can appear in $\binom{k}{\ell}$ positions, and the remaining $c-\ell$ common edges can be partitioned into $k-\ell$ ordered parts of size at least 2 in $\binom{c-k-1}{k-\ell-1}$ ways. Common segments of length at least 2 will already have a fixed orientation, because their sequential edge labels will need to be increasing with respect to both A and B. On the other hand, a common segment of length 1 could be traversed in either the same direction by both paths, or in opposite directions (as in the case of the second common segment in Figures 5.2 and 5.3). Therefore, by considering relative orientations, we gain another factor of exactly 2^ℓ.
2. Now that the sequence of lengths and directions has been fixed for all common segments, it remains to choose an order in which the k common segments, the m edges of A, and the m edges of B appear. For this, we construct labeled profiles from strings of m *A*'s, m *B*'s, and k *C*'s. For example, if we have already fixed the first common segment to have length 2, and the other common segment to have length 1, traversed in both directions, then the string *BABCAABBACAB* corresponds precisely to Figure 5.3. (Note that a single C represents an entire common segment, whose length we have already fixed, and not a single common edge.) There are at most $\binom{2m+k}{m,m,k}$ such strings of *A*'s, *B*'s, and *C*'s.

The above two-step procedure immediately implies the claimed upper bound on $|\mathcal{L}(c,k,\ell)|$ in Lemma 5.4.1. Our next objective is to show that this bound is asymptotically correct for $c \leq \log n$. The second step overestimates $|\mathcal{L}(c,k,\ell)|$ because of two possible illegal interactions between adjacent common segments: (a) we cannot have two consecutive *C*'s, separated by all

A's or all *B*'s, and (b) no consecutive *C*'s can be separated by exactly one *A* and exactly one *B*. We will show that the number of such strings is an $o(1)$-fraction of the total number of strings which appear in the second step.

First, we control the number of strings which have two *C*'s which are separated only by *B*'s. For this, fix one of the $k-1$ gaps between common segments, and suppose that there are exactly $m-i$ *B*'s in the gap, with $0 \leq i \leq m$. Then, those strings are in bijective correspondence with the strings with exactly m *A*'s, exactly i *B*'s, and exactly $k-1$ *C*'s: the bijection is realized by the deletion of a segment *BB*...*BC* (with $m-i$ *B*'s) after the *C* which corresponds to the beginning of the gap. Thus, the total number of strings which have two *C*'s separated only by *B*'s is at most

$$\begin{aligned}(k-1)\sum_{i=0}^{m}\binom{m+i+k-1}{m,i,k-1} &= (k-1)\binom{m+k-1}{k-1}\sum_{i=0}^{m}\binom{m+k-1+i}{i} \\ &= (k-1)\binom{m+k-1}{k-1}\binom{2m+k}{m} \\ &= (k-1)\frac{k}{m+k}\binom{2m+k}{m,m,k}. \end{aligned} \tag{5.9}$$

The same bound applies for the total number of strings with two *C*'s separated only by *A*'s. Similarly, if two *C*'s are separated by exactly one *A* and one *B*, there are two possible orderings (*AB*, *BA*) between the *C*'s, and so the total number of such strings is at most

$$\begin{aligned}(k-1)2\binom{2m+k-3}{m-1,m-1,k-1} &= (k-1)2\frac{(2m+k-3)!}{(m-1)!(m-1)!(k-1)!} \\ &= \frac{2(k-1)m^2k}{(2m+k)(2m+k-1)(2m+k-2)}\cdot\frac{(2m+k)!}{m!m!k!} \end{aligned} \tag{5.10}$$

When $k \leq c \leq \log n$, both (5.9) and (5.10) are of order $\frac{k^2}{m}\binom{2m+k}{m,m,k}$, and $\frac{k^2}{m} = o(1)$. Therefore the true number of choices to be made in the second step is indeed $(1-o(1))\binom{2m+k}{m,m,k}$ for small c and k, as claimed. Finally, to obtain the rougher approximation for the second part of the lemma, we forget about the value of ℓ, and get an upper bound by assuming that all k segments can be reversed, even if they don't have length 1. Effectively, we use $2^\ell \leq 2^k$ and $\sum_\ell \binom{k}{\ell}\binom{c-k-1}{k-\ell-1} = \binom{c-1}{k-1}$. □

Proof of Lemma 5.4.2. Since the number of pairs (A, B) that fit L is the same as the number that fit P, the underlying profile of L, we may as well forget about the labeling, and count paths that fit a profile $P \in \mathcal{P}(c, k, \ell)$.

The first part of the lemma is immediate: there are $n!$ ways to choose the n vertices of A, and $(n - c - k)!$ ways to choose the remaining $n - c - k$ vertices of B. However, this mistakenly counts some pairs of paths that don't fit the profile P. For example, if the profile in Figure 5.2 embeds into K_n by sending v_5 and v_{10} to the same vertex $v \in K_n$, and sending v_6 and v_{11} to the same vertex $w \in K_n$, then the embedded paths no longer correspond to the intersection profile in the figure, because additional common segments are created. So, to prove the second part of the lemma, we must estimate the probability that in a random permutation of the $n - c - k$ vertices of B which are not on common segments with A, no new common segments are created between the embedded paths.

We first consider the case $c = k = 0$, which corresponds to the probability that in a random permutation of $\{1, 2, \ldots, n\}$, no two consecutive elements are adjacent. Wolfowitz [53] has shown that asymptotically, the number of adjacent consecutive elements has the Poisson distribution with mean 2, and therefore we obtain the desired probability of e^{-2}.

For the general case, suppose that the n vertices of A have been fully embedded into K_n. Exactly $n - c - k$ of them correspond to vertices of A which are not shared by B in the profile diagram. Following the natural order for A in the profile, label those $n - c - k$ embedded vertices (in K_n) by $1, 2, \ldots, n - c - k$. Then, each embedding of the remaining $n - c - k$ vertices of B (which completes the embedding of the two paths A and B) corresponds precisely to a distinct permutation of $\{1, 2, \ldots, n - c - k\}$, because both A and B are Hamiltonian paths, and thus each use all of the vertices. Here, the permutation is the order in which the vertices $\{1, 2, \ldots, n - c - k\}$ are visited when the embedded B is traversed in its natural order. Permutations with adjacent consecutive elements still approximately correspond to embeddings which create extraneous common segments, and it remains to quantify the error in the approximation,

which arises at junctions with common segments.

When A is traced in its natural order, there are either $k-1$ or k vertices of A which come immediately before the start of a common segment. (There are $k-1$ if the first vertex of A is already part of a common edge between the two paths, and k otherwise.) Let $i_j \in \{1, 2, \ldots, n\}$ with $j = 1, 2, \ldots, k$ be the label in K_n of the embedded vertex corresponding to the vertex of A which comes immediately before the start of the j-th common segment. If there are only $k-1$ such vertices, then leave i_1 undefined, and ignore all references to it in the remainder of this argument.

In terms of the i_j's, permutations σ with adjacent consecutive elements correspond to embeddings with extraneous common segments, except when for some j, we have that (a) i_j and $i_j + 1$ are adjacent in σ, or (b) the vertex of B which immediately precedes the j-th common segment maps to i_j, or (c) the vertex of B which comes right after the j-th common segment maps to $i_j + 1$. To see this, observe that (a) identifies a "false positive," in which the elements are adjacent in σ, but are actually separated by a common segment in B's traversal. On the other hand, (b) and (c) represent the "false negatives," in which the j-th common segment is unduly extended. Fortunately, a union bound over all j shows that the probability of (a) happening is at most $\frac{2k}{n-c-k-1}$, the probability of (b) happening is at most $\frac{k}{n-c-k}$, and the probability of (c) has the same bound. All of these quantities are $o(1)$ for $k \leq c \leq \log n$, and therefore the probability that no new common segments are created differs by $o(1)$ from the probability that no consecutive elements occur, and is also asymptotically e^{-2}. □

5.4.2 Estimating S_1

To simplify the expressions involved, we use the notation $(n)_k$ for the falling power $\frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$. This satisfies $(n)_k \sim n^k$ for $k = o(\sqrt{n})$. In particular, for $c \leq \log n$, we have $k \leq \log n$ as well; therefore for falling powers linear in c and k, we may freely use this

asymptotic relation. Starting from (5.6), we have:

$$\begin{aligned}
S_1 &\sim e^{-2} \sum_{c=0}^{\log n} \sum_{k,\ell} 2^\ell \binom{k}{\ell} \binom{c-k-1}{k-\ell-1} \binom{2(n-c-1)+k}{n-c-1, n-c-1, k} \frac{n!(n-c-k)!}{(2n-c-2)!} \\
&= e^{-2} \sum_{c=0}^{\log n} \sum_{k,\ell} 2^\ell \binom{k}{\ell} \binom{c-k-1}{k-\ell-1} \frac{(2n-2c+k-2)!n!(n-c-k)!}{k!(n-c-1)!^2(2n-c-2)!} \\
&= e^{-2} \sum_{c=0}^{\log n} \sum_{k,\ell} 2^\ell \binom{k}{\ell} \binom{c-k-1}{k-\ell-1} \frac{1}{k!} \cdot \frac{(n)_{c+1}}{(n-c-1)_{k-1}(2n-c-2)_{c-k}} \\
&\sim e^{-2} n^2 \sum_{c=0}^{\log n} \sum_{k,\ell} \binom{k}{\ell} \binom{c-k-1}{k-\ell-1} \frac{2^{\ell-c+k}}{k!}.
\end{aligned}$$

Splitting off the $e^{-2}n^2$ factor, we are left with a sum of the first $\log n$ terms of an infinite series that does not otherwise depend on n. It turns out that this series converges to a constant C as $n \to \infty$, and we now compute this limit. Recall that the constraints on k and ℓ in the inner sum are that $0 \le \ell \le k \le c$ and, furthermore, that $c \ge \ell + 2(k-\ell) = 2k - \ell$ (which is a stronger bound than $c \ge k$). Therefore

$$\begin{aligned}
C &= \sum_{c=0}^{\infty} \sum_{k,\ell} \binom{k}{\ell} \binom{c-k-1}{k-\ell-1} \frac{2^{\ell-c+k}}{k!} \\
&= \sum_{k=0}^{\infty} \frac{2^k}{k!} \sum_{\ell=0}^{k} 2^\ell \binom{k}{\ell} \sum_{c=2k-\ell}^{\infty} \binom{c-k-1}{k-\ell-1} 2^{-c} \\
&= \sum_{k=0}^{\infty} \frac{2^k}{k!} \sum_{\ell=0}^{k} 2^\ell \binom{k}{\ell} 2^{-2k+\ell} \sum_{j=0}^{\infty} \binom{j+(k-\ell-1)}{k-\ell-1} 2^{-j},
\end{aligned}$$

where we re-parameterized the final sum as $j = c - (2k - \ell)$. The final summation is now conveniently in the form of the following power series identity:

$$\sum_{j=0}^{\infty} \binom{j+m-1}{m-1} z^j = \frac{1}{(1-z)^m}.$$

Therefore,

$$\begin{aligned}
C &= \sum_{k=0}^{\infty} \frac{2^k}{k!} \sum_{\ell=0}^{k} 2^\ell \binom{k}{\ell} 2^{-2k+\ell} \cdot 2^{k-\ell} \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{\ell=0}^{k} 2^\ell \binom{k}{\ell} = \sum_{k=0}^{\infty} \frac{3^k}{k!} = e^3,
\end{aligned}$$

which implies that $S_1 \sim e^{-2}n^2 C = en^2$, as claimed.

5.4.3 Estimating S_2

Let $a_{c,k}$ be one of the summands in (5.7). Then

$$\frac{a_{c,k+1}}{a_{c,k}} = \frac{2(c-k)}{k(k+1)} \cdot \frac{2n-2c+k-1}{n-c-k}. \tag{5.11}$$

Our goal is to simplify the upper bound on S_2 by selecting, for each c, the k that maximizes $a_{c,k}$, and then using this maximum in place of all the terms with that value of c.

First consider k such that $k \le \frac{1}{2}(n-c)$. In this case, the second factor of (5.11) is bounded between 1 and 5: on one hand, $(2n-2c+k-1)-(n-c-k) = n-c+2k-1 \ge 0$, and on the other hand, $(2n-2c+k-1)-5(n-c-k) = 6k-3(n-c)-1 \le 0$. Therefore we have

$$\frac{2(c-k)}{k(k+1)} \le \frac{a_{c,k+1}}{a_{c,k}} \le \frac{10(c-k)}{k(k+1)}.$$

If k maximizes $a_{c,k}$ and lies in this range, then $2(c-k) \le k(k+1)$ and therefore $k \ge (1-o(1))\sqrt{2c}$; otherwise, $10(c-k-1) \ge k(k-1)$ and therefore $k \le (1-o(1))\sqrt{10c}$. We may safely and concisely say $\sqrt{c} < k < 4\sqrt{c}$.

On the other hand, if $k \ge \frac{1}{2}(n-c)$, since S_2 only runs c up to $\frac{9n}{10}$, we have $k \ge \frac{1}{2}(n-c) \ge n/20$ in the denominator of (5.11), and we always have $k \le c \le n$ in the numerator. So,

$$\frac{a_{c,k+1}}{a_{c,k}} < \frac{2n}{k^2} \cdot \frac{3n}{n-c-k} \le \frac{2n}{n^2/400} \cdot \frac{3n}{n-c-k} = \frac{2400}{n-c-k},$$

which is less than 1 as long as $k \le n-c-2400$. Therefore the maximizing k is either in the range found above, or between $n-c-2400$ and $n-c$ (since $k \le n-c$ always).

Let S_2' be the result of replacing in S_2 all terms $a_{c,k}$ by a_{c,k^*} where $\sqrt{c} < k^*(c) < 4\sqrt{c}$ is the maximizing k from the range $0 \le k \le n-c-2400$. Then

$$\begin{aligned}
S_2' &< \sum_{c=\log n}^{9n/10} c \cdot 2^{k^*} \binom{c-1}{k^*-1} \binom{2(n-c-1)+k^*}{n-c-1, n-c-1, k^*} \frac{n!(n-c-k^*)!}{(2n-c-2)!} \\
&= \sum_{c=\log n}^{9n/10} \frac{k^* 2^{k^*}}{k^*!} \binom{c}{k^*} \frac{(n)_{c+1}}{(n-c-1)_{k^*-1}(2n-c-2)_{c-k^*}} \\
&< \sum_{c=\log n}^{9n/10} k^* \left(\frac{2e}{k^*}\right)^{k^*} \left(\frac{ce}{k^*}\right)^{k^*} \frac{(n)_{c+1}}{(n-c-1)_{k^*-1}(2n-c-2)_{c-k^*}}
\end{aligned}$$

$$< n^2 \sum_{c=\log n}^{9n/10} k^*(2e^2)^{k^*} \frac{(n-2)_{c-k^*}}{(2n-c-2)_{c-k^*}} \cdot \frac{(n-c+k^*-2)_{k^*-1}}{(n-c-1)_{k^*-1}}.$$

We now eliminate the powers of n in the summand. In the first fraction, since $c \le \frac{9}{10}n$, $2n-c \ge \frac{11}{10}n$, and therefore $\frac{(n-2)_{c-k^*}}{(2n-c-2)_{c-k^*}} \le \left(\frac{10}{11}\right)^{c-k^*}$. In the second fraction, since k^* is less than $\frac{1}{2}(n-c)$, $n-c+k^*-2 < \frac{3}{2}(n-c-1)$, and therefore $\frac{(n-c+k^*-2)_{k^*-1}}{(n-c-1)_{k^*-1}} < (3/2)^{k^*}$. Thus

$$\begin{aligned} S_2' &< n^2 \sum_{c=\log n}^{9n/10} k^* \left(2e^2 \cdot \frac{11}{10} \cdot \frac{3}{2}\right)^{k^*} \left(\frac{10}{11}\right)^c \\ &= n^2 \sum_{c=\log n}^{9n/10} k^* \left(\frac{33e^2}{10}\right)^{k^*} \left(\frac{10}{11}\right)^c \\ &< n^2 \sum_{c=\log n}^{\infty} 4\sqrt{c} \left(\frac{33e^2}{10} \cdot \left(\frac{10}{11}\right)^{\sqrt{c}/4}\right)^{4\sqrt{c}}. \end{aligned}$$

The sum is the tail of a convergent series in c, and therefore $S_2' = o(n^2)$.

To show that $S_2 = o(n^2)$, it remains to consider the terms $a_{c,k}$ for which $n-c-k \le 2400$, as these are potentially not dominated by a_{c,k^*}. For this case, we consider a second ratio:

$$\frac{a_{c,k}}{a_{c+1,k}} = \left(\frac{c-k+1}{c}\right) \cdot \frac{(2n-2c+k-2)(2n-2c+k-3)}{(n-c+1)(n-c-1)} \cdot \frac{(n-c-k)}{(2n-c-2)}.$$

Here, $n-c-k \le 2400$ and $c-k+1 \le c$; all other factors are $\Theta(n)$ because $c \le \frac{9n}{10}$, and so the overall ratio is $O(n^{-1})$. Therefore, once n surpasses some absolute constant, all of these $a_{c,k}$ with $n-c-k \le 2400$ satisfy $a_{c,k} \le a_{n-k,k}$, and there are at most $2400n$ of them. It remains to control $a_{n-k,k}$ in the range $k \ge n-c-2400 \ge (1-o(1))\frac{n}{10}$, where we used $c \le \frac{9n}{10}$. For those, we have

$$\begin{aligned} a_{n-k,k} &< 2^k \binom{n-k-1}{k-1} \binom{3k-2}{k-1,k-1,k} \frac{n!}{(n+k-2)!} \\ &< 2^k 2^{n-k-1} 3^{3k-2} \cdot \frac{1}{(n+k-2)_{k-2}} \\ &< \frac{54^n}{(n+k-2)_{k-2}} \\ &< \frac{54^n}{n^{(1-o(1))n/10}} = o(1). \end{aligned}$$

Therefore, the total contribution of these residual $a_{c,k}$ is at most $2400n \cdot o(1)$, and $S_2 = S_2' + o(n) = o(n^2)$, as claimed.

5.4.4 Estimating S_3

Let $d = n - c$ and consider the sum for $1 \le d \le \frac{n}{10}$. Then from (5.8) we get

$$\begin{aligned} S_3 &\le \sum_{k=1}^{n/10} \sum_{d=k}^{n/10} 2^k \binom{n-d-1}{k-1} \binom{2d+k-2}{d-1, d-1, k} \frac{n!(d-k)!}{(n+d-2)!} \\ &\le \sum_{k=1}^{n/10} \sum_{d=k}^{n/10} \frac{2^k}{(k-1)!} \cdot \frac{(n-d-1)!}{(n-d-k)!} \cdot 3^{2d+k-2} \cdot \frac{n!(d-k)!}{(n+d-2)!} \\ &= \sum_{k=1}^{n/10} \sum_{d=k}^{n/10} \frac{6^k \cdot 9^{d-1}}{(k-1)!} \cdot \frac{(n-d-1)_{k-1}(d-k)!}{(n+d-2)_{d-2}} \\ &\le \sum_{k=1}^{n/10} \sum_{d=k}^{n/10} \frac{6^k \cdot 9^{d-1} \cdot (d-k)!}{(k-1)! \cdot n^{d-k-1}}. \end{aligned}$$

Let $b_{d,k}$ be a term of this sum; since $d - k \le d \le \frac{n}{10}$,

$$\frac{b_{d,k}}{b_{d-1,k}} = \frac{9(d-k)}{n} \le \frac{9}{10}.$$

Therefore an upper bound on S_3 is:

$$S_3 \le \sum_{k=1}^{n/10} b_{k,k} \sum_{d=k}^{n/10} \left(\frac{9}{10}\right)^{d-k} \le 10 \sum_{k=1}^{\infty} \frac{6^k \cdot 9^{k-1}}{(k-1)! \cdot n^{-1}} = 60e^{54}n = o(n^2)\,,$$

which completes our proof.

Chapter 6

Distance-Uniform Graphs with Large Diameter

6.1 Background

We say that an n-vertex graph is *ϵ-distance-uniform* if there is a value d, called the *critical distance*, such that, for every vertex v, all but ϵn of the other vertices are at distance exactly d from v. This notion is introduced by Alon, Demaine, Hajiaghayi, and Leighton in [2], motivated by the analysis of network creation games.

In a network creation game, according to a model proposed by Fabrikant et al. in [20], a graph is constructed and modified by independently acting nodes. Each node must pay a creation cost to add nodes to the graph, but once the graph is built, each node must pay usage costs depending on how well-connected it is to other nodes in the graph. (This can be modeled in various ways, but we will take the usage cost to be a node's total distance to all other nodes.) Research in this area focuses on comparing the total cost to the nodes when this game is played to the optimal value of this cost over all graphs. The ratio between these quantities is known as the price of anarchy, a term first used by Koutsoupias and Papadimitriou [35].

Alon et al. simplify the problem by starting with an existing graph, and allowing each node

v to perform edge swaps, adding one edge vw and deleting another edge vw'; this lets us separate creation costs from usage costs. A graph is said to be in sum equilibrium when no node can decrease its total distance to other nodes by any edge swap. If this network creation game terminates, it ends with a graph in sum equilibrium; as a result, the price of anarchy is given by comparing the worst-case and best-case sum equilibrium graph.

Alon et al. show that sufficiently large graph powers of a graph in sum equilibrium will result in distance-uniform graphs; if the critical distance is large, then the original sum equilibrium graph imposed a high total cost on its nodes. This application motivates the already natural question: in an ϵ-distance-uniform graph with n vertices and critical distance d, what is the relationship between the parameters ϵ, n, and d? Specifically, can we derive an upper bound on d in terms of ϵ and n? Up to a constant factor, this is equivalent to finding an upper bound on the diameter of the graph, which must be between d and $2d$ as long as $\epsilon > \frac{1}{2}$.

Random graphs provide one example of distance-uniform graphs. In [10], Bollobás shows that for sufficiently large $p = p(n)$, the diameter of the random Erdős–Rényi random graph $\mathcal{G}_{n,p}$ is asymptotically almost surely concentrated on one of two values. In fact, from every vertex v in $\mathcal{G}_{n,p}$, the breadth-first search tree expands by a factor of $O(np)$ at every layer, reaching all or almost all vertices after about $\log_r n$ steps. Such a graph is also expected to be distance-uniform: the biggest layer of the breadth-first search tree will be much bigger than all previous layers.

More precisely, suppose that we choose $p(n)$ so that the average degree $r = (n-1)p$ satisfies two criteria: that $r \gg (\log n)^3$, and that for some d, $r^d/n - 2\log n$ approaches a constant C as $n \to \infty$. Then it follows from Lemma 3 in [10] that (with probability $1 - o(1)$) for every vertex v in $\mathcal{G}_{n,p}$, the number of vertices at each distance $k < d$ from v is $O(r^k)$. It follows from Theorem 6 in [10] that the number of vertex pairs in $\mathcal{G}_{n,p}$ at distance $d+1$ from each other is Poisson with mean $\frac{1}{2}e^{-C}$, so there are only $O(1)$ such pairs with probability $1 - o(1)$. As a result, such a random graph is ϵ-distance-uniform with $\epsilon = O(\frac{\log n}{r})$, and critical distance $d = \log_r n + O(1)$.

This example provides a compelling image of what distance-uniform graphs look like: if

the breadth-first search tree from each vertex grows at the same constant rate, then most other vertices will be reached in the same step. In any graph that is distance-uniform for a similar reason, the critical distance d will be at most logarithmic in n. In fact, Alon et al. conjecture that all distance-uniform graphs have diameter $O(\log n)$.

Alon et al. prove an upper bound of $O(\frac{\log n}{\log \epsilon^{-1}})$ in a special case: for ϵ-distance-uniform graphs with $\epsilon < \frac{1}{4}$ that are Cayley graphs of Abelian groups. In this case, if G is the Cayley graph of an Abelian group A with respect to a generating set S, one form of Plünnecke's inequality (see, e.g., [49]) says that the sequence

$$|\underbrace{S + S + \cdots + S}_{k}|^{1/k}$$

is decreasing in k. Since $S, S + S, S + S + S, \ldots$ are precisely the sets of vertices which can be reached by $1, 2, 3, \ldots$ steps from 0, this inequality quantifies the idea of constant-rate growth in the breadth-first search tree; Theorem 15 in [2] makes this argument formal.

In this chapter, we disprove Alon et al.'s conjecture by constructing distance-uniform graphs that do not share this behavior, and whose diameter is exponentially larger than these examples. We also prove an upper bound on the critical distance (and diameter) showing our construction to be best possible, in one asymptotic sense, for a wide range of ϵ. Specifically, we show the following two results:

Theorem 6.1.1. *In any ϵ-distance-uniform graph with n vertices and $\epsilon = 2^{-\Omega(\sqrt{\log n})}$, the critical distance d satisfies*

$$d = 2^{O(\frac{\log n}{\log \epsilon^{-1}})}.$$

Theorem 6.1.2. *For any ϵ and n with $\frac{1}{n} \leq \epsilon \leq \frac{1}{\log n}$, there exists an ϵ-distance-uniform graph on n vertices with critical distance*

$$d = 2^{\Omega\left(\frac{\log n}{\log \epsilon^{-1}}\right)}.$$

Combined, these results prove that the maximum critical distance is $2^{\Theta(\frac{\log n}{\log \epsilon^{-1}})}$ whenever they both apply. Unfortunately, we were unable to extend this upper bound on d to large values of

ϵ: for example, when ϵ is constant as $n \to \infty$. Of course, since a $\frac{1}{\log n}$-distance-uniform graph is also $\frac{1}{2}$-distance-uniform, Theorem 6.1.2 provides a lower bound of $d = 2^{\Omega(\frac{\log n}{\log \log n})}$ for any $\epsilon > \frac{1}{\log n}$.

The family of graphs used to prove Theorem 6.1.2 is interesting in its own right. We give two different interpretations of the underlying structure of these graphs. First, we describe a combinatorial game, generalizing the well-known Tower of Hanoi puzzle, whose transition graph is ϵ-distance-uniform and has large diameter. Second, we give a geometric interpretation, under which each graph in the family is the skeleton of the convex hull of an arrangement of points on a high-dimensional sphere.

6.2 Upper bound

Before proceeding to the proof of Theorem 6.1.1, we begin with a simple argument that is effective for an ϵ which is very small:

Proposition 6.2.1. *The minimum degree $\delta(G)$ of an ϵ-distance-uniform graph G satisfies $\delta(G) \geq \epsilon^{-1} - 1$.*

Proof. Suppose that G is ϵ-distance-uniform, n is the number of vertices of G, and d is the critical distance: for any vertex v, at least $(1-\epsilon)n$ vertices of G are at distance exactly d from v.

Let v be an arbitrary vertex of G, and fix an arbitrary breadth-first search tree T, rooted at v. We define the *score* of a vertex w (relative to T) to be the number of vertices at distance d from v which are descendants of w in the tree T.

There are at least $(1-\epsilon)n$ vertices at distance d from v, and all of them are descendants of some vertex in the neighborhood $N(v)$. Therefore the total score of all vertices in $N(v)$ is at least $(1-\epsilon)n$.

On the other hand, if $w \in N(v)$, each vertex counted by the score of w is at distance $d-1$ from w. Since at least $(1-\epsilon)n$ vertices are at distance d from w, at most ϵn vertices are at

distance $d-1$, and therefore the score of w is at most ϵn.

In order for $|N(v)|$ scores of at most ϵn to sum to at least $(1-\epsilon)n$, $|N(v)|$ must be at least $\frac{(1-\epsilon)n}{\epsilon n} = \epsilon^{-1} - 1$. □

This proposition is enough to show that in a $\frac{1}{\sqrt{n}}$-distance-uniform graph, the critical distance is at most 2. Choose a vertex v: all but $\sqrt{n}$ of the vertices of G are at the critical distance d from v, and $\sqrt{n}-1$ of the vertices are at distance 1 from v by Proposition 6.2.1. The remaining uncounted vertex is v itself. It's impossible to have $d \geq 3$, as that would leave no vertices at distance 2 from v.

For larger ϵ, the bound of Proposition 6.2.1 becomes ineffective, but we can improve in it by iterating the same argument.

Proposition 6.2.2. *Let $a_k = \frac{4^k-1}{3}$, satisfying $a_1 = 1$ and $a_{k+1} = 4a_k + 1$.*

In an ϵ-distance-uniform G, for each vertex v, and for any k such that a_k is less than the critical distance d, the number of vertices within distance at most a_k from v is at least $2^{-k^2}(\epsilon^{-1} - 1)^k$.

Proof. We induct on k. Proposition 6.2.1 provides the $k = 1$ case of this theorem.

Suppose the proposition holds for some value of k, and that $a_{k+1} < d$. Once again, we choose an arbitrary vertex v and an arbitrary breadth-first search tree T, rooted at v, and assign each vertex w a score relative to T.

Next, we assign each vertex w a *metascore* equal to the sum of the scores of all vertices within distance a_k from w. Suppose w is at a distance of at least $2a_k + 1$ from v. Then taking up to a_k steps from w and then following the shortest path in T to depth d will yield a vertex at distance $d-1$ or less from w. Therefore, at most ϵn vertices of G are descendants of a vertex within distance a_k from w.

This does not, however, mean that the metascore of w is at most ϵn: these ϵn vertices can be counted multiple times when they are descendants of several vertices within distance a_k from w.

Each vertex at depth d has at most $2a_k + 1$ ancestors within distance a_k from w. Taking this into account, we conclude that the metascore of w is at most $(2a_k + 1)\epsilon n$.

Now sum the metascores of all vertices at a distance between $2a_k+1$ and $a_{k+1} = 4a_k+1$ from v. If x is a vertex at distance $3a_k + 1$ from v, then the score of x is included in the metascore of at least $2^{-k^2}(\epsilon^{-1} - 1)^k$ vertices; the scores of all vertices at distance $3a_k + 1$ add to at least $(1-\epsilon)n$, and therefore the total of all metascores we have summed is at least $2^{-k^2}(\epsilon^{-1} - 1)^k(1 - \epsilon)n$.

Since each vertex contributes at most $(2a_k + 1)\epsilon n$ to this total, we conclude that there are at least

$$\frac{2^{-k^2}(\epsilon^{-1} - 1)^k(1 - \epsilon)n}{(2a_k + 1)\epsilon n} = \frac{2^{-k^2}}{2a_k + 1}(\epsilon^{-1} - 1)^{k+1} > \frac{2^{-k^2}}{4^k}(\epsilon^{-1} - 1)^{k+1} > 2^{-(k+1)^2}(\epsilon^{-1} + 1)^{k+1}$$

vertices at a distance between $2a_k + 1$ and $4a_k + 1$, and therefore within distance a_{k+1}, from v. □

The result of Proposition 6.2.2 is, in particular, a lower bound on the number of vertices in G, in terms of d and ϵ, which we can hope to reverse to get an upper bound on d in terms of ϵ and n. The following propositon makes this argument precise.

Proposition 6.2.3. *In an ϵ-distance-uniform graph with critical distance d, the number of vertices is at least*

$$\min\left\{2^{\Omega((\log \epsilon^{-1})^2)}, 2^{\Omega(\log d \log \epsilon^{-1})}\right\}.$$

Proof. Let G be ϵ-distance-uniform with critical distance d.

For any k with $a_k < d$, Proposition 6.2.2 implies that G has at least $2^{-k^2}(\epsilon^{-1} - 1)^k$ vertices; this is maximized when $k = \log_4(\epsilon^{-1} - 1)$. We will choose some value of k such that $k \leq \log_4(\epsilon^{-1} - 1)$; then $2^{-k^2} > 2^{-k\log_4(\epsilon^{-1}-1)} = (\epsilon^{-1} - 1)^{-k/2}$, so $2^{-k^2}(\epsilon^{-1} - 1)^k > (\epsilon^{-1} - 1)^{k/2} = 2^{k\log_4(\epsilon^{-1}-1)}$, and the guarantee of Proposition 6.2.2 is that G has at least $2^{k\log_4(\epsilon^{-1}-1)}$ vertices.

Let k_1 be the value of k that satisfies $\frac{4^{k_1}-1}{3} < d \leq \frac{4^{k_1+1}-1}{3}$: k_1 is the largest value of k that can be chosen in Proposition 6.2.2, since $a_{k_1} < d \leq a_{k_1+1}$, and satisfies $k_1 = \log_4 d - O(1)$.

If $k_1 \leq \log_4(\epsilon^{-1} - 1)$, then it makes sense to use $k = k_1$ in the proposition. In this case, we conclude that G has at least $2^{k_1 \log_4(\epsilon^{-1}-1)} = 2^{\Omega(\log d \log \epsilon^{-1})}$ vertices.

On the other hand, if $k_1 > \log_4(\epsilon^{-1} - 1)$, then we achieve better results by using $k_2 = \lfloor \log_4(\epsilon^{-1} - 1) \rfloor$ in the proposition: for larger k, the bound we get on n will start decreasing. In this case, we conclude that G has at least $2^{k_2 \log_4(\epsilon^{-1}-1)} = 2^{\Omega((\log \epsilon^{-1})^2)}$ vertices. □

To obtain the upper bound in Theorem 6.1.1, we assume that ϵ is small enough that the first bound of Proposition 6.2.3 is less than n: the upper bound on ϵ we derive from this assumption has the form $\epsilon = 2^{-\Omega(\sqrt{\log n})}$.

In this case, the first bound of Proposition 6.2.3 does not hold; therefore the second bound must hold, and $n = 2^{\Omega(\log d \log \epsilon^{-1})}$. Solving for d, we conclude that $d = 2^{O(\frac{\log n}{\log \epsilon^{-1}})}$, as desired.

6.3 Lower bound

6.3.1 The Hanoi game

We define a *Hanoi state* to be a finite sequence of nonnegative integers $\vec{x} = (x_1, x_2, \ldots, x_k)$ such that, for all $i > 1$, $x_i \neq x_{i-1}$. Let

$$\mathcal{H}_{r,k} = \big\{\vec{x} \in \{0, 1, \ldots, r\}^k : \vec{x} \text{ is a Hanoi state}\big\}.$$

For convenience, we also define a *proper Hanoi state* to be a Hanoi state $\vec{x}$ with $x_1 \neq 0$, and $\mathcal{H}^*_{r,k} \subset \mathcal{H}_{r,k}$ to be the set of all proper Hanoi states. While everything we prove will be equally true for Hanoi states and proper Hanoi states, it's more convenient to work with $\mathcal{H}^*_{r,k}$, because $|\mathcal{H}^*_{r,k}| = r^k$.

In the *Hanoi game on* $\mathcal{H}_{r,k}$, an initial state $\vec{a} \in \mathcal{H}_{r,k}$ and a final state $\vec{b} \in \mathcal{H}_{r,k}$ are chosen. The state $\vec{a}$ must be transformed into $\vec{b}$ via a sequence of moves of two types:

1. An *adjustment* of $\vec{x} \in \mathcal{H}_{r,k}$ changes x_k to any value in $\{0, 1, \ldots, r\}$ other than x_{k-1}. For example, $(1, 2, 3, 4)$ can be changed to $(1, 2, 3, 0)$ or $(1, 2, 3, 5)$, but not $(1, 2, 3, 3)$.

2. An *involution* of $\vec{x} \in \mathcal{H}_{r,k}$ finds the longest tail segment of $\vec{x}$ on which the values x_k and x_{k-1} alternate, and swaps x_k with x_{k-1} in that segment. For example, $(1, 2, 3, 4)$ can be changed to $(1, 2, 4, 3)$, or $(1, 2, 1, 2)$ to $(2, 1, 2, 1)$.

We define the Hanoi game on $\mathcal{H}^*_{r,k}$ in the same way, but with the added requirement that all states involved should be proper Hanoi states. This means that involutions (or, in the case of $k = 1$, adjustments) that would change x_1 to 0 are forbidden.

The name "Hanoi game" is justified because its structure is similar to the structure of the classical Tower of Hanoi puzzle. In fact, though we have no need to prove this, the Hanoi game on $\mathcal{H}^*_{3,k}$ is isomorphic to a Tower of Hanoi puzzle with k disks.

It's well-known that the k-disk Tower of Hanoi puzzle can be solved in $2^k - 1$ moves, moving a stack of k disks from one peg to another. In [32], a stronger statement is shown: only $2^k - 1$ moves are required to go from any initial state to any final state. A similar result holds for the Hanoi game on $\mathcal{H}_{r,k}$:

Lemma 6.3.1. *The Hanoi game on $\mathcal{H}_{r,k}$ (or $\mathcal{H}^*_{r,k}$) can be solved in at most $2^k - 1$ moves for any initial state $\vec{a}$ and final state $\vec{b}$.*

Proof. We induct on k to show the following stronger statement: for any initial state $\vec{a}$ and final state $\vec{b}$, a solution of length at most $2^k - 1$ exists for which any intermediate state $\vec{x}$ has $x_1 = a_1$ or $x_1 = b_1$. This auxiliary condition also means that if $\vec{a}, \vec{b} \in \mathcal{H}^*_{r,k}$, all intermediate states will also stay in $\mathcal{H}^*_{r,k}$.

When $k = 1$, a single adjustment suffices to change $\vec{a}$ to $\vec{b}$, which satisfies the auxiliary condition.

For $k > 1$, there are two possibilities when changing $\vec{a}$ to $\vec{b}$:

- If $a_1 = b_1$, then consider the Hanoi game on $\mathcal{H}_{r,k-1}$ with initial state $(a_2, a_3, \ldots, a_k)$ and final state $(b_2, b_3, \ldots, b_k)$. By the inductive hypothesis, a solution using at most $2^{k-1} - 1$ moves exists.

  Apply the same sequence of adjustments and involutions in $\mathcal{H}_{r,k}$ to the initial state $\vec{a}$. This

has the effect of changing the last $k-1$ entries of $\vec{a}$ to $(b_2, b_3, \ldots, b_k)$. To check that we've obtained $\vec{b}$, we need to verify that the first entry is left unchanged.

The auxiliary condition of the inductive hypothesis tells us that all intermediate states have $x_2 = a_2$ or $x_2 = b_2$. Any move that leaves x_2 unchanged also leaves x_1 unchanged. A move that changes x_2 must be an involution swapping the values a_2 and b_2; however, $x_1 = a_1 \neq a_2$, and $x_1 = b_1 \neq b_2$, so such an involution also leaves x_1 unchanged.

Finally, the new auxiliary condition is satisfied, since we have $x_1 = a_1 = b_1$ for all intermediate states.

- If $a_1 \neq b_1$, begin by taking $2^{k-1} - 1$ moves to change $\vec{a}$ to $(a_1, b_1, a_1, b_1, \ldots)$ while satisfying the auxiliary condition, as in the first case.

  An involution takes this state to $(b_1, a_1, b_1, a_1, \ldots)$; this continues to satisfy the auxiliary condition.

  Finally, $2^{k-1} - 1$ more moves change this state to $\vec{b}$, as in the first case, for a total of $2^k - 1$ moves. □

If we obtain the same results as in the standard Tower of Hanoi puzzle, why use the more complicated game in the first place? The reason is that in the classical problem, we cannot guarantee that any starting state would have a final state $2^k - 1$ moves away. With the rules we define, as long as the parameters are chosen judiciously, each state $\vec{a} \in \mathcal{H}_{r,k}$ is part of many pairs $(\vec{a}, \vec{b})$ for which the Hanoi game requires $2^k - 1$ moves to solve.

The following lemma almost certainly does not characterize such pairs, but provides a simple sufficient condition that's strong enough for our purposes.

Lemma 6.3.2. *The Hanoi game on $\mathcal{H}_{r,k}$ (or $\mathcal{H}^*_{r,k}$) requires exactly $2^k - 1$ moves to solve if $\vec{a}$ and $\vec{b}$ are chosen with disjoint support: that is, $a_i \neq b_j$ for all i and j.*

Proof. Since Lemma 6.3.1 proved an upper bound of $2^k - 1$ for all pairs $(\vec{a}, \vec{b})$, we only need to prove a lower bound in this case.

Once again, we induct on k. When $k = 1$, a single move is necessary to change $\vec{a}$ to $\vec{b}$ if $\vec{a} \neq \vec{b}$, verifying the base case.

Consider a pair $\vec{a}, \vec{b} \in \mathcal{H}_{r,k}$ with disjoint support, for $k > 1$. Moreover, assume that $\vec{a}$ and $\vec{b}$ are chosen so that, of all pairs with disjoint support, $\vec{a}$ and $\vec{b}$ require the least number of moves to solve the Hanoi game. (Since we are proving a lower bound on the number of moves necessary, this assumption is made without loss of generality.)

In a shortest path from $\vec{a}$ to $\vec{b}$, every other move is an adjustment: if there were two consecutive adjustments, the first adjustment could be skipped, and if there were two consecutive involutions, they would cancel out and both could be omitted. Moreover, the first move is an adjustment: if we began with an involution, then the involution of $\vec{a}$ would be a state closer to $\vec{b}$ yet still with disjoint support to $\vec{b}$, contrary to our initial assumption. By the same argument, the last move must be an adjustment.

Given a state $\vec{x} \in \mathcal{H}_{r,k}$, let its *abbreviation* be $\vec{x}' = (x_1, x_2, \ldots, x_{k-1}) \in \mathcal{H}_{r,k-1}$. An adjustment of $\vec{x}$ has no effect on $\vec{x}'$, since only x_k is changed. If $x_k \neq x_{k-2}$, then an involution of $\vec{x}$ is an adjustment of $\vec{x}'$, changing its last entry x_{k-1} to x_k. Finally, if $x_k = x_{k-2}$, then an involution of $\vec{x}$ is also an involution of $\vec{x}'$.

Therefore, if we take a shortest path from $\vec{a}$ to $\vec{b}$, omit all adjustments, and then abbreviate all states, we obtain a solution to the Hanoi game on $\mathcal{H}_{r,k-1}$ that takes $\vec{a}'$ to $\vec{b}'$. By the inductive hypothesis, this solution contains at least $2^{k-1} - 1$ moves, since $\vec{a}'$ and $\vec{b}'$ have disjoint support. Therefore the shortest path from $\vec{a}$ to $\vec{b}$ contains at least $2^{k-1} - 1$ involutions. Since the first, last, and every other move is an adjustment, there must be 2^{k-1} adjustments as well, for a total of $2^k - 1$ moves. □

Now let the *Hanoi graph* $G^*_{r,k}$ be the graph with vertex set $\mathcal{H}^*_{r,k}$ and edges joining each state to all the states that can be obtained from it by a single move. Since an adjustment can be reversed by another adjustment, and an involution is its own inverse, $G^*_{r,k}$ is an undirected graph.

For any state $\vec{a} \in \mathcal{H}^*_{r,k}$, there are at least $(r-k)^k$ other states with disjoint support to $\vec{a}$, out of

$|\mathcal{H}^*_{r,k}| = r^k$ other states, forming a $\left(1 - \frac{k}{r}\right)^k > 1 - \frac{k^2}{r}$ fraction of all the states. By Lemma 6.3.2, each such state $\vec{b}$ is at distance $2^k - 1$ from $\vec{a}$ in the graph $G^*_{r,k}$, so $G^*_{r,k}$ is ϵ-distance uniform with $\epsilon = \frac{k^2}{r}$, $n = r^k$ vertices, and critical distance $d = 2^k - 1$.

Having established the graph-theoretic properties of $G^*_{r,k}$, we now prove Theorem 6.1.2 by analyzing the asymptotic relationship between these parameters.

Proof of Theorem 6.1.2. Begin by assuming that $n = 2^{2^m}$ for some m. Choose a and b such that $a + b = m$ and

$$\frac{2^{2b}}{2^{2^a}} \leq \epsilon < \frac{2^{2(b+1)}}{2^{2^{a-1}}},$$

which is certainly possible since $\frac{2^0}{2^{2^m}} = \frac{1}{n} \leq \epsilon$ and $\frac{2^{2m}}{2^{2^0}} > 1 \geq \epsilon$. Setting $r = 2^{2^a}$ and $k = 2^b$, the Hanoi graph $G^*_{r,k}$ has n vertices and is ϵ-distance uniform, since $\frac{k^2}{r} \leq \epsilon$. Moreover, our choice of a and b guarantees that $\epsilon < \frac{4k^2}{\sqrt{r}}$, or $\log \epsilon^{-1} \geq \frac{1}{2} \log r - 2 \log 2k$. Since $n = r^k$, $\log n = k \log r$, so

$$\log \epsilon^{-1} \geq \frac{1}{2k} \log n - 2 \log 2k.$$

We show that $k \geq \frac{\log n}{6 \log \epsilon^{-1}}$. Since $\epsilon \leq \frac{1}{\log n}$, this is automatically true if $k \geq \frac{\log n}{6 \log \log n}$, so assume that $k < \frac{\log n}{6 \log \log n}$. Then

$$\frac{1}{3k} \log n > 2 \log \log n > 2 \log 2k,$$

so

$$\log \epsilon^{-1} \geq \frac{1}{2k} \log n - 2 \log 2k > \frac{1}{2k} \log n - \frac{1}{3k} \log n = \frac{1}{6k} \log n,$$

which gives us the desired inequality $k \geq \frac{\log n}{6 \log \epsilon^{-1}}$. The Hanoi graph $G^*_{r,k}$ has critical distance $d = 2^k - 1 = 2^{\Omega(\frac{\log n}{\log \epsilon^{-1}})}$, so the proof is finished in the case that n has the form 2^{2^m} for some n.

For a general n, we can choose m such that $2^{2^m} \leq n < 2^{2^{m+1}} = \left(2^{2^m}\right)^2$, which means in particular that $2^{2^m} \geq \sqrt{n}$. If $\epsilon < \frac{2}{\sqrt{n}}$, then the requirement of a critical distance of $2^{\Omega(\frac{\log n}{\log \epsilon^{-1}})}$ is only a constant lower bound, and we may take the graph K_n. Otherwise, by the preceding argument, there is a $\frac{\epsilon}{2}$-distance-uniform Hanoi graph with 2^{2^m} vertices; its critical distance d satisfies

$$d \geq 2^{\Omega\left(\frac{\log \sqrt{n}}{\log(\epsilon/2)^{-1}}\right)} = 2^{\Omega\left(\frac{\log n}{\log \epsilon^{-1}}\right)}.$$

To extend this to an n-vertex graph, take the blow-up of the 2^{2^m}-vertex Hanoi graph, replacing every vertex by either $\lfloor n/2^{2^m} \rfloor$ or $\lceil n/2^{2^m} \rceil$ copies.

Whenever v and w were at distance d in the original graph, the copies of v and w will be at distance d in the blow-up. The difference between floor and ceiling may slightly ruin distance uniformity, but the graph started out $\frac{\epsilon}{2}$-distance-uniform, and $\lceil n/2^{2^m} \rceil$ differs from $\lfloor n/2^{2^m} \rfloor$ at most by a factor of 2. Even in the worst case, where for some vertex v the $\frac{\epsilon}{2}$-fraction of vertices not at distance d from v all receive the larger number of copies, the resulting n-vertex graph will be ϵ-distance-uniform. □

6.3.2 Points on a sphere

In this section, we identify $G_{r,k}$, the graph of the Hanoi game on $\mathcal{H}_{r,k}$, with a graph that arises from a geometric construction.

Fix a dimension r. We begin by placing $r+1$ points on the r-dimensional unit sphere arbitrarily in general position (though, for the sake of symmetry, we may place them at the vertices of an equilateral r-simplex). We identify these points with a graph by taking the 1-skeleton of their convex hull. In this starting configuration, we simply get K_{r+1}.

Next, we define a truncation operation on a set of points on the r-sphere. Let $\delta > 0$ be sufficiently small that a sphere of radius $1-\delta$, concentric with the unit sphere, intersects each edge of the 1-skeleton in two points. The set of these intersection points is the new arrangement of points obtained by the truncation; they all lie on the smaller sphere, and for convenience, we may scale them so that they are once again on the unit sphere.

Proposition 6.3.1. *Starting with a set of $r+1$ points on the r-dimensional sphere and applying k truncations produces a set of points such that the 1-skeleton of their convex hull is isomorphic to the graph $G_{r,k}$.*

Proof. We induct on k. When $k=1$, the graph we get is K_{r+1}, which is isomorphic to $G_{r,1}$.

From the geometric side, we add an auxiliary statement to the induction hypothesis: given

points p, q_1, q_2 such that, in the associated graph, p is adjacent to both q_1 and q_2, there is a 2-dimensional face of the convex hull containing all three points. This is easily verified for $k = 1$.

Assuming that the induction hypotheses are true for $k-1$, fix an isomorphism of $G_{r,k-1}$ with the set of points after $k-1$ truncations, and label the points with the corresponding vertices of $G_{r,k-1}$. We claim that the graph produced after one more truncation has the following structure:

1. A vertex that we may label $(\vec{x}, \vec{y})$ for every ordered pair of adjacent vertices of $G_{r,k-1}$.
2. An edge between $(\vec{x}, \vec{y})$ and $(\vec{y}, \vec{x})$.
3. An edge between $(\vec{x}, \vec{y})$ and $(\vec{x}, \vec{z})$ whenever both are vertices of the new graph.

The first claim is immediate from the definition of truncation: we obtain two vertices from the edge between $\vec{x}$ and $\vec{y}$. We choose to give the name $(\vec{x}, \vec{y})$ to the vertex closer to $\vec{x}$. The edge between $\vec{x}$ and $\vec{y}$ remains an edge, and now joins the vertices $(\vec{x}, \vec{y})$ and $(\vec{y}, \vec{x})$, verifying the second claim.

By the auxiliary condition of the induction hypothesis, the vertices labeled $\vec{x}$, $\vec{y}$, and $\vec{z}$ lie on a common 2-face whenever $\vec{x}$ is adjacent to both $\vec{y}$ and $\vec{z}$. After truncation, $(\vec{x}, \vec{y})$ and $(\vec{x}, \vec{z})$ will also be on this 2-face; since they are adjacent along the boundary of that face, and extreme points of the convex hull, they are joined by an edge, verifying the third claim.

To finish the geometric part of the proof, we verify that the auxiliary condition remains true. There are two cases to check. For a vertex labeled $(\vec{x}, \vec{y})$, if we choose the neighbors $(\vec{x}, \vec{z})$ and $(\vec{x}, \vec{w})$, then any two of them are joined by an edge, and therefore they must lie on a common 2-dimensional face. If we choose the neighbors $(\vec{x}, \vec{z})$ and $(\vec{y}, \vec{x})$, then the points continue to lie on the 2-dimensional face inherited from the face through $\vec{x}$, $\vec{y}$, and $\vec{z}$ of the previous convex hull.

Now it remains to construct an isomorphism between the 1-skeleton graph of the truncation, which we'll call T, and $G_{r,k}$. We identify the vertex $(\vec{x}, \vec{y})$ of T with the vertex $(x_1, x_2, \ldots, x_{k-1}, y_{k-1})$ of $G_{r,k}$. Since $x_{k-1} \neq y_{k-1}$ after any move in the Hanoi game, this k-tuple really is a Hanoi state. Conversely, any Hanoi state $\vec{z} \in \mathcal{H}_{r,k}$ corresponds to a vertex of T: let $\vec{x} = (z_1, z_2, \ldots, z_{k-1})$, and let $\vec{y}$ be the state obtained from $\vec{x}$ by either an adjustment of z_{k-1} to z_k, if $z_k \neq z_{k-2}$, or else

an involution, if $z_k = z_{k-2}$. Therefore the map we define is a bijection between the vertex sets.

Both T and $G_{r,k}$ are r-regular graphs, therefore it suffices to show that each edge of T corresponds so an edge in $G_{r,k}$. Consider an edge joining $(\vec{x}, \vec{y})$ with $(\vec{x}, \vec{z})$ in T. This corresponds to vertices $(x_1, x_2, \ldots, x_{k-1}, y_{k-1})$ and $(x_1, x_2, \ldots, x_{k-1}, z_{k-1})$ in $G_{r,k}$; these are adjacent, since we can obtain one from the other by an adjustment.

Next, consider an edge joining $(\vec{x}, \vec{y})$ to $(\vec{y}, \vec{x})$. If $\vec{x}$ and $\vec{y}$ are related by an adjustment in $G_{r,k-1}$, then they have the form $(x_1, \ldots, x_{k-2}, x_{k-1})$ and $(x_1, \ldots, x_{k-2}, y_{k-1})$. The vertices corresponding to $(\vec{x}, \vec{y})$ and $(\vec{y}, \vec{x})$ in $G_{r,k}$ are $(x_1, \ldots, x_{k-2}, x_{k-1}, y_{k-1})$ and $(x_1, \ldots, x_{k-2}, y_{k-1}, x_{k-1})$, and one can be obtained from the other by an involution.

Finally, if $\vec{x}$ and $\vec{y}$ are related by an involution in $G_{r,k-1}$, then that involution swaps x_{k-1} and y_{k-1}. Therefore such an involution in $G_{r,k}$ will take $(x_1, \ldots, x_{k-1}, y_{k-1})$ to $(y_1, \ldots, y_{k-1}, x_{k-1})$, and the vertices corresponding to $(\vec{x}, \vec{y})$ and $(\vec{y}, \vec{x})$ are adjacent in $G_{r,k}$. $\square$

Chapter 7

Conclusion

I began this thesis with a discussion of combinatorial results proven by other mathematicians nearly a century ago. It seems appropriate to end it with a discussion of the problems on which I hope to see progress—and which I intend to work on personally—in the future.

The results of Chapter 3 prove improved upper bounds on the Hales–Jewett number $\mathrm{HJ}(4, 2)$. A more ambitious goal of mine is to achieve improved upper bounds on $\mathrm{HJ}(t, 2)$ for arbitrary t, extending this result. Recent work on the Hales–Jewett theorem has focused on density results, which strengthen the statement at the cost of returning to Ackermann-type bounds on the dimension n. On the other hand, our approach avoids iterated use of the pigeonhole principle; if the challenges that currently prevent generalization from $t = 4$ can be solved, it is unlikely that the resulting bound would grow much faster than the exponentially-growing lower bound.

I am also interested in making progress on the deterministic side of the Chvátal and Komlós problem mentioned in Chapter 5. This problem asks to determine the largest $m(n)$ such that every edge-ordering of the edges of K_n contains an increasing path of length at least $m(n)$. This problem has remained open for decades, and the last few years have seen a burst of progress: in addition to already-mentioned work by Milans [42] and Martinsson [41], a recent result of de Silva et al. [16] extends the problem to the hypercube graph and to the random graph $\mathcal{G}_{n,p}$.

I could go on discussing my hopes for problems already discussed in this thesis. However,

what I look forward to the most is something that, by definition, I cannot describe. It is the certainty that no matter which directions the field of combinatorics proceeds in, it will open the way for new questions that are more exciting than anything we have already thought about.

Bibliography

[1] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, Hoboken, NJ, 2008. 3.3.2, 5.1

[2] Noga Alon, Erik D. Demaine, Mohammad T. Hajiaghayi, and Tom Leighton. Basic network creation games. *SIAM J. Discrete Math.*, 27(2):656–668, 2013. (document), 6.1

[3] J. Barkley. Improved lower bound on an Euclidean Ramsey problem. 2008. arXiv:0811.1055. 2.1.2

[4] Tomasz Bartnicki, Jarosław Grytczuk, H. A. Kierstead, and Xuding Zhu. The map-coloring game. *Amer. Math. Monthly*, 114(9):793–803, 2007. 4.1

[5] E. R. Berlekamp. A construction for partitions which avoid long arithmetic progressions. *Canad. Math. Bull.*, 11:409–414, 1968. 3.1

[6] Hans L. Bodlaender. On the complexity of some coloring games. *Internat. J. Found. Comput. Sci.*, 2(2):133–147, 1991. 4.1

[7] Tom Bohman, Alan Frieze, and Benny Sudakov. The game chromatic number of random graphs. *Random Structures Algorithms*, 32(2):223–235, 2008. 4.1

[8] B. Bollobás. *Random graphs*. Academic Press, London, 1985. 5.1

[9] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1(4):311–316, 1980. 4.1

[10] Béla Bollobás. The diameter of random graphs. *Trans. Amer. Math. Soc.*, 267(1):41–52,

1981. 1.2, 6.1

[11] A. R. Calderbank, F. R. K. Chung, and D. G. Sturtevant. Increasing sequences with nonzero block sums and increasing paths in edge-ordered graphs. *Discrete Math.*, 50(1):15–28, 1984. 5.1

[12] Timothy J. Carlson, Neil Hindman, and Dona Strauss. An infinitary extension of the Graham-Rothschild parameter sets theorem. *Trans. Amer. Math. Soc.*, 358(7):3239–3262, 2006. 2.1.2

[13] V. Chvátal. Some unknown van der Waerden numbers. In *Combinatorial Structures and their Applications (Proc. Calgary Internat. Conf., Calgary, Alta., 1969)*, pages 31–33. Gordon and Breach, New York, 1970. 3.1

[14] V. Chvátal and J. Komlós. Some combinatorial theorems on monotonicity. *Canad. Math. Bull.*, 14:151–157, 1971. 5.1

[15] C. Cooper and A. M. Frieze. On the number of Hamilton cycles in a random graph. *J. Graph Theory*, 13(6):719–735, 1989. 5.1

[16] Jessica De Silva, Theodore Molla, Florian Pfender, Troy Retter, and Michael Tait. Increasing paths in edge-ordered graphs: the hypercube and random graph. *Electron. J. Combin.*, 23(2):Paper 2.15, 9, 2016. 7

[17] P. Erdös. Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.*, 53:292–294, 1947. 1.2

[18] P. Erdős and G. Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935. 5.1

[19] Geoffrey Exoo. A Euclidean Ramsey problem. *Discrete Comput. Geom.*, 29(2):223–227, 2003. 2.1.2

[20] Alex Fabrikant, Ankur Luthra, Elitza Maneva, Christos H. Papadimitriou, and Scott Shenker. On a network creation game. In *Proceedings of the twenty-second annual sympo-*

sium on Principles of distributed computing, pages 347–351. ACM, 2003. 6.1

[21] J. Fox, J. Pach, B. Sudakov, and A. Suk. Erdős–Szekeres-type theorems for monotone paths and convex bodies. *Proc. Lond. Math. Soc. (3)*, 105(5):953–982, 2012. 5.1

[22] Alan Frieze, Simcha Haber, and Mikhail Lavrov. On the game chromatic number of sparse random graphs. *SIAM J. Discrete Math.*, 27(2):768–790, 2013. 1.3, 4.1

[23] Martin Gardner. Mathematical games. *Scientific American*, 244(4). 4.1

[24] Martin Gardner. In which joining sets of points leads into diverse (and diverting) paths. *Sci. Amer.*, 237(5):18–28, 1977. 2.1.1

[25] R. Glebov and M. Krivelevich. On the number of Hamilton cycles in sparse random graphs. *SIAM J. Discrete Math.*, 27(1):27–42, 2013. 5.1

[26] S. W. Golomb and P. Gaal. On the number of permutations of n objects with greatest cycle length k. *Adv. in Appl. Math.*, 20(1):98–107, 1998. 5.3.4

[27] R. L. Graham and D. J. Kleitman. Increasing paths in edge ordered graphs. *Period. Math. Hungar.*, 3:141–148, 1973. Collection of articles dedicated to the memory of Alfréd Rényi, II. (document), 1.2, 5.1

[28] R. L. Graham and B. L. Rothschild. Ramsey's theorem for n-parameter sets. *Trans. Amer. Math. Soc.*, 159:257–292, 1971. (document), 1.1, 2.1.1, 2.1.2, 2.2.3, 2.2.1

[29] R. L. Graham, B. L. Rothschild, and J. H. Spencer. *Ramsey Theory*. Wiley, New York, 1990. 2.2.3

[30] A. W. Hales and R. I. Jewett. Regularity and positional games. *Trans. Amer. Math. Soc.*, 106:222–229, 1963. 1.1, 3.1

[31] Neil Hindman and Eric Tressler. The first nontrivial Hales-Jewett number is four. *Ars Combin.*, 113:385–390, 2014. 3.1

[32] Andreas M. Hinz. Shortest paths between regular states of the Tower of Hanoi. *Inform.*

Sci., 63(1-2):173–181, 1992. 6.3.1

[33] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, New York, 2000. 5.1

[34] K. Kalmanson. On a theorem of Erdős and Szekeres. *J. Combinatorial Theory Ser. A*, 15:343–346, 1973. 5.1

[35] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *STACS 99 (Trier)*, volume 1563 of *Lecture Notes in Comput. Sci.*, pages 404–413. Springer, Berlin, 1999. 6.1

[36] Mikhail Lavrov. An upper bound for the Hales-Jewett number $\mathrm{HJ}(4, 2)$. *SIAM J. Discrete Math.*, 30(2):1333–1342, 2016. 1.3

[37] Mikhail Lavrov, Mitchell Lee, and John Mackey. Improved upper and lower bounds on a geometric Ramsey problem. *European J. Combin.*, 42:135–144, 2014. 1.3

[38] Mikhail Lavrov and Po-Shen Loh. Increasing Hamiltonian paths in random edge orderings. *Random Structures Algorithms*, 48(3):588–611, 2016. 1.3

[39] Mikhail Lavrov and Po-Shen Loh. Distance-uniform graphs with large diameter. *arXiv preprint arXiv:1703.01477*, 2017. 1.3

[40] B. F. Logan and L. A. Shepp. A variational problem for random Young tableaux. *Advances in Math.*, 26(2):206–222, 1977. 5.1

[41] Anders Martinsson. Most edge-orderings of K_n have maximal altitude. *arXiv preprint arXiv:1605.07204*, 2016. 5.1.1, 5.1.3, 7

[42] Kevin Milans. Monotone paths in dense edge-ordered graphs. *arXiv preprint arXiv:1509.02143*, 2015. 1.2, 5.1, 7

[43] Guy Moshkovitz and Asaf Shapira. Ramsey theory, integer partitions and a new proof of the Erdős-Szekeres theorem. *Adv. Math.*, 262:1107–1129, 2014. 5.1

[44] F. P. Ramsey. On a Problem of Formal Logic. *Proc. London Math. Soc.*, S2-30(1):264. 1.1

[45] R. W. Robinson and N. C. Wormald. Almost all regular graphs are Hamiltonian. *Random*

Structures Algorithms, 5(2):363–374, 1994. 4.1

[46] Saharon Shelah. Primitive recursive bounds for van der Waerden numbers. *J. Amer. Math. Soc.*, 1(3):683–697, 1988. (document), 2.1.1, 2.1.2, 2.2.3, 2.5, 2.5, 3.1

[47] M. Steele. Variations on the monotone subsequence theme of Erdős and Szekeres. In *Discrete probability and algorithms*, pages 111–131. Springer, 1995. 5.1

[48] Tibor Szabó and Gábor Tardos. A multidimensional generalization of the Erdős-Szekeres lemma on monotone subsequences. *Combin. Probab. Comput.*, 10(6):557–565, 2001. 5.1

[49] Terence Tao and Van H. Vu. *Additive combinatorics*, volume 105 of *Cambridge studies in advanced mathematics*. Cambridge University Press, 2006. 6.1

[50] A. M. Veršik and S. V. Kerov. Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux. *Dokl. Akad. Nauk SSSR*, 233(6):1024–1027, 1977. 5.1

[51] P. Winkler. Puzzled: Delightful graph theory. *Commun. ACM*, 51(8):104–104, August 2008. 4.1

[52] P. Winkler. Puzzled: Solutions and sources. *Commun. ACM*, 51(9):103–103, September 2008. 4.1, 5.1

[53] J. Wolfowitz. Note on runs of consecutive elements. *Ann. Math. Statistics*, 15:97–98, 1944. 5.4.1

[54] N. C. Wormald. Models of random regular graphs. In *Surveys in combinatorics, 1999 (Canterbury)*, volume 267 of *London Math. Soc. Lecture Note Ser.*, pages 239–298. Cambridge Univ. Press, Cambridge, 1999. 4.1