Topics in Extremal and Probabilistic Combinatorics

by

Natalie C. Behague

A thesis submitted to the University of London for the degree of Doctor of Philosophy

School of Mathematical Sciences
Queen Mary, University of London
United Kingdom

May 2020

Statement of Originality

I, Natalie C. Behague, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Natalie Behague

1st June 2020

Details of collaboration and publications:

The work in Chapter 2 was conducted in collaboration with Dr. Robert Johnson. All other chapters are solely my own work.

The results of Chapters 1 and 3 were previously published as follows:

1. Natalie C. Behague. Hypergraph Saturation Irregularities. *Electronic Journal of Combinatorics*, 25, 2018.

2. Natalie C. Behague. Semi-perfect 1-Factorizations of the Hypercube. *Discrete Mathematics*, 342(6), 2019.

Abstract

This thesis encompasses several problems in extremal and probabilistic combinatorics.

Chapter 1. Tuza's famous conjecture on the saturation number states that for r-uniform hypergraphs F the value $\mathrm{sat}(F,n)/n^{r-1}$ converges. I answer a question of Pikhurko concerning the asymptotics of the saturation number for families of hypergraphs, proving in particular that $\mathrm{sat}(\mathcal{F},n)/n^{r-1}$ need not converge if $\mathcal{F}$ is a family of r-uniform hypergraphs.

Chapter 2. Černý's conjecture on the length of the shortest reset word of a synchronizing automaton is arguably the most long-standing open problem in the theory of finite automata. We consider the minimal length of a word that resets some k-tuple. We prove that for general automata if this is finite then it is $\Theta\left(n^{k-1}\right)$. For synchronizing automata we improve the upper bound on the minimal length of a word that resets some triple.

Chapter 3. The existence of perfect 1-factorizations has been studied for various families of graphs, with perhaps the most famous open problem in the area being Kotzig's conjecture which states that even-order complete graphs have a perfect 1-factorization. In my work I focus on another well-studied family of graphs: the hypercubes. I answer almost fully the question of how close (in some particular sense) to perfect a 1-factorization of the hypercube can be.

Chapter 4. The k-nearest neighbour random geometric graph model puts vertices randomly in a d-dimensional box and joins each vertex to its k nearest neighbours. I find significantly improved upper and lower bounds on the threshold for connectivity for the k-nearest neighbour graph in high dimensions.

Acknowledgments

This work was supported by an EPSRC doctoral studentship.

The most important thank you goes to my supervisor, Robert Johnson, who has been a purveyor of interesting problems, a sharp-eyed proof-reader, and a fount of useful advice, both mathematical and otherwise. Thank you for always encouraging me try just a little longer — a lot of this would never have been solved if I hadn't been made to corral my ideas when I thought I was at a dead end.

I am grateful to my second supervisor Mark Walters, without whom the final chapter of this thesis would not exist, for setting me off in a fruitful direction and for invaluable discussions on the topic.

Thank you also to David Ellis, Bill Jackson, Mark Jerrum, Justin Ward, Felix Fischer and the whole combinatorics group at Queen Mary for showing such an interest in my work. I always enjoyed our mini seminars, annual reviews and of course the CSG.

I would like to thank Ali, Ben, Oliver, Rhys, Will and all of my friends and colleagues among the PhD community at Queen Mary for innumerable interesting conversations both at lunchtime and down the pub. Thank you for the maths, the beer and the boardgames.

Thank you to the rest of my friends: to Ruairidh, Jovan, Ben, Róisin, Ash, Dan, John; to the Yogis, my literal and metaphorical tag team; to so many people I can't list you all. I could maybe have done it without you but it wouldn't have been nearly as much fun.

Finally, thank you to my Mum, Dad and sisters Kat and Emily for a lifetime of love and support. Most pertinently, thank you for taking me in and keeping me sane when a global crisis hit just as I was writing this thesis.

Table of Contents

List of Figures

List of Tables

Introduction

Extremal combinatorics is the study of combinatorial objects and the possible values their parameters can take. Combinatorial objects are discrete structures which are usually finite (though arbitrarily large), with examples including graphs, hypergraphs, set systems and automata. In some cases these structures might be randomly generated, which is one way a probabilistic element can come in.

A typical problem in extremal combinatorics asks how large or small some parameter can be, given a structural property the object satisfies. A simple example might be to ask what is the maximum number of edges a graph can have, given it contains no triangles. Often solutions to these problems require two complementary parts: an argument that shows how possible behaviour is limited, along with a construction to show that these limits can be attained.

Each of the four chapters of this thesis focuses on a different problem; as such, each chapter is self-contained. We start each chapter with an introduction to the problem in question and some background on the area, thus this introduction serves merely as a brief overview of what is to come and an observation of some commonalities between the problems.

Chapter 1: *Hypergraph Saturation Irregularities* looks at the asymptotics of the saturation number for families of hypergraphs.

For a fixed graph F, Turan's number $\mathrm{ex}(F, n)$ is the maximum number of edges in

an F-free graph on n vertices. In any maximal F-free graph, adding any new edge must create a copy of F as a subgraph, which inspires the following definition: we say a graph G is *F-saturated* if it does not contain any copies of F but adding any new edge creates some copy of F. Then Turan's number can be defined equivalently as the maximum number of edges in an F-saturated graph on n vertices.

The saturation number $\mathrm{sat}(F, n)$ is obtained by replacing maximum by minimum; that is, $\mathrm{sat}(F, n)$ is the minimum number of edges in an F-saturated graph on n vertices. It forms an interesting counterpoint to the Turan number — the saturation number is in many ways less well-behaved. For example, we know that the Turan density $\lim_{n\to\infty} \mathrm{ex}(F, n)/n^2$ exists. Tuza [31] conjectured that $\mathrm{sat}(F, n)/n$ must tend to a limit as n tends to infinity, but this conjecture is still open.

The definition of saturation extends naturally to families of graphs. Pikhurko [25] disproved a strengthening of Tuza's conjecture by finding a finite family $\mathcal{F}$ of graphs such that $\mathrm{sat}(\mathcal{F}, n)/n$ does not converge as n tends to infinity. Pikhurko then asked whether a similar behaviour can occur for families of r-uniform hypergraphs. We shall see in Chapter 1 that the answer to this question is yes.

Chapter 2: *Synchronizing Automata and Černý's Conjecture* concerns reset words of automata. An automaton consists of a finite set of states and a finite set of *transition functions*, which are functions from the set of states to itself. We are interested in the results of applying a sequence of transitions to the set of states – we call such a sequence of transitions a *word* of the automaton. We say that a word is a *reset word* if it sends every state to the same point, and we call an automaton *synchronizing* if it has a reset word.

Černý conjectured that if an automaton on n states is synchronizing then there exists a reset word of length at most $(n-1)^2$. This is arguably the most famous open problem in the theory of finite automata. Currently, the best upper bound known on the length of a minimal reset word is $\approx 0.1654n^3$ [27].

We focus our attention on the minimum length of a word synchronizing some k -set: that is, sending some set of k states to the same point. In Chapter 2 we improve the best known bounds on the minimum length of a word synchronizing a 3-set, 4-set and 5-set of any synchronizing automaton. Furthermore, we show that for a non-synchronizing automaton it could require a word of length $\Theta\left(n^{k-1}\right)$ to synchronize a k-set, which is the worst possible.

Chapter 3: *Semi-perfect 1-Factorizations of the Hypercube* looks at 1-factorizations of the hypercube. A 1-factorization of a graph is a partition of the edges of the graph into disjoint perfect matchings $M_1, M_2, \ldots, M_n$. We say that a 1-factorization $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$ is a perfect factorization if every pair $M_i \cup M_j$ with i, j distinct forms a Hamilton cycle. A 1-factorization $\mathcal{M}$ is called *semi-perfect* if $M_1 \cup M_i$ forms a Hamilton cycle for all $i \neq 1$.

The existence or non-existence of perfect 1-factorizations has been studied for various families of graphs, with perhaps the most famous open problem in the area being Kotzig's conjecture [18] which states that the complete graph K_{2n} has a perfect 1-factorization. We focus on another well-studied family of graphs: the hypercubes Q_d.

Craft [2] conjectured that for every integer $d \geq 2$ there is a semi-perfect 1-factorization of Q_d. This was proved independently by Gochev and Gotchev [14] and by Královič and Královič [19] in the case where d is odd, and settled for d even by Chitra and Muthusamy [8]. Gochev and Gotchev in fact went further and defined $\mathcal{M}$ to be k-semi-perfect if $M_i \cup M_j$ forms a Hamilton cycle for every $1 \leq i \leq k$ and $k+1 \leq j \leq d$. They proved that there is a k-semi-perfect factorization of Q_d whenever k and d are both even with $k < d$.

It turns out there is no perfect 1-factorization of the hypercube, which is a corollary of a result due to Laufer [20]. An analysis of Laufer's proof reveals that a k-semi-perfect factorization is the best we can hope for: the matchings must split into two classes with no two matchings in the same class forming a Hamilton cycle. In light of this

observation, the only remaining question is whether for any k and d there is a k-semi-perfect factorization of Q_d In Chapter 3 we answer this question in the affirmative for almost every pair k, d, with only the case $k = 3, d = 6$ left unresolved.

Finally, Chapter 4: *Connectivity of High Dimension k-Nearest-Neighbour Graphs* studies the threshold for connectivity of a particular random geometric graph model. The k-nearest neighbour random geometric graph $G = G(d, n, k)$ is defined as follows: take a d-dimensional cube of volume n and let $\mathcal{P}$ be a Poisson process of density 1 in the cube. Then put an edge between each point of $\mathcal{P}$ and its k nearest neighbours.

This graph has been well-studied in the 2-dimensional setting where the the threshold for connectivity is $\Theta(\log n)$ [3, 35]. Using very simple generalisations of the arguments in the 2-dimensional setting it is easy to show that for fixed d, the threshold for connectivity is still $\Theta(\log n)$. These arguments give weak bounds on how the coefficient of $\log n$ depends on d: if $k = \Omega\left(\frac{1}{\log d} \log n\right)$ then G is connected with high probability and if $k = O\left(\frac{1}{e^d} \log n\right)$ then G is disconnected with high probability.

Given the difference in terms of d between these bounds, one natural question is to ask how the threshold for connectivity depends on the dimension d. In Chapter 4 we improve the bounds substantially. Precisely, we show that if $k \geq \frac{2.467}{d} \log n$ then $G(d, n, k)$ is connected with high probability and if $k \leq \frac{0.102}{d \log d} \log n$ then $G(d, n, k)$ is disconnected with high probability. We also establish bounds on the threshold for connectivity of the similarly-defined directed graph $\overrightarrow{G}$.

All of these problems come from extremal combinatorics and as a result they share a common theme of examining the properties of potentially large discrete structures, whether they are graphs, hypergraphs or automata.

Beyond this, a common thread that emerges from the work is a use of intricate constructions, which are utilised in a variety of ways. In Chapter 1 the constructions are counterexamples: we use them to show that the saturation number doesn't have to have nicely behaved asymptotics for forbidden families of hypergraphs. Chapter 2 ends with

an explicit construction of automata with large rendezvous time, while in Chapter 3 we construct semi-perfect 1-factorizations to show that they can exist. The constructions in Chapter 4 are neither examples nor counterexamples, but rather tools used to bound the probability of small connected components of the geometric random graph.

In each of these cases, it is not sufficient to merely come up with the constructions, though this can itself be difficult. It takes further work to prove that the constructions have the desired properties or to bound the probability of the construction occurring, as relevant.

Chapter 1

Hypergraph Saturation Irregularities

1.1 Introduction

For a fixed graph F and an integer n, Turán's extremal number $\mathrm{ex}(F,n)$ is the maximum number of edges in a graph on n vertices that does not contain F as a subgraph.

$$\mathrm{ex}(F,n) = \max\{e(G) : |G| = n \text{ and } G \text{ is } F\text{-free}\}.$$

It is not hard to see that the Turán density $\lim_{n\to\infty} \mathrm{ex}(F,n)/\binom{n}{2}$ must exist. The density $\mathrm{ex}(F,n)/\binom{n}{2}$ must be decreasing with n: removing a vertex of minimum degree from an F-free graph on n vertices gives a denser F-free graph on $n-1$ vertices. As the density is bounded between 0 and 1 the limit must therefore exist.

The Turán graph $T_{n,r}$ is a complete r-partite graph on n vertices with the r parts as equal in size as possible. This is clearly K_{r+1}-free, and Turán [30] proved that it has the maximal number of edges.

Theorem 1.1 (Turán)**.** $\mathrm{ex}(K_r,n) = e(T_{n,r-1}) \leq \left(1 - \frac{1}{r-1}\right)\frac{n^2}{2}$ *for all* r,n.

Turán's Theorem gives the Turán density for complete graphs. The Erdős-Stone Theorem [11] generalises Turán's Theorem to give the value of the Turán density for all graphs.

Theorem 1.2 (Erdős-Stone)**.** *For any graph F, we have* $\mathrm{ex}(F,n) = \left(1 - \frac{1}{\chi(F)-1} + o(1)\right)\binom{n}{2}$ *where $\chi(F)$ is the chromatic number of F.*

Turán's Theorem and the Erdős-Stone Theorem are foundation stones of extremal graph theory, and Turán's number $\mathrm{ex}(F,n)$ has been very well studied.

Suppose that instead that we want to ask about the minimum number of edges rather than the maximum. As it stands this is a trivial question, since if F has at least one edge an n-vertex graph with no edges is F-free and so the answer is always zero. Note however that if a graph is F-free and has the maximum number of edges possible then adding any edge to the graph must create a copy of F. This inspires the following definition.

For a fixed graph F, we say that a graph G is *F-saturated* if G does not contain F as a subgraph but adding any edge to G would create a copy of F. We can write down an equivalent definition of the extremal number in terms of saturation.

$$\mathrm{ex}(F,n) = \max\{e(G) : |G| = n \text{ and } G \text{ is } F\text{-saturated}\}.$$

Replacing maximum with minimum gives the definition of the *saturation number* $\mathrm{sat}(F,n)$:

$$\mathrm{sat}(F,n) = \min\{e(G) : |G| = n \text{ and } G \text{ is } F\text{-saturated}\}.$$

As an example, consider the case when F is a triangle. The triangle-free graph with the maximum number of edges is a complete bipartite graph with parts of size $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$, giving that $\mathrm{ex}(F,n) = \left\lfloor \frac{n^2}{4} \right\rfloor$.

On the other hand, the triangle-free graph on n vertices with the minimum number of edges is a star. Clearly the star is triangle-saturated as it does not contain a triangle and

adding any edge would create one. It also has the minimum number of edges: the star on n vertices has $n-1$ edges and any graph on n vertices with fewer edges is disconnected. Adding an edge from one connected component to another could not possibly create a triangle and so any disconnected graph is not triangle-saturated. This tells us that $\mathrm{sat}(F,n) = n-1$.

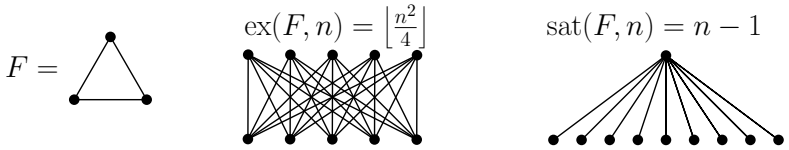


Figure 1.1: An example of the extremal number and the saturation number for a triangle.

The saturation number behaves quite differently to the extremal number. For example, the simple argument that the Turán density exists does not translate to the saturation number. Removing a vertex from an F-saturated graph does not necessarily give an F-saturated graph (in contrast with F-free) and so we cannot conclude that the density is decreasing.

We know from the Erdős-Stone Theorem that when F is not bipartite the extremal number is order n^2. We saw that for the triangle the saturation number was, in contrast, linear in n. Kászonyi and Tuza [17] proved the following:

Theorem 1.3 (Kászonyi, Tuza)**.** *For every graph F, we have* $\mathrm{sat}(F,n) = O(n)$.

In contrast to the extremal number and the Erdős-Stone Theorem we do not know whether the saturation number grows like cn where c is some constant depending on F. Tuza [31] conjectured that it does.

Conjecture 1.1 (Tuza)**.** *For every 2-graph F the limit* $\lim_{n\to\infty} \frac{\mathrm{sat}(F,n)}{n}$ *exists.*

Tuza's conjecture is probably the biggest open question concerning saturation. Bollobás [4] proved that $\mathrm{sat}(K_k,n) = \binom{n}{2} - \binom{n-k+2}{2}$, so the conjecture holds for complete graphs. For more information and other results relating to the saturation number and Tuza's conjecture, see surveys from Pikhurko [25] and from Faudree, Faudree and Schmitt

[12].

We can generalise the notion of saturation to families of graphs. For a family $\mathcal{F}$ of graphs (called a forbidden family), a graph G is called *$\mathcal{F}$-saturated* if it does not contain any graph in $\mathcal{F}$ as a subgraph, but adding any edge creates a copy of some graph $F \in \mathcal{F}$ as a subgraph of G. We define the saturation number in the same way as before:

$$\mathrm{sat}(\mathcal{F}, n) = \min\{e(G) : |G| = n \text{ and } G \text{ is } \mathcal{F}\text{-saturated}\}.$$

Note that we can also generalise the extremal number in this way, and the same argument showing the Turán density exists still works.

For a family $\mathcal{F}$ of graphs we have $\mathrm{sat}(\mathcal{F}, n) = O(n)$ [17], just as we did for single graphs. However, the generalisation of Tuza's conjecture to finite families of graphs is not true: an example of a finite family $\mathcal{F}$ where $\mathrm{sat}(\mathcal{F}, n)/n$ does not tend to a limit was given by Pikhurko [25].

Theorem 1.4 (Pikhurko)**.** *There exists a finite family $\mathcal{F}$ of graphs such that the limit $\lim_{n\to\infty} \frac{\mathrm{sat}(F,n)}{n}$ does not exist.*

Sketch of Proof. Fix some constant $k \geq 4$. The idea is to construct a family such that $\mathrm{sat}(\mathcal{F}, n)$ has different behaviour depending on whether n is divisible by k.

The family $\mathcal{F}$ contains the 'dumb-bell' graph that is two copies of K_k and a single edge between them. The other graphs in $\mathcal{F}$ are, for each $1 \leq i \leq k-1$, the union of two copies of K_k intersecting in i common vertices. See Figure 1.2 for an example where $k = 4$.

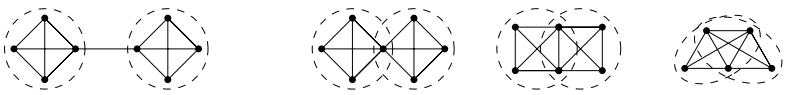


Figure 1.2: An example of Pikhurko's family for $k = 4$.

It is easy to check that when k divides n the graph consisting of $\frac{n}{k}$ disjoint copies of

K_k is $\mathcal{F}$-saturated.

When k does not divide n, Pikhurko shows that here is no such 'nice' construction and any $\mathcal{F}$-saturated graph G on n vertices must contain many extra edges. To get a handle on the structure of G notice that any copies of K_k within G must be disjoint. This further implies that not all vertices of G can be contained within a K_k, since k does not divide n. Combining this observation with a counting argument completes the proof. □

Pikhurko asked whether a similar result holds for r-uniform hypergraphs.

An r-uniform hypergraph, or *r-graph*, is a pair, $(V(H), E(H))$, of *vertices* and *edges* where the edge set $E(H)$ is a collection of r-element subsets of the vertex set $V(H)$. We have $|H| = |V(H)|$ and $e(H) = |E(H)|$. When the context is clear we will refer to r-graphs simply as graphs. Note that a 2-graph is just a graph in the usual sense.

For a family $\mathcal{F}$ of r-graphs, it was shown by Pikhurko [24] that $\mathrm{sat}(\mathcal{F}, n) = O\left(n^{r-1}\right)$ when the family contains only a finite number of graphs. When $r > 2$ this is still open for infinite families of r-graphs, although when $r = 2$ Tuza's result applies to both finite and infinite families.

Pikhurko's result leads to the following generalisation of Tuza's conjecture to r-graphs, first posed by Pikhurko [24].

Conjecture 1.2. *For every r-graph F the limit $\lim_{n\to\infty} \frac{\mathrm{Sat}(F,n)}{n^{r-1}}$ exists.*

Let $K_k^{(r)}$ be the complete r-graph on k vertices (that is, the edge set of $K_k^{(r)}$ is all sets of vertices of size r). Bollobás [4] proved that $\mathrm{sat}(K_k^{(r)}, n) = \binom{n}{r} - \binom{n-k+r}{r}$ for all $k \geq r$, so the conjecture holds for the complete r-graphs.

As in the 2-graph case we can further generalise this conjecture by replacing the single r-graph F with a finite family of r-graphs $\mathcal{F}$. Our main aim in this paper is to prove that this generalised conjecture is not true — that is, for all r there exists a finite

family of r-graphs $\mathcal{F}$ such that $\mathrm{sat}(\mathcal{F}, n)/n^{r-1}$ does not tend to a limit. This resolves a question of Pikhurko (problem 7 in [25]).

Theorem 1.5. For all $r \geq 2$ there exists a family $\mathcal{F}$ of r-graphs and a constant $k \in \mathbb{N}$ such that

$$\mathrm{sat}(\mathcal{F}, n) = \begin{cases} O(n) & \text{if } k \mid n \\ \Omega\left(n^{r-1}\right) & \text{if } k \nmid n \end{cases}$$

In particular, for any $l \in \{1, 2, \ldots, r-1\}$, we have that $\frac{\mathrm{sat}(\mathcal{F},n)}{n^l}$ does not converge.

We prove Theorem 1.5 in Section 1.2. As in Pikhurko's proof for the 2-graph case, the idea of the proof will be to choose a constant k and to define a forbidden family $\mathcal{F}$ such that when k divides n there is a 'nice' construction of an $\mathcal{F}$-saturated graph with few edges; and when k does not divide n, an $\mathcal{F}$-saturated graph requires comparatively many edges.

Our proof of Theorem 1.5 uses a family $\mathcal{F}$ which grows in size with r. In a variation of the theorem, proved in Section 1.3, we show that we can reduce the size of the forbidden family to be independent of r.

Theorem 1.6. For all $r \geq 3$ there exists a family $\mathcal{F}$ of four r-graphs such that $\frac{\mathrm{sat}(\mathcal{F},n)}{n^{r-1}}$ does not converge.

In reducing the family to a constant size we lose the large gap between the asymptotics that we had in Theorem 1.5. In particular, for a choice of constant k, we still have that if $k \nmid n$ then $\mathrm{sat}(\mathcal{F}, n) = \Omega\left(n^{r-1}\right)$, but if $k|n$ we only have $\mathrm{sat}(\mathcal{F}, n) = O\left(n^{r-2}\right)$ (as opposed to the $O(n)$ we had before). It would be interesting to know whether this extreme oscillation between $O(n)$ and $\Omega\left(n^{r-1}\right)$ genuinely does require an unbounded family or whether this is just an artefact of the construction.

Consider, with respect to a family satisfying the requirements of Theorem 1.5 or Theorem 1.6, the set of integers n where $\mathrm{sat}(\mathcal{F}, n)$ is $O(n)$. This set has low density: specifically, density $1/k$ where k grows with r. A second variation of the theorem gives a

forbidden family such that the the set of integers n where $\text{sat}(\mathcal{F}, n)$ is $O(n)$ has density $1/2$. This is proved in Section 1.4.

Theorem 1.7. For all $r \geq 2$ there exists a family $\mathcal{F}$ of r-graphs such that

$$\text{sat}(\mathcal{F}, n) = \begin{cases} O(n) & \text{if } n \text{ is even} \\ \Omega\left(n^{r-1}\right) & \text{if } n \text{ is odd.} \end{cases}$$

We end with some open problems in Section 1.5.

1.2 A Proof of the Main Theorem

In this section we prove Theorem 1.5 by giving an explicit construction of such a family and showing that it has the required properties.

Theorem 1.5. *For all $r \geq 2$ there exists a family $\mathcal{F}$ of r-graphs and a constant $k \in \mathbb{N}$ such that*

$$\text{sat}(\mathcal{F}, n) = \begin{cases} O(n) & \textit{if } k \mid n \\ \Omega\left(n^{r-1}\right) & \textit{if } k \nmid n \end{cases}$$

In particular, for any $l \in \{1, 2, \ldots, r-1\}$, we have that $\frac{\text{sat}(\mathcal{F},n)}{n^l}$ does not converge.

Proof. Fix any integer $k > r$ and take $\mathcal{F}$ to be the family of all of the following r-graphs:

a) For each $1 \leq i \leq k-1$, the graph F_i consisting of two copies of $K_k^{(r)}$ intersecting in exactly i vertices.

b) For each $(x_1, x_2, \ldots, x_t)$ with $\Sigma x_i = r$ and $1 \leq x_1 \leq x_2 \leq \ldots \leq x_t \leq (r-1)$, the graph $H_{(x_1,x_2,\ldots,x_t)}$ consisting of t disjoint copies of $K_k^{(r)}$ and an edge E meeting the i^{th} copy of $K_k^{(r)}$ in x_i vertices. We refer to E as the bridge edge.

An example of the family $\mathcal{F}$ for $r = 5$ and $k = 7$ can be seen in figure 1.3, where the vertices surrounded by a dashed line represent a copy of $K_k^{(r)}$, and vertices grouped by

a solid line represent a bridge edge.

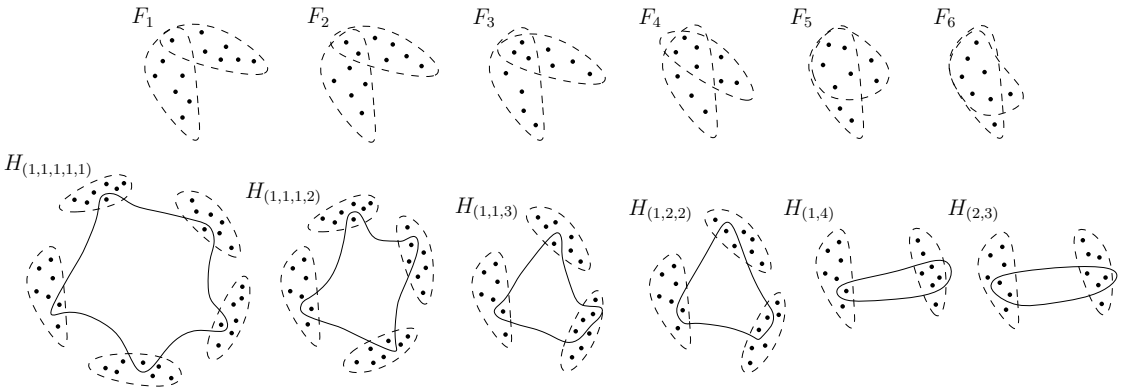


Figure 1.3: The family $\mathcal{F}$ of r-graphs for $r = 5$ and $k = 7$

First, let us deal with the case where k divides n. Take the graph consisting of $\frac{n}{k}$ disjoint copies of $K_k^{(r)}$. It is easy to see that this is $\mathcal{F}$-saturated and thus

$$\mathrm{sat}(\mathcal{F}, n) \leq \frac{n}{k}\binom{k}{r} = O(n).$$

Note that in fact $\mathrm{sat}(\mathcal{F}, n)$ is equal to $\frac{n}{k}\binom{k}{r}$ (although we do not require this), as can be shown by an argument similar to the one that follows.

Now suppose that $k \nmid n$ and let $G = (V, E)$ be a graph on n vertices that is $\mathcal{F}$-saturated. We will show that $e(G) = \Omega\left(n^{r-1}\right)$. Let A be the set of all vertices of G that are contained in a $K_k^{(r)}$, and $B = V \setminus A$ be all vertices not contained in any $K_k^{(r)}$.

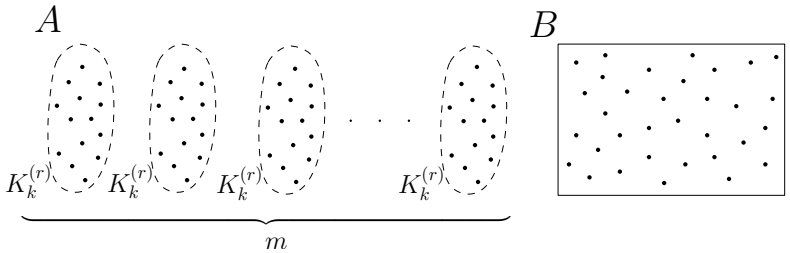


Figure 1.4: The structure of the graph G

Note that the subgraph of G induced by A must consist of m disjoint copies of $K_k^{(r)}$ for some $m \geq 0$, by the choice of family $\mathcal{F}$. This implies that B is not empty, since k

does not divide n. Note also that if an r-set intersecting B is not in $E(G)$, then adding this edge to G must create a copy of $K_k^{(r)}$ — it must create some graph in $\mathcal{F}$ and it cannot form a bridge between two or more $K_k^{(r)}$s by definition of B.

We make the following two claims about the number of edges in G:

Claim 1.5.1. G contains at least $\binom{mk}{r-1} - m\binom{k}{r-1}$ edges consisting of $r-1$ vertices in A and one vertex in B.

Claim 1.5.2. Suppose $|B| \geq k-1$. Then G contains at least $\binom{|B|}{r-1}\frac{k-r}{r}$ edges.

We can use these two claims to deduce the result. One of A and B contains at least half the vertices in G. If $|A| = mk \geq \frac{n}{2}$, then using claim 1.5.1 the number of edges in G is $\Omega\left(n^{r-1}\right)$. If $|B| \geq \frac{n}{2}$, then using claim 1.5.2 the number of edges in G is $\Omega\left(n^{r-1}\right)$.

All that is left is to prove the two claims.

Proof of Claim. [Proof of Claim 1.5.1.] This holds trivially if m is 0 or 1, so we may assume $m \geq 2$. Fix a set X of $r-1$ vertices in A, not all in the same copy of $K_k^{(r)}$. We will show that G contains at least one edge containing all vertices of X together with a vertex in B. This proves the claim since there are $\binom{mk}{r-1} - m\binom{k}{r-1}$ such sets X.

Fix some x in B (note that B is non-empty) and suppose that the r-set $X \cup \{x\}$ is not in $E(G)$. Since G is $\mathcal{F}$-saturated, adding $X \cup \{x\}$ as an edge must create a copy of $K_k^{(r)}$ on some vertex set K.

Note that for y in $A \setminus X$, the r-set $X \cup \{y\}$ is not in $E(G)$, as otherwise it would form a copy of some $H_{(x_1,\ldots,x_t)}$ in $\mathcal{F}$. Thus the vertices in $K \setminus (X \cup \{x\})$ cannot be in A.

Hence $K \setminus (X \cup \{x\})$ is contained entirely in B, and so G contains $k - r > 1$ edges that consist of all vertices of X together with a vertex in B.

Proof of Claim. [Proof of Claim 1.5.2.] Fix a set X of $r-1$ vertices in B. We will show

that X is contained in at least $k-r$ edges of G.

Suppose first that we have that $X \cup \{y\}$ is an edge for all y in $B \setminus X$. Then X is in $|B| - |X| \geq k - r$ edges as required.

Otherwise, there exists some y in $B \setminus X$ such that $X \cup \{y\}$ is not in $E(G)$. Then adding the edge $X \cup \{y\}$ must create a copy of $K_k^{(r)}$ in G, since B contains no $K_k^{(r)}$s and so this edge cannot be a bridge edge. Then we have that X is contained in $k-r$ edges in that $K_k^{(r)}$.

Thus every $(r-1)$-set in B is contained in at least $k-r$ edges. Each edge in G contains at most r different $(r-1)$-sets in B, and so the total number of edges in G is at least

$$\binom{|B|}{r-1}\frac{k-r}{r}.$$

□

Note that the size of the family $\mathcal{F}$ used in this proof is $(r-1)+(p(r)-1)$, where p is the partition function. We have $\log(p(r)) = \Theta(\sqrt{r})$, and so the size of the family grows exponentially with $\sqrt{r}$.

1.3 A Forbidden Family of Constant Size

Recall that Tuza's conjecture (and its generalisation to r-graphs) concerns a forbidden family of just one r-graph. The size of the forbidden family in Theorem 1.5 grows with r, the size of each edge. It is natural then to ask whether there exists a family of constant size, independent of r, which has this same non-convergence property.

We will prove that there is such a family, using the family $\mathcal{F}$ consisting of the following four r-graphs:

F: Two $K_k^{(r)}$s intersecting in one vertex,

H: r disjoint copies of $K_k^{(r)}$ and an edge intersecting each $K_k^{(r)}$,

I_2: One $K_k^{(r)}$ and an edge intersecting it in exactly two vertices, and

I_{r-1}: One $K_k^{(r)}$ and an edge intersecting it in exactly $r-1$ vertices.

An example of the family $\mathcal{F}$ for $r = 5$ and $k = 15$ can be seen in figure 1.5, where the vertices surrounded by a dashed line represent a copy of $K_k^{(r)}$, and vertices grouped by a solid line represent an extra edge.

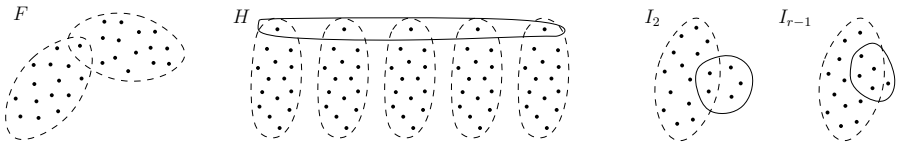


Figure 1.5: The family $\mathcal{F}$ of r-graphs for $r = 5$ and $k = 15$

This family was obtained by considering the two types of graphs we had in our previous family, and finding a smaller set of graphs that fulfil the same role for each.

The previous family contained all of the graphs $H_{(x_1,\dots,x_t)}$ to ensure that the graph consisting of disjoint copies of $K_k^{(r)}$ was $\mathcal{F}$-saturated. In particular, this meant that when $k \mid n$ there was an $\mathcal{F}$-saturated graph of size $O(n)$. We will keep $H = H_{(1,1,\dots,1)}$ to ensure there are no edges intersecting r different copies of $K_k^{(r)}$, and replace the other $p(r) - 2$ graphs by I_2 and I_{r-1}. With this smaller family we can no longer find a graph of size $O(n)$ that is $\mathcal{F}$-saturated, so we lose the large gap in asymptotics that we had in Theorem 1.5. However, we can construct (for n divisible by k) an $\mathcal{F}$-saturated graph that has size $O\left(n^{r-2}\right)$.

The previous family contained all of the graphs F_i to ensure that all of the copies of $K_k^{(r)}$ in an $\mathcal{F}$-saturated graph must be disjoint. For k sufficiently large, these $r-1$ different graphs can be replaced by the three graphs I_2, I_{r-1} and $F = F_1$, which achieve the same goal.

Theorem 1.6. *For all $r \geq 3$ there exists a family $\mathcal{F}$ of four r-graphs such that $\frac{\mathrm{sat}(\mathcal{F},n)}{n^{r-1}}$ does not converge.*

Proof of Theorem 1.6. Fix $r \geq 3$ and $k \geq \max\{r+1, 2r-4\}$.

Let $\mathcal{F}$ be the set containing the four r-graphs F, H, I_2 and I_{r-1}, as defined earlier.

First, we will construct an example of an $\mathcal{F}$-saturated graph with $O\left(n^{r-2}\right)$ edges when k divides n.

Claim 1.6.1. $\mathrm{sat}(\mathcal{F}, n) = O\left(n^{r-2}\right)$ when $k \mid n$.

Proof of Claim. Let G be a graph consisting of $\frac{n}{k} = m$ disjoint copies of $K_k^{(r)}$, together with all other edges *except* those which:

i) intersect r of the $K_k^{(r)}$s, each in one vertex, or

ii) intersect one of the $K_k^{(r)}$s in exactly two vertices, or

iii) intersect one of the $K_k^{(r)}$s in exactly $r-1$ vertices.

Clearly, adding any r-set not in $E(G)$ to the graph G creates one of the graphs in the family $\mathcal{F}$ — graphs I_2, I_{r-1} and H respectively.

We will show that G contains no other $K_k^{(r)}$s except for the original ones, thus proving that G does not contain any graph in $\mathcal{F}$. Suppose G did contain another $K_k^{(r)}$. Using that $k \geq 2r-2$, we have that one of the following three cases holds:

- the new $K_k^{(r)}$ intersects all of the original $K_k^{(r)}$s in at most one vertex, and thus it contains an edge of type (i);
- the new $K_k^{(r)}$ and one of the original $K_k^{(r)}$s intersect in between two and $k-(r-2)$ vertices, and thus it contains an edge of type (ii); or
- the new $K_k^{(r)}$ and one of the original $K_k^{(r)}$s intersect in $r-1$ or more vertices, and thus it contains an edge of type (iii).

Whichever case we are in, we have a contradiction. Thus G is $\mathcal{F}$-saturated.

We now need to calculate the size of G. The number of r-sets meeting exactly t of the $K_k^{(r)}$s is $O\left(n^t\right)$. Note that G does not contain any edges intersecting more than $r-2$ of the $K_k^{(r)}$s and thus we have $e(G) = O\left(n^{r-2}\right)$ and Claim 1.6.1 follows.

Now we will consider the case where k does not divide n. We want to show that $\text{sat}(\mathcal{F}, n) = \Omega\left(n^{r-1}\right)$ when $k \nmid n$.

Let $G = (V, E)$ be a graph on n vertices that is $\mathcal{F}$-saturated. Let A be the set of all vertices of G that are contained in a $K_k^{(r)}$, and $B = V \setminus A$ be all vertices not contained in any $K_k^{(r)}$.

The choice of family $\mathcal{F}$ implies that all of the copies of $K_k^{(r)}$ contained in A must be disjoint:

- F forbids two $K_k^{(r)}$s intersecting in exactly one vertex.
- I_2 forbids two $K_k^{(r)}$s intersecting in at least two vertices and each containing at least $r-2$ vertices not in the intersection.
- I_{r-1} forbids two $K_k^{(r)}$s intersecting in at least $r-1$ vertices and each containing at least one vertex not in the intersection.

Since k was chosen with $k \geq 2r-4$, any two intersecting copies of $K_k^{(r)}$s fall in one of these three categories.

Thus A consists of m disjoint copies of $K_k^{(r)}$ for some m, together with some extra edges that go between them. Since k does not divide n, we can conclude that A is not all of $V(G)$, or equivalently that B is non-empty.

We make the following two claims about the number of edges in G:

Claim 1.6.2. If $m \geq r-1$ then G contains at least $\binom{m}{r-1}k^{r-1}$ edges consisting of $r-1$ vertices in A and one vertex in B.

Claim 1.6.3. If $|B| \geq k-1$, then G contains $\binom{|B|}{r-1}\frac{k-r}{r}$ edges.

We can use these two claims to deduce the result. One of A and B contains at least half the vertices in G. If $|A| = mk \geq \frac{n}{2}$, then using claim 1.6.2 the number of edges in G is at least $\Omega\left(n^{r-1}\right)$. If $|B| \geq \frac{n}{2}$, then using claim 1.6.3 the number of edges in G is at least $\Omega\left(n^{r-1}\right)$.

All that is left is to prove the two claims.

Proof of Claim. [Proof of Claim 1.6.2.] Fix $r-1$ vertices $v_1, \ldots, v_{r-1}$, each in a different copy of $K_k^{(r)}$ in A. We will show that G contains at least one edge containing all of $v_1, \ldots v_{r-1}$ together with a vertex in B.

If all possible such edges exist, then we are done (recall that B is non-empty).

Otherwise, there is some x in B such that the r-set $\{x, v_1, \ldots, v_{r-1}\}$ is not in $E(G)$. Adding this edge must create a graph in $\mathcal{F}$, and so it must create a new copy of $K_k^{(r)}$ (it cannot be any of the 'extra' edges in I_2, I_{r-1} or H).

Consider this new $K_k^{(r)}$. Suppose for a contradiction it contains some vertex y in $A \setminus \{v_1 \ldots v_{r-1}\}$. If y is in the same original $K_k^{(r)}$ as one of the v_is then the edge $\{y, v_1, \ldots v_{r-1}\}$ creates a copy of I_2. If y is not in the same original $K_k^{(r)}$ as any of the v_is then the edge $\{y, v_1, \ldots v_{r-1}\}$ creates a copy of H. Thus $\{y, v_1, \ldots v_{r-1}\}$ is not in $E(G)$, and we have a contradiction.

Thus the other vertices of this new $K_k^{(r)}$ are in B, and G contains $k-r > 1$ edges containing all of $v_1, \ldots v_{r-1}$ together with a vertex in B.

Proof of Claim. [Proof of Claim 1.6.3.] Fix a set X of $r-1$ vertices in B. We will show that X is contained in at least $k-r$ edges in G.

Suppose first that $X \cup \{y\}$ is an edge for all y in $B \setminus X$. Then X is in $|B| - |X| \geq k-r$ edges as required.

Otherwise, there exists some y in $B \setminus X$ such that $X \cup \{y\}$ is not in $E(G)$. Note that adding the edge $X \cup \{y\}$ must create some graph in $\mathcal{F}$. It must thus create a $K_k^{(r)}$, since B contains no $K_k^{(r)}$s and so it cannot be the 'extra' edge in I_2, I_{r-1} or H. Then we have that X is contained in $k-r$ other edges in that $K_k^{(r)}$.

Thus every $r-1$ set in B is contained in at least $k-r$ edges. Each edge in G contains at most r different $(r-1)$-sets in B, and so the total number of edges in G is at least

$$\binom{|B|}{r-1}\frac{k-r}{r}.$$

□

1.4 Obtaining an Small Saturation Number on a Denser Set

In both Theorem 1.5 and Theorem 1.6, the saturation number is asymptotically small ($O\left(n^{r-2}\right)$) when n is divisible by k and asymptotically large ($\Theta\left(n^{r-1}\right)$) for all other values of n. Since k is at least as big as r, this means that the set of values where the saturation number is asymptotically small has low density — less than $1/r$. It is natural to ask whether it is possible to have a forbidden family where the saturation number has different asymptotics for complementary subsets of the naturals of equal density. For example, could we ensure $\mathrm{sat}(\mathcal{F}, n)$ is asymptotically small on even numbers and asymptotically large on odd numbers?

It turns out that it is possible.

Theorem 1.7. *For all $r \geq 2$ there exists a family $\mathcal{F}$ of r-graphs such that*

$$\mathrm{sat}(\mathcal{F}, n) = \begin{cases} O(n) & \text{if } n \text{ is even} \\ \Omega\left(n^{r-1}\right) & \text{if } n \text{ is odd.} \end{cases}$$

The proof will use a family similar to the one in Theorem 1.5. However, rather than just using $K_k^{(r)}$ as a base graph, we will take two base graphs of different even orders. The family will contain all possible intersections of the two base graphs, and all graphs consisting of disjoint unions of copies of the base graphs together with a bridge edge.

For large even n there is an $\mathcal{F}$-saturated graph on n vertices that uses few edges; namely taking disjoint copies of the base graphs. However, for odd n we will need to use many more edges.

A first attempt at choosing the two base graphs might be to take $K_k^{(r)}$ and $K_{k+2}^{(r)}$ for some even k. However, $K_{k+2}^{(r)}$ contains two $K_k^{(r)}$s intersecting in $k-2$ vertices, which is a graph we would want to include in our forbidden family. This is a problem, as if that graph was forbidden, all copies of $K_{k+2}^{(r)}$ would be forbidden too.

Instead, we will take $K_k^{(r)}$ and any graph on $k+2$ vertices which has certain helpful properties: one of which is that it contains only one copy of $K_k^{(r)}$.

Proof. Fix any even integer $k > r+1$. Let $K = K_k^{(r)}$ and let L be an r-graph which satisfies the following properties:

- L has $k+2$ vertices;
- Every vertex of L is contained in at least one edge;
- L contains exactly one copy of $K_k^{(r)}$; and
- For any edge e of L and any $(r-1)$-sized subset s of the edge e, there is another edge e' in L such that $e \cap e' = s$.

One such L consists of a $K_k^{(r)}$ and two $K_{k-1}^{(r)}$s with a common intersection of $k-2$ vertices. It is easy to see that this has the required properties.

We call K and L the *base graphs*. An example of the these two graphs for $k = 18$ can be seen in figure 1.6, where if a set of vertices is surrounded by a dashed line then all edges contained in that set exist.

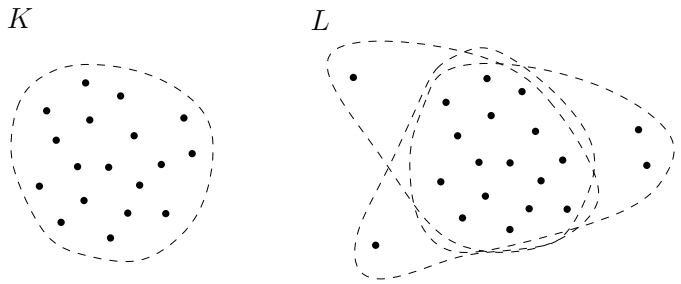


Figure 1.6: The two graphs under consideration for $k = 14$.

Take $\mathcal{F}$ to be the family containing all of the following r-graphs:

a) Every graph comprising t disjoint graphs $H_1, H_2, \dots H_t$ (for $2 \leq t \leq r$) where each H_i is a base graph, together with an edge E meeting each H_i in $x_i \geq 1$ vertices such that $\sum_{i=1}^{t} x_i = r$. We call E a bridge edge.

b) Every graph comprising two base graphs on vertex sets V_1 and V_2 with non-empty intersection and neither contained in the other — that is, $V_1 \cap V_2, V_1 \setminus V_2$ and $V_2 \setminus V_1$ all non-empty.

c) Every graph comprising L plus a single extra edge on the same vertex set.

First, let us deal with the case where n is even. For all n sufficiently large, (in particular, at least $\frac{k(k-2)}{2}$), we can write n as a sum $ak + b(k+2)$ for some $a, b \in \mathbb{N}$.

Take G to be a graph on n vertices consisting of a disjoint copies of K and b disjoint copies of L. It is clear that adding any edge will create a graph in the family $\mathcal{F}$: adding a missing edge between base graphs creates a graph of type (a), and adding a missing

edge within a copy of L creates a graph of type (c). Thus

$$\begin{aligned}\mathrm{sat}(\mathcal{F}, n) \leq e(G) &\leq a\binom{k}{r} + b\binom{k+2}{r} \\ &\leq (ak + b(k+2))\frac{1}{r}\binom{k+1}{r-1} \\ &= O(n)\end{aligned}$$

Now suppose that n is odd and let $G = (V, E)$ be a graph on n vertices that is $\mathcal{F}$-saturated.

Let A be the set of all vertices of G that are contained in a copy of one of the base graphs, and $B = V \setminus A$ be all vertices not contained in any copy of a base graph.

Note that the subgraph induced on A must consist of disjoint copies of the two base graphs, by the choice of family $\mathcal{F}$. This implies that B is not empty, since n is odd and both base graphs have an even number of vertices.

Let X be an r-set meeting B that is not in $E(G)$. Adding the edge X to G must create some graph in $\mathcal{F}$. The edge X cannot form a bridge between two $K_k^{(r)}$s by definition of B, and it also cannot add an extra edge to an existing copy of L for the same reason. Thus adding such an edge must create a copy of one of the base graphs, K or L.

We make the following two claims about the number of edges in G:

Claim 1.7.1. G contains at least

$$\binom{\frac{|A|}{k+2}}{r-1}k^{r-1}$$

edges consisting of $r-1$ vertices in A and one vertex in B.

Claim 1.7.2. If $|B| \geq k-1$ then G contains at least $\binom{|B|}{r-1}\frac{k-r}{r}$ edges.

We can use these two claims to deduce the result. One of A and B contains at least half the vertices in G. If $|A| \geq \frac{n}{2}$, then using claim 1.7.1 the number of edges in G is at

least $\Omega\left(n^{r-1}\right)$. If $|B| \geq \frac{n}{2}$, then using claim 1.7.2 the number of edges in G is $\Omega\left(n^{r-1}\right)$.

All that is left is to prove the two claims.

Proof of Claim. [Proof of Claim 1.7.1.] Fix $r-1$ vertices $v_1, \ldots, v_{r-1}$, each in a different base graph in A (of which there are at least $\frac{|A|}{k+2}$). We will show that G contains at least one edge containing all of $v_1, \ldots v_{r-1}$ together with a vertex in B. Then the number of edges between A and B is at least the desired amount.

If all possible such edges exist, then we are done (recall that B is non-empty).

Otherwise, there is some x in B such that the r-set $\{x, v_1, \ldots, v_{r-1}\}$ is a missing edge. Adding this edge must create one of the graphs in $\mathcal{F}$. It cannot be a bridge edge as B contains no copies of L. It also cannot be the extra edge in a copy of 'L plus an edge', as no vertex in B is contained in a copy of L. Thus adding the edge must create a new copy of one of the base graphs, $K_k^{(r)}$ or L.

The other vertices of this new base graph cannot be in A: if y is in $A \setminus \{v_1 \ldots v_{r-1}\}$, then $\{y, v_1, \ldots v_{r-1}\}$ is a non-edge, otherwise it serves as a bridge edge and G contains a graph of type (a) in the family $\mathcal{F}$.

So the other vertices of this new base graph are all in B. We then have that there is at least one edge containing all of $v_1, \ldots v_{r-1}$ together with a vertex in B: this is obviously true if the base graph was $K_k^{(r)}$, and true by the properties insisted upon earlier if the base graph was L.

Proof of Claim. [Proof of Claim 1.7.2.] To apply a similar proof to before, we first want to show that if Y is an r-set in B that is not in $E(G)$, then adding Y to G creates a new copy of $K_k^{(r)}$. Suppose for contradiction this is not the case.

Adding Y must create a graph in the family $\mathcal{F}$, so Y must create one of:

- a graph of type (a) in $\mathcal{F}$;

- a copy of the graph L; or
- a copy of the graph L plus an edge.

However, no vertex in B is in a copy of one of the base graphs. This implies both that Y cannot be a bridge edge between copies of the base graph and also that Y cannot be an extra edge added to a copy of L. Thus we must have that Y creates a copy of L.

Note that L contains a copy of $K_k^{(r)}$. Since Y does not create a $K_k^{(r)}$, this $K_k^{(r)}$ must already exist. However, then Y must intersect this $K_k^{(r)}$ in $r-2$ vertices, contradicting that no vertex in B is contained within a copy of $K_k^{(r)}$.

Now, fix a set X of $r-1$ vertices in B. We will show that X is contained in at least $k-r$ edges in G.

Suppose first that $X \cup \{y\}$ is an edge for all y in $B \setminus X$. Then X is in $|B|-|X| \geq k-r$ edges as required.

Otherwise, there exists some y in $B \setminus X$ such that $X \cup \{y\}$ is not in $E(G)$. Note that by the above argument, adding the edge $X \cup \{y\}$ must create a copy of $K_k^{(r)}$ and so we have that X is contained in $k-r$ other edges in that $K_k^{(r)}$.

Thus every $r-1$ set in B is contained in at least $k-r$ edges. Each edge in G contains at most r different $(r-1)$-sets in B, and so the total number of edges in G is at least

$$\binom{|B|}{r-1}\frac{k-r}{r}.$$

□

1.5 Open Questions

The main questions still left unanswered are the two conjectures in Section 1; that is, Tuza's conjecture and its generalisation to r-graphs. The questions that follow are all variations on and generalizations of these conjectures.

In Theorem 1.6, we defined H to be the graph consisting of r disjoint copies of $K_k^{(r)}$ and an edge intersecting each $K_k^{(r)}$. This seems to somehow be the key graph in ensuring a nice construction when k divides n and so we might guess that H is a counterexample to Tuza's conjecture. Unfortunately, $\frac{\mathrm{Sat}(H,n)}{n^{r-1}}$ does tend to a limit.

In the case $r = 2$, we use an argument similar to one by Pikhurko (example 4, [25]).

Proposition 1.8. *Let H be the dumb-bell graph consisting of two copies of K_k joined by a bridge edge. For $k \geq 3$, we have that $\frac{\mathrm{Sat}(H,n)}{n} = \left(\frac{k-1}{2}\right) n + O(1)$.*

Proof. First, we show that $\frac{\mathrm{Sat}(H,n)}{n} \leq \left(\frac{k-1}{2}\right) n + O(1)$. Write $n = mk + c$ where $0 \leq c < k$ and let G be the graph on n vertices consisting of $m - 1$ disjoint K_ks and one K_{k+c}. This graph G is certainly H-saturated and has $\frac{k-1}{2}n + O(1)$ edges.

Next, we show that $\frac{\mathrm{Sat}(H,n)}{n} \geq \left(\frac{k-1}{2}\right) n + O(1)$.

Let G be an H-saturated graph on n vertices. First, note that if G has minimum degree $\geq k - 1$ then $e(G) \geq \left(\frac{k-1}{2}\right) n$ and we are done immediately.

Thus suppose that there exists a vertex $v \in G$ of degree $\leq k - 2$. For any x that is not a neighbour of v, adding the edge vx must create a copy of H. Since v has $\leq k - 2$ other neighbours the edge vx cannot be a bridge edge and so it must complete a copy of K_k. In particular, v must have exactly $k - 2$ other neighbours and x must be connected to all of them. This holds for all x that are not in the neighbourhood of v, and so G has at least $(k - 2)(n - (k - 2)) > \frac{k-1}{2}n - (k - 2)^2$ edges. □

For $r > 2$, we can show something more general. We call an r-graph a *generalized*

dumb-bell if it consists of r disjoint complete graphs $K_{k_1}, K_{k_2}, \ldots, K_{k_r}$ and an edge intersecting each complete graph in exactly one vertex.

Proposition 1.9. *Suppose an r-graph H is a generalized dumb-bell. Then* $\mathrm{Sat}(H,n) = O(n)$ *and in particular, for $r \geq 3$ we have that* $\frac{\mathrm{Sat}(H,n)}{n^{r-1}} \to 0$ *as n tends to infinity.*

Proof. Let H consist of r disjoint complete graphs $K_{k_1}, K_{k_2}, \ldots, K_{k_r}$ with $k_1 \leq k_2 \leq \ldots < k_r$, and an edge intersecting each complete graph in exactly one vertex.

Let $m = \sum_{i=1}^{r} k_i$ and let

$$l = \max\left\{\sum_{i=1}^{r-1} k_i, \sum_{i=\lceil \frac{r}{2} \rceil + 1}^{r} k_i\right\}.$$

Note that $m - l = \min\left\{k_r, \sum_{i=1}^{i=\lceil \frac{r}{2} \rceil} k_i\right\}$ is greater than 1 for $r \geq 3$, so long as H is not just a single edge (in which case we are done).

Consider a graph G consisting of disjoint complete graphs where each complete graph has size $\geq l$ and $< m$.

It is not hard to see that G is H-saturated. Since each complete graph has size $< m$, we know G is H-free. Let A be an r-set that is not an edge of G. Let A intersect the jth complete graph K_{t_j} in a_j vertices, where by reordering the K_{t_j} we assume that $a_1 \geq a_2 \geq \ldots \geq a_s > 0$ and $a_j = 0$ for $j > s$. We have that $s \geq 2$ and $a_1 + a_2 + \ldots a_s = r$.

Note that by our choice of l we have

$$\begin{aligned} k_1 + k_2 + \ldots k_{a_1} &\leq l \leq t_1, \\ k_{(a_1+1)} + k_{(a_1+2)} + \ldots + k_{(a_1+a_2)} &\leq l \leq t_2, \\ &\vdots \\ k_{(a_1+a_2+\ldots+a_{s-1}+1)} + k_{(a_1+a_2+\ldots+a_{s-1}+2)} + \ldots + k_r &\leq l \leq t_s. \end{aligned}$$

Thus adding the edge A to G will create a copy of H.

Finally, note that $l \leq m-2 \leq m-1 < m$ and so any G consisting of disjoint complete graphs on $m-2$ or $m-1$ vertices is H-saturated. For n sufficiently large we can always find such a graph G on n vertices and we have $e(G) < \binom{m-1}{r}\frac{n}{m-2} = O(n)$. □

In the case $r = 2$ we know the precise asymptotic behaviour of $\mathrm{Sat}(H, n)$. When $r > 2$ we know that $\mathrm{Sat}(H, n) = O(n)$ but it seems difficult to work out the coefficient of n in $\mathrm{Sat}(H, n)$. To do so we would need an argument giving a matching lower bound on the number of edges in an H-saturated graph.

Pikhurko gave an example of a family $\mathcal{F}$ of r-graphs where $\mathrm{Sat}(\mathcal{F}, n) = O(n)$ but $\mathrm{Sat}(\mathcal{F}, n)/n$ does not tend to any limit as n tends to infinity (Example 6 in [25]). One could ask whether the same thing is possible with a single graph rather than a family.

Question 1.3. *For all r does there exists an r-graph F such that $\mathrm{Sat}(F, n) = O(n)$ but $\mathrm{Sat}(F, n)/n$ does not tend to a limit as n tends to infinity?*

The generalized dumb-bell H might be a contender for a positive answer to this question.

A similar question could be asked at other scales, that is, for any $1 \leq t \leq r-1$ we can ask whether there is an F for which $\mathrm{Sat}(F, n) = O(n^t)$ but $\mathrm{Sat}(F, n)/n^t$ does not converge. In the case $t = r-1$ this is precisely Tuza's conjecture for r graphs.

We do not have a good guess for a single r-graph which is a counterexample to Tuza's conjecture, since H does not work. Instead we ask the weaker question of whether there is a smaller family with $\mathrm{sat}(\mathcal{F}, n)/n^{r-1}$ not converging.

Question 1.4. *Does there exists a family $\mathcal{F}$ containing fewer than four r-graphs such that $\frac{\mathrm{sat}(\mathcal{F},n)}{n^{r-1}}$ does not tend to a limit as n tends to infinity?*

Observe that the forbidden family used in section 1.3 actually only contains three hypergraphs when $r = 3$, so we have a positive answer for this special case.

The smallest 2-graph family given by Pikhurko that has $\mathrm{sat}(\mathcal{F}, n)/n$ not converging contains four graphs. A recent result of Chakraborti and Loh [7] improves this to a family of three graphs with $\mathrm{sat}(\mathcal{F}, n)/n$ not converging. Adapting this construction to r-graphs could be a good place to start looking for a smaller family to answer question 1.4.

A natural question that arises from this work is whether it is possible to combine the results of Theorems 1.6 and 1.7. It seems that it is non-trivial to combine the constructions in the two proofs to get a family with both desired properties.

Question 1.5. *Does there exist for all r a bounded size forbidden family of r-graphs where $\mathrm{sat}(\mathcal{F}, n)$ is asymptotically small on even numbers and asymptotically large on odd numbers?*

In going from a forbidden family of size that grows with r in Theorem 1.5 to a family of constant size in Theorem 1.6, we lost the large gap in the asymptotics for $\mathrm{sat}(\mathcal{F}, n)$. That is, in the case when n is divisible by k, the construction of a saturated graph with few edges went from having $\Theta(n)$ edges to having $\Theta\left(n^{r-2}\right)$ edges. Is it possible to retain the large difference in asymptotics and still decrease the size of the family? This seems difficult, especially if we try to reduce the family to a single graph.

Question 1.6. *Let F be an r-graph. Can $\mathrm{Sat}(F, n)$ be $O(n)$ for some infinite sequence of values of n and $\Omega(n^{r-1})$ for some other infinite sequence?*

An example of a class of r-graphs where the saturation number is $O(n)$ are the generalized dumb-bell graphs defined above.

If Tuza's conjecture is true, it would imply that the answer to Question 1.6 is 'no'. However, it may be easier to provide a negative answer to Question 1.6 than to prove Tuza's conjecture directly, and an answer might help provide ideas towards a full proof.

Chapter 2

Synchronizing Automata and Černý's Conjecture

2.1 Introduction

A (deterministic, finite) automaton Ω consists of a finite set of *states* (usually labelled $[n] = \{1, 2, \ldots, n\}$) and a finite set of *transition functions*, which are functions from the set of states to itself.

We shall be interested in the results of applying a sequence of transition functions to the set of states. We call such a sequence of transition functions a *word* of the automaton. The words of the automaton form a monoid, generated by the transition functions, which acts on the set of states.

We say that a word w of the automaton is a *reset word* if it sends every state to the same point; that is if $w(i) = w(j)$ for all i, j. We call an automaton *synchronizing* if it has a reset word.

Conjecture 2.1 (Černý's Conjecture)**.** *Suppose an automaton on n states is synchronizing. Then the automaton has a reset word of length at most $(n-1)^2$.*

This conjecture comes from a particular family of automata, which we shall refer to as the Černý automata. For each $n \geq 2$, we define an automata with states $\{1, 2, \ldots, n\}$ and two transition functions a and b, defined as follows:

$$a(i) = i + 1 \pmod{n} \qquad b(i) = \begin{cases} 2 & \text{if } i = 1 \\ i & \text{otherwise} \end{cases}$$

Figure 2.1 shows the Černý automaton for $n = 4$, which has shortest reset word $baaabaaab$. It is not too hard to check that the shortest reset word for the Černý automaton on n states has length $(n-1)^2$. Thus if Černý's conjecture were true it would be best possible.

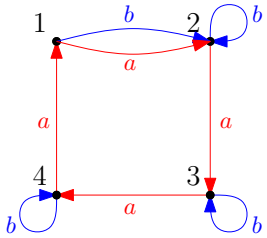


Figure 2.1: The Černý automaton for $n = 4$.

Černý's conjecture has been shown to hold for certain classes of automata, including orientable automata [10], automata where one transition function is a cyclic permutation of the states [9], and automata where the underlying digraph is Eulerian [16]. For a survey of these and other results see [32]. It remains open to prove the conjecture for all automata.

One can easily obtain a naive upper bound on the length of a shortest reset word. Note that for any pair of states there is some word sending them to a single state, since the automaton is synchronising. Applying the shortest such word, we will never pass through the same pair of states twice, or we could have found a shorter word. Thus the shortest word sending a given pair of states to a single state is of length at most $\binom{n}{2}$. Applying this repeatedly gives a reset word of length at most $\frac{n(n-1)^2}{2}$.

A better upper bound for the length of a minimal reset word comes from a result due

to Frankl and Pin [13] [26].

Theorem 2.1 (Frankl–Pin)**.** *Consider a synchronizing automaton with state set Ω of size n. Let $S \subseteq \Omega$ be a set of size k where $k \geq 2$. There exists a word w of length at most $\binom{n-k+2}{2}$ such that $|w(S)| < k$.*

Sketch of Proof. Let S be a set of size $k \geq 2$. Let w be a word of minimum length such that $|w(S)| < k$. Write w as a product of transition functions $w = f_m f_{m-1} \dots f_2 f_1$ where m is the length of w.

Let $S_0 = S$ and let $S_i = f_i(S_{i-1})$ for $i = 1, 2, \dots m$, so we have $S_m = w(S)$. There must be some pair of states $x \neq y$ in S such that $|w(\{x, y\})| = 1$. Let $P_0 = \{x, y\}$ and let $P_i = f_i(P_{i-1})$.

Since w is of minimal length we have $|S_i| = k$ and $|P_i| = 2$ for $i = 0, 1, \dots, m-1$, with $P_i \subseteq S_i$. We must also have that $P_j \not\subseteq S_i$ for $i < j$, else we would have that the shorter word $w' = f_m f_{m-1} \dots f_{j+1} f_i \dots f_2 f_1$ has $|w'(s)| < k$. We can rewrite these conditions in terms of S_i and the complements P_i^c as follows:

1. $|S_i| = k$ and $P_i^c = n - 2$ for $0 \leq i \leq m - 1$,
2. $S_i \cap P_i^c = \emptyset$ for $0 \leq i \leq m - 1$, and
3. $S_i \cap P_j^c \neq \emptyset$ for $0 \leq i < j \leq m - 1$.

If the final condition were replaced by $S_i \cap P_j^c \neq \emptyset$ for all pairs i, j, then Bollobás' two families theorem [4] (also known as the set pair method) would show that the number of set-pairs m is at most $\binom{(n-2)+k}{2}$. Frankl used linear algebra and symmetric tensor products to prove that the same holds in this more general case. □

Applying Theorem 2.1 repeatedly, we have that the length of a shortest reset word for an n-state synchronizing automaton is at most $\sum_{i=2}^{n} \binom{n-i+2}{2} = \frac{n^3-n}{6}$.

This was the best known upper bound until relatively recently. Slight improvements

to the constant factor have now been found: Szykuła [29] obtained an upper bound of $\approx \frac{114}{685}n^3 + O(n^2)$ and Shitov [27] refines this method to obtain an upper bound of $\approx 0.1654n^3 + o\left(n^3\right)$.

Let Ω be an automaton on $[n]$. The *transition graph* $\mathcal{T}(\Omega)$ has vertices the non-empty subsets of $[n]$, and for each set S and each transition function f a directed edge from S to $f(S)$ with label f. Figure 2.2 shows the transition graph for the Černý automaton in figure 2.1. The subsets of size k form the *kth layer* of the transition graph, written L_k.

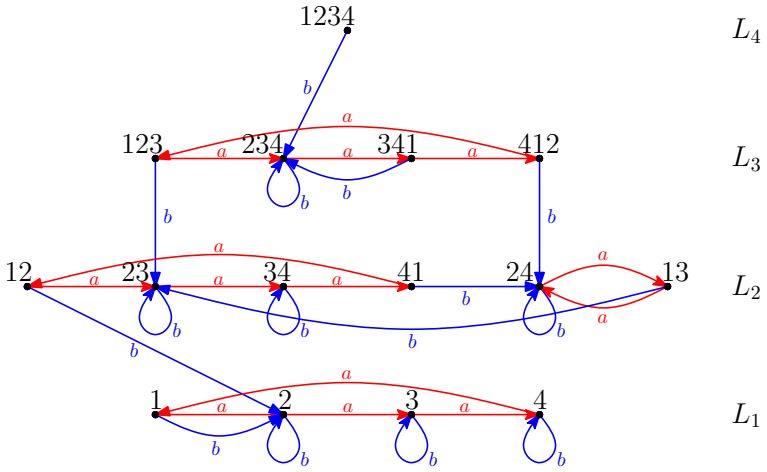


Figure 2.2: The transition graph for the Černý automaton on 4 vertices.

Now Černý's conjecture can be restated in terms of the transition graph:

Conjecture 2.1′ (Černý's conjecture)**.** *Let Ω be an automaton on $[n]$. If in the transition graph $\mathcal{T}(\Omega)$ there exists a path from $[n]$ to a vertex in L_1, then there exists such a path of length at most $(n-1)^2$.*

Conjecture 2.1′ suggests the following questions:

Question 2.2. *What is the minimal value of* $\mathrm{rdv}(k,n)$ *such for any synchronizing automaton Ω on $[n]$ there is a path in the transition graph $\mathcal{T}(\Omega)$ from L_k to L_1 of length at most* $\mathrm{rdv}(k,n)$*?*

Question 2.3. *What is the minimal value of* $\mathrm{RDV}(k,n)$ *such for any synchronizing automaton Ω on $[n]$ and for any k-set S there is a path in the transition graph $\mathcal{T}(\Omega)$*

from S to L_1 of length at most $\mathrm{RDV}(k,n)$*?*

Given an automaton Ω and a set of states S, we call S *synchronizable* if there exists a path from S to a singleton in the transition graph $\mathcal{T}(\Omega)$. Let the weight $t(S)$ of a set S be the shortest path from S to a singleton if S is synchronizable and ∞ otherwise. We define

$$m(k,\Omega) = \min\{t(S) : S \in L_k\}$$
$$M(k,\Omega) = \max\{t(S) : S \in L_k, S \text{ synchronizable}\}.$$

Then $\mathrm{rdv}(k,n)$ is the maximum of $m(k,\Omega)$ taken over all synchronizing automata and $\mathrm{RDV}(k,n)$ is the maximum of $M(k,\Omega)$ again taken over all synchronizing automata. It is clear that $\mathrm{rdv}(k,n) \leq \mathrm{RDV}(k,n)$ and $\mathrm{rdv}(k,n) \leq 1 + \mathrm{RDV}(k-1,n)$.

Answering either of these questions in the case $k = n$ is equivalent to answering Černý's conjecture. Note that finding a lower bound on $\mathrm{rdv}(k,n)$ or $\mathrm{RDV}(k,n)$ requires a construction of a suitable automaton with all k-sets having large weight or one k-set having large weight respectively; while finding an upper bound on $\mathrm{rdv}(k,n)$ or $\mathrm{RDV}(k,n)$ requires an argument about all synchronizing automata.

We know that $\mathrm{rdv}(2,n) = 1$ as all synchronising automata must have a transition function sending a pair to a singleton. We have $\mathrm{RDV}(2,n) \leq \binom{n}{2}$ since at worst some pair must travel through every other pair before being squashed to a singleton. In fact, $\mathrm{RDV}(2,n) = \binom{n}{2}$, where the example is given by the Černý automaton on $[n]$. Consider the pair $(2, \lfloor \frac{n}{2} \rfloor + 2)$ in this atuomaton: when n is even it takes $\frac{n}{2}$ steps to reduce the distance between the points on the cycle to $\frac{n}{2} - 1$ and n steps for each further reduction by one; when n is odd it takes n steps to reduce the distance between the points by one each time.

The Černý automaton also gives lower bounds for general k. We have that the minimum weight k-set is $\{1, 2, \ldots, k\}$ with weight $(k-2)n + 1$ and so $\mathrm{rdv}(k,n) \geq$

	lower bound	upper bound
$\mathrm{rdv}(3,n)$	$n+3$	$\frac{3-\sqrt{5}}{4}n^2+O(n)$
$\mathrm{rdv}(k,n)$	$(k-2)n+1$	$\frac{1}{2}\left\lfloor\frac{k-1}{2}\right\rfloor n^2+O(n)$
$\mathrm{RDV}(k,n)$	$\frac{k-1}{k}n(n-1)$	$\frac{k-1}{2}n^2+O(n)$

Table 2-A: Upper and lower bounds on $\mathrm{rdv}(k,n)$ and $\mathrm{RDV}(k,n)$.

$(k-2)n+1$. Using the fact that it takes n moves to get two states one step closer to each other on the cycle, the k-set with states equally spaced around the circle has weight $\geq \frac{(k-1)n(n-1)}{k}$ and so $\mathrm{RDV}(k,n) \geq \frac{(k-1)n(n-1)}{k}$.

For upper bounds, we can apply Theorem 2.1, which gives $\mathrm{rdv}(n,k) \leq 1+\sum_{i=2}^{k-1}\binom{n-i+2}{2}$ and $\mathrm{RDV}(n,k) \leq \sum_{i=2}^{k}\binom{n-i+2}{2}$.

An equivalent version of question 2.2 was asked by Gonze and Jungers [15], particularly for the case $k=3$. They call $\mathrm{rdv}(3,n)$ the *triple rendezvous time*, which inspires our notation rdv. We will sometimes refer to $\mathrm{rdv}(k,n)$ as the k-set rendezvous time.

Gonze and Jungers give a construction showing that $\mathrm{rdv}(3,n) \geq n+3$. They also claim a proof that $\mathrm{rdv}(3,n) \leq \frac{n(n+4)}{4}$, but this proof appears to be fundamentally flawed. We will outline the error in this proof in Section 2.2.

Our main result is an improved upper bound for the triple rendezvous time $\mathrm{rdv}(3,n)$ of $\frac{3-\sqrt{5}}{4}n^2+\frac{5-\sqrt{5}}{4}n \approxeq 0.19098n^2+O(n)$. We prove this upper bound in Section 2.2, together with a simple argument almost halving the upper bound on $\mathrm{rdv}(k,n)$ given by Frankl–Pin. We will also apply the techniques used on $\mathrm{rdv}(3,n)$ to further improve the upper bound for $\mathrm{rdv}(4,n)$ and $\mathrm{rdv}(5,n)$.

Table 2-A summarises what is known about $\mathrm{rdv}(3,n)$, $\mathrm{rdv}(k,n)$ and $\mathrm{RDV}(k,n)$, with the new results highlighted in red.

We can also ask similar questions over all automata, not just synchronizing automata.

Question 2.4. *What is the minimal value of* $\mathrm{rdv}^*(k,n)$ *such that for any automaton* Ω

on $[n]$*, if in the transition graph* $\mathcal{T}(\Omega)$ *there exists a path from some vertex in* L_k *to a vertex in* L_1*, then there is such a path of length at most* $\mathrm{rdv}^*(k,n)$*?*

Question 2.5. *What is the minimal value of* $\mathrm{RDV}^*(k,n)$ *such that for any* Ω *an automaton on* $[n]$*, if any* k*-set has a path to some vertex in* L_1*, then that* k*-set has such a path of length at most* $\mathrm{RDV}^*(k,n)$*?*

In particular, $\mathrm{rdv}^*(k,n)$ is the maximum of $m(k,\Omega)$ taken over all automata Ω with at least one synchronizable k-set, and $\mathrm{RDV}^*(k,n)$ is the maximum of $M(k,\Omega)$ over the same collection of automata.

Again we have that answering either question in the case $k = n$ is again equivalent to Černý's conjecture. Note that $\mathrm{rdv}^*(k,n) \leq \mathrm{RDV}^*(k,n)$ and $\mathrm{rdv}^*(k,n) \leq 1 + \mathrm{RDV}^*(k-1,n)$. A naive upper bound on $\mathrm{rdv}^*(k,n)$ is $1 + \sum_{i=2}^{k-1} \binom{n}{i}$, since a shortest word down to a singleton will take a set through each of set of size $< k$ at most once.

A very slightly improved upper bound can be obtained by noting that an automaton is synchronizing if and only if for every pair of states u, v there is a word w with $w(u) = w(v)$. If the automaton is synchronizing then we can use the Frankl–Pin bound. If not, then there is pair u, v that cannot be sent to the same state and any set containing both u and v is not synchronizable. The shortest path will not pass through any of these sets and so $\mathrm{rdv}^*(k,n) \leq 1 + \sum_{i=2}^{k-1} \left(\binom{n}{i} - \binom{n-2}{i-2} \right)$. In either case, for fixed k we have $\mathrm{rdv}^*(k,n) = O(n^{k-1})$ and by the same argument $\mathrm{RDV}^*(k,n) = O(n^k)$.

In section 2.3 we show that this is, surprisingly, best possible — that is, if k is fixed then the answer to question 2.4 is $\Theta\left(n^{k-1}\right)$. Since $\mathrm{RDV}^*(k,n) \geq \mathrm{rdv}^*(k+1,n) - 1$, we also get that $\mathrm{RDV}^*(k,n) = \Theta\left(n^k\right)$.

Table 2-B summarises what is known about $\mathrm{rdv}^*(k,n)$ and $\mathrm{RDV}^*(k,n)$. The new contributions are highlighted in red.

	lower bound	upper bound
$\mathrm{rdv}^*(3,n)$	$\frac{1}{8}n^2$	$\frac{1}{2}n(n-1)$
$\mathrm{rdv}^*(k,n)$	$\frac{4}{3}\left(\frac{n}{4k}\right)^{k-1}$	$\binom{n}{k-1}+O\left(n^{k-2}\right)$
$\mathrm{RDV}^*(k)$	$\frac{4}{3}\left(\frac{n}{4(k+1)}\right)^{k}-1$	$\binom{n}{k}+O\left(n^{k-1}\right)$

Table 2-B: Upper and lower bounds on $\mathrm{rdv}^*(k,n)$ and $\mathrm{RDV}^*(k,n)$.

2.2 Upper Bounds on the Rendezvous Time

Frankl–Pin gives trivially that for $2 \le k \le n$, the k-rendezvous time ω is at most $1+\sum_{i=2}^{k-1}\binom{n-i+2}{2}$. The following simple adaptation of Frankl–Pin's result improves on this bound for $k \ge 4$.

Theorem 2.2. *For all n and all $2 \le k \le n$ the k-set rendezvous time $\mathrm{rdv}(k,n)$ is at most*

$$\sum_{i=1}^{\lfloor \frac{k}{2} \rfloor}\binom{i+1}{2}+\sum_{i=1}^{\lceil \frac{k}{2} \rceil -1}\binom{n-i+1}{2}.$$

In particular, for fixed k and n sufficiently large given k we have that

$$\mathrm{rdv}(k,n) < \left\lfloor \frac{k-1}{2} \right\rfloor \frac{n^2}{2}.$$

Proof. By Frankl–Pin, there exists a word w that takes $[n]$ to a set S of size $n-\lfloor \frac{k}{2} \rfloor$ of length at most

$$\sum_{i=n-\lfloor \frac{k}{2} \rfloor+1}^{n}\binom{n-i+2}{2}=\sum_{i=1}^{\lfloor \frac{k}{2} \rfloor}\binom{i+1}{2}.$$

By the pigeonhole principle, there are at least $n-2\lfloor \frac{k}{2} \rfloor$ points in S with exactly one point in their preimage under w. Take T to be $n-k$ such points and let $R=S\setminus T$. We have $|R|=\lceil \frac{k}{2} \rceil$ and $|w^{-1}(R)|=n-|w^{-1}(T)|=n-|T|=k$.

By Frankl–Pin again, we can find a word w' that takes R to a singleton of length at

most

$$\sum_{i=2}^{\lceil \frac{k}{2} \rceil} \binom{n-i+2}{2} = \sum_{i=1}^{\lceil \frac{k}{2} \rceil - 1} \binom{n-i+1}{2}.$$

Concatenating $w'w$ gives the required word. □

We can also obtain an improved rendezvous bound for $k = 3, 4$ and 5. The triple rendezvous time $\mathrm{rdv}(3, n)$ was studied in particular by Gonze and Jungers [15] who claimed a proof that it was bounded by $\frac{n^2}{4}$, a proof which we believe to be fundamentally flawed.

Theorem 2.3. *For all $n \geq 3$, we have* $\mathrm{rdv}(3, n) \leq \frac{3-\sqrt{5}}{4} n^2 + \frac{3}{2} n$.

Note that $\frac{3-\sqrt{5}}{4} \approxeq 0.19098$, so this does significantly better than the $1 + \binom{n}{2}$ given by Frankl–Pin, as well as the $\frac{n^2}{4}$ claimed by Gonze and Jungers. The proof introduces ideas that will be further built on to improve the bounds for $\mathrm{rdv}(4, n)$ and $\mathrm{rdv}(5, n)$.

Proof. First, note that if $n = 3$ then the triple rendezvous time is at most 4 and the result is trivially true. Thus we may assume that $n \geq 4$.

The rank of a word w is the number of points in the image $\operatorname{Im} w = w([n])$. Let r be the minimum rank over all words of length at most n. Note that by Frankl–Pin there is a word of length $4 = \binom{2}{2} + \binom{3}{2}$ that takes $[n]$ to a set of size $n - 2$. Since $n \geq 4$ we thus have that $r \leq n - 2$.

Let w be a word of length $\leq n$ of minimal rank r. If $r < \frac{n}{2}$ then by the pigeonhole principle there must be some triple sent to a singleton by w and so we have triple rendezvous time at most n. We may therefore assume that $r \geq \frac{n}{2} \geq 2$.

Claim 2.3.1. There exists a word of length $\leq n + \binom{r+2}{2}$ that takes some triple to a singleton.

Proof of Claim. Let w be a word of length $\leq n$ of minimal rank r. If there is some triple which w sends to a singleton then we are done, so we can assume that w sends at most

two points to the same point.

Let $S = \{x : \exists y \neq z \text{ with } w(y) = w(z) = x\}$ be the set of points with two pre-images under w. Let $T = \operatorname{Im} w - S$ be the set of all points with a unique pre-image under w. We have that $|S| + |T| = r$ and $2|S| + |T| = |w^{-1}([n])| = n$, from which we obtain $|S| = n - r$.

By Frankl–Pin there exists a word w' of length $\leq \binom{n-(n-r)+2}{2} = \binom{r+2}{2}$ such that $|w'(S)| < |S|$. In particular, there exist $x \neq y$ in S with $w'(x) = w'(y) = z$. Take u, v with $w(u) = w(v) = x$ and s, t with $w(s) = w(t) = y$. The word $w'w$ has length at most $n + \binom{r+2}{2}$ and $w'w(\{u, v, s, t\}) = w'\{x, y\}) = \{z\}$ so in this case $w'w$ sends some 4-set to a single point.

Claim 2.3.2. There exists a word of length $\leq n + \frac{(n-r)n}{2}$ that takes some triple to a singleton.

Proof of Claim. Let C be the minimal non-empty set such that $f(C) \subseteq C$ for all transition functions f. In particular, if w is any synchronising word and x is the vertex with $w([n]) = x$ then $C = \{y : \exists w' \text{ with } w'(x) = y\}$. Note that C is strongly connected, by which we mean that for any pair of points $u, v \in C$ we can find a word that sends u to v and a word that sends v to u. Let $m = |C|$ and let $E = \{(u, v) : u, v \in C, u \neq v\}$ be the set of pairs of points in C.

We call an pair $(u, v) \in E$ *good* if there exists a word w_{uv} of length $\leq n$ with $|w_{uv}^{-1}(\{u, v\})| \geq 3$. We will count the number of good pairs, splitting into three cases depending on the value of m.

First, suppose that $m > n - r + 1$. We can find a point $z \in C$ and a transition function f with $|f^{-1}(z)| \geq 2$. Since C is strongly connected, for each $v \in C$ there is some word w_v of length $\leq m - 1$ with $w_v(z) = v$. In particular, $(w_v f)^{-1}(v) \supseteq f^{-1}(z)$ where $w_v f$ is a word of length $\leq m$. For all $a \in (\operatorname{Im}(w_v f) - v) \cap C$ the pair (v, a) is good.

Since $w_v f$ is a word of length $\leq m \leq n$, the rank of $w_v f$ is $\geq r$ and so $|\,(\mathrm{Im}\,(w_v f) - v) \cap C| \geq r - 1 - (n - m) > 0$. Every vertex in C is in at least $m - 1 - n + r$ good pairs and so the number of good pairs is at least $\frac{(m-1-n+r)m}{2} > 0$.

The number of pairs in E that are not good is at most $\binom{m}{2} - \frac{(m-1-n+r)m}{2} = \frac{(n-r)m}{2}$. We conclude that there is some word w of length at most $\frac{(n-r)m}{2}$ that sends some good pair (u, v) to a singleton x, where the worst case scenario is having to pass through every not good pair in C first.

By definition of good, we can find a word w_{uv} of length $\leq n$ with $|w_{uv}^{-1}(\{u, v\})| \geq 3$. Then ww_{uv} is a word of length at most $n + \frac{(n-r)m}{2}$ where $(ww_{uv})^{-1}(x) \supseteq w_{uv}^{-1}(\{u, v\})$ has size at least 3. In particular, the claim holds when $m > n - r + 1$.

Now, suppose that $2 \leq m \leq n - r + 1 \leq n - 1$. There must be some $y \notin C$ and some transition function f such that $f(y) \in C$. We have that $f(C \cup \{y\}) \subseteq C$ and so by the pigeonhole principle there is some x in C with two pre-images under f. If x is the only point in $\mathrm{Im}\, f \cap C$ then $f^{-1}(x) \supseteq C \cup \{y\}$ has size $m + 1 \geq 3$, and we get that f takes a triple down to the single point x. Otherwise, there is another point y in $\mathrm{Im}\, f \cap C$. We can find a word w of length at most $\binom{m}{2}$ that takes (x, y) to a singleton, where the worst case scenario is having to pass through every other pair in C. Now wf is a word of length at most $1 + \binom{m}{2}$ that takes some triple to a singleton.

Finally, suppose $m = 1$, so $C = \{x\}$ for some point x. Note that $f(x) = x$ for all f. There is some point $y \neq x$ and some transition function f such that $f(y) = x$. The rank of f is at least r, so there is some point z in $\mathrm{Im}\, f - x$ and some word w of length at most $n - r + 1$ with $w(z) = x$, where the worst case scenario is having to pass through every point in $[n] \setminus \mathrm{Im}\, f$. Now wf is a word of length at most $2 + n - r$ that takes some triple to the singleton x.

To summarise, we can find a word taking some triple to a singleton of the following

length:

$$\begin{cases} m + \frac{(n-r)m}{2} & \text{if } m > n - r + 1 \\ 1 + \binom{m}{2} & \text{if } 2 \leq m \leq n - r + 1 \\ 2 + n - r & \text{if } m = 1 \end{cases}$$

each of which is at most $n + \frac{(n-r)n}{2}$ as required.

Combining the results of these two claims, we have that the triple rendezvous time is at most $\min\left\{n + \binom{r+2}{2}, n + \frac{(n-r)n}{2}\right\}$.The former is increasing in r while the latter is decreasing in r and so to find the maximum we look for the r where they are equal. This is when $(r+2)(r+1) = (n-r)n$, which occurs when $r = \frac{-n-3+\sqrt{5n^2+6n+1}}{2}$ (subject to $r \geq 0$).

Substituting this in gives that the triple rendezvous time is at most

$$\begin{aligned} n + \frac{\left(3n + 3 - \sqrt{5n^2 + 6n + 1}\right) n}{4} &\leq n + \frac{\left(3n + 3 - \sqrt{5}\left(n + \frac{1}{\sqrt{5}}\right)\right) n}{4} \\ &= n + \frac{\left((3 - \sqrt{5})n + 2\right) n}{4} \\ &= \frac{3 - \sqrt{5}}{4} n^2 + \frac{3}{2} n. \end{aligned}$$

□

In a similar way we can improve the upper bounds on the 4-set and 5-set rendezvous times.

Theorem 2.4. *For all $n \geq 4$, we have*

$$\mathrm{rdv}(4, n) \leq \mathrm{rdv}(3, n) + (2 - \sqrt{3})n^2 + 2n - 1 \leq \left(\frac{11 - \sqrt{5} - 4\sqrt{3}}{4}\right) n^2 + \frac{7}{2} n - 1.$$

Note that $\frac{11-\sqrt{5}-4\sqrt{3}}{4} \approxeq 0.4589$, so this is again an improvement on the $4+\binom{n}{2}$ given by Theorem 2.2.

Theorem 2.5. *For all $n \geq 5$, we have*

$$\operatorname{rdv}(5,n) \leq \operatorname{rdv}(4,n) + \frac{4-\sqrt{7}}{4}n^2 + \frac{3}{2}n - 1 \leq \left(\frac{15-\sqrt{5}-4\sqrt{3}-\sqrt{7}}{4}\right)n^2 + 5n - 2.$$

Note that $\frac{15-\sqrt{5}-4\sqrt{3}-\sqrt{7}}{4} \approxeq 0.7975$, so this is again an improvement on the bound of $4+\binom{n}{2}+\binom{n-1}{2}$ given by Theorem 2.2.

To prove both Theorems 2.4 and 2.5 we will need two lemmas. In the case $k=2$ the lemmas correspond precisely to claims 2.3.1 and 2.3.2 in the proof of Theorem 2.3 and we prove each lemma in an anlogous way.

Lemma 2.6. *Fix $2 \leq k \leq n-1$ and $l \geq 1$. Let r be the minimal rank over all words of length $\leq l$. Suppose that $r \leq \frac{n-2\left\lceil \frac{k}{2} \right\rceil}{\left\lfloor \frac{k}{2} \right\rfloor}$ and let $s = \frac{n-\left\lfloor \frac{k}{2} \right\rfloor r}{\left\lceil \frac{k}{2} \right\rceil} \geq 2$. Then*

$$\operatorname{rdv}(k+1,n) \leq l + \binom{n-s+2}{2}.$$

Proof. Let w be a word of length $\leq l$ of minimal rank r. If there is some $(k+1)$-set which w sends to a singleton then we are done, so we can assume that w sends at most k points to the same point.

Let $S = \{x : \left|w^{-1}(x)\right| \geq \left\lceil \frac{k+1}{2} \right\rceil\}$ be the set of points with at least $\left\lceil \frac{k+1}{2} \right\rceil$ pre-images under w. Let $T = \operatorname{Im} w - S$ be the set of points with $\leq \left\lfloor \frac{k}{2} \right\rfloor$ pre-images under w. We have that $|S|+|T| = r$. We also have that $n = |w^{-1}([n])| = |w^{-1}(S)|+|w^{-1}(T)| \leq k|S| + \left\lfloor \frac{k}{2} \right\rfloor |T|$. Putting these together, we have $k|S|+\left\lfloor \frac{k}{2} \right\rfloor (r-|S|) \leq n$ and so $|S| \geq \frac{n-\left\lfloor \frac{k}{2} \right\rfloor r}{\left\lceil \frac{k}{2} \right\rceil} \geq 2$.

By Frankl–Pin there exists a word w' of length $\leq \binom{n-|S|+2}{2}$ such that $|w'(S)| < |S|$. In particular, there exist $x \neq y$ in S with $w'(x) = w'(y) = z$. We have that $\left|(w'w)^{-1}(z)\right| = |w^{-1}(x,y)| \geq 2\left\lceil \frac{k+1}{2} \right\rceil$, so $w'w$ sends some $(k+1)$-set to a single point.

The length of the word $w'w$ is $l + \binom{n-|S|+2}{2} \leq l + \binom{n-s+2}{2}$ where $s = \frac{n-\lfloor \frac{k}{2} \rfloor r}{\lceil \frac{k}{2} \rceil}$. □

Lemma 2.7. *Fix $n \geq 4$ and $k \leq n-1$. Let r be the minimal rank over all words of length $\leq \mathrm{rdv}(k,n) + n - 1$. Then $\mathrm{rdv}(k+1,n) \leq \mathrm{rdv}(k,n) + n - 1 + \frac{(n-r)n}{2}$.*

Proof. Let C be the minimal non-empty set such that $f(C) \subseteq C$ for all transition functions f. Note that since C is minimal it must be strongly connected. Let $m = |C|$ and let $E = \{(u,v) : u,v \in C, u \neq v\}$ be the set of pairs of points in C.

Let $\mathrm{rdv}(k,n) = l$. We call an pair $(u,v) \in E$ *good* if there exists a word w_{uv} of length $\leq l+n-1$ with $|w_{uv}^{-1}(\{u,v\})| \geq k+1$. We will count the number of good pairs, splitting into two cases depending on the value of m.

Case 1: $m > n - r + 1$.

There is a word ω of length $l = \mathrm{rdv}(k,n)$ that sends some k-set to a singleton x. We can then find a word ω' of length at most $n-m$ that sends x to a point $z \in C$. In particular, $|(\omega'\omega)^{-1}(z)| \geq k$ where $\omega'\omega$ is a word of length $\leq l+n-m$.

Since C is strongly connected, for each $v \in C$ there is some word ω_v of length $\leq m-1$ with $\omega_v(z) = v$. In particular, $(\omega_v\omega'\omega)^{-1}(v) \supseteq (\omega'\omega)^{-1}(z)$ where $\omega_v\omega'\omega$ is a word of length $\leq l+n-1$. For all $a \in (\mathrm{Im}\,(\omega_v\omega'\omega) - v) \cap C$ the pair (v,a) is good.

Since $\omega_v\omega'\omega$ is a word of length $\leq l+n-1$, the rank of $\omega_v\omega'\omega$ is $\geq r$ and so $|\,(\mathrm{Im}\,(\omega_v\omega'\omega) - v) \cap C| \geq r-1-(n-m) > 0$. Every vertex in C is in at least $m-1-n+r$ good pairs and so the number of good pairs is at least $\geq \frac{(m-1-n+r)m}{2} > 0$.

The number of pairs in E that are not good is at most $\binom{m}{2} - \frac{(m-1-n+r)m}{2} = \frac{(n-r)m}{2}$. We conclude that there is some word w of length at most $\frac{(n-r)m}{2}$ that sends some good pair (u,v) to a singleton x, where the worst case scenario is having to pass through every not good pair in C first.

By definition of good, we can find a word w_{uv} of length $\leq n$ with $|w_{uv}^{-1}(\{u,v\})| \geq k+1$.

Then ww_{uv} is a word of length at most $l+n-1+\frac{(n-r)n}{2}$ where $(ww_{uv})^{-1}(x) \supseteq w_{uv}^{-1}(\{u,v\})$ has size at least $k+1$. In particular, the claim holds when $m > n-r+1$.

Case 2: $m \leq n-r+1$.

We will show that there is a word of length $\leq l + 2(n-m) + \binom{m}{2}$ that takes some $(k+1)$-set to a singleton.

As before, we can find a point $z \in C$ and a word w of length $\leq l + (n-m)$ with $|(w)^{-1}(z)| \geq k$. If z is the only point in $\mathrm{Im}(w)$ then w is a synchronizing word sending $n \geq k+1$ points down to a singleton.

So suppose $\mathrm{Im}(w) - z$ is non-empty and take some $v \in \mathrm{Im}(w)$, $v \neq z$. We can find a word w' of length $\leq n-m$ that takes v to a vertex $y \in C$.

If $w'(z) = y$ then the word $w'w$ of length $\leq l + 2(n-m)$ has $|(w'w)^{-1}(y)| = |w^{-1}(z,v)| \geq k+1$.

Else, $y, w'(z)$ are two distinct vertices in C and we can find a word w'' of length at most $\binom{m}{2}$ that takes $\{y, w'(z)\}$ to a singleton. Then $w''w'w$ is a word of length $\leq l + 2(n-m) + \binom{m}{2}$ that has takes some $(k+1)$-set to a singleton.

We have that there is a word of length $\leq l + 2(n-m) + \binom{m}{2}$ taking some $(k+1)$-set to a singleton. To prove that the bound as stated in the lemma holds, it suffices to show that the following quantity is positive.

$$\left(n-1+\frac{(n-r)n}{2}\right) - \left(2(n-m)+\binom{m}{2}\right) = \frac{n(n-r-2)-m^2+5m-2}{2}$$

Note that by Frankl–Pin there is some word of length $\leq \binom{2}{2} + \binom{3}{2} = 4$ of rank $n-2$. Since $l+n-1 \geq 4$, this implies that $r \leq n-2$, and so $n-r-2 \geq 0$.

If $-m^2+5m-2 \geq 0$ then we are done, so in particular we are done if $m \leq 4$. If

$m \geq 5$, then we have

$$\begin{aligned}\frac{n(n-r-2)-m(m-5)-2}{2} &\geq \frac{(m+r-1)(m-3)-m^2+5m-2}{2} \\ &= \frac{(m-3)r+m+1}{2} \geq 0\end{aligned}$$

□

We use these lemmas to prove the Theorems.

Proof of Theorem 2.4. Fix $n \geq 4$. Let $l = \mathrm{rdv}(3,n) + n - 1$ and let r be the minimal rank of a word of length at most l.

Applying the $k = 3$ case of Lemmas 2.6 and 2.7 we get

$$\mathrm{rdv}(4,n) \leq \begin{cases} l + \frac{1}{2}\left(\frac{n+r}{2}+2\right)\left(\frac{n+r}{2}+1\right) & \text{if } r \leq n-4 \\ l + \frac{1}{2}(n-r)n & \text{for all } r \end{cases}$$

If $r > n-4$ then $\frac{1}{2}(n-r)n < 2n$.

If $r \leq n-4$, the first bound is increasing with r (for $r \geq 0$) and the second is decreasing with r so the maximum is obtained where the two are equal, that is when $\left(\frac{n+r}{2}+2\right)\left(\frac{n+r}{2}+1\right) = (n-r)n$. Rearranging gives $r^2 + 6(n+1)r - (3n^2 - 6n - 8) = 0$. Solving for r, we get that the maximum is obtained when

$$r = -3(n+1) + \sqrt{12n^2+12n+1}$$

Thus we have that the maximum is

$$\begin{aligned}\frac{(n-r)n}{2} &= \frac{\left(4n+3-\sqrt{12n^2+12n+1}\right)n}{2} \\ &\leq \left(2n+\frac{3}{2}-\sqrt{3}\left(n+\frac{1}{\sqrt{12}}\right)\right)n \\ &= (2-\sqrt{3})n^2+n\end{aligned}$$

Putting this together with the bound on $\mathrm{rdv}(3,n)$ from Theorem 2.3 we get the final bound

$$\mathrm{rdv}(4,n) \leq \left(\frac{3-\sqrt{5}}{4}+2-\sqrt{3}\right)n^2+\frac{7}{2}n-1.$$

□

Proof of Theorem 2.5. Fix $n \geq 5$. Let $l = \mathrm{rdv}(4,n)+n-1$ and let r be the minimal rank of a word of length at most l.

Applying the $k=4$ case of Lemmas 2.6 and 2.7 we get

$$\mathrm{rdv}(5,n) \leq \begin{cases} l+\frac{1}{2}\left(\frac{n+2r}{2}+2\right)\left(\frac{n+2r}{2}+1\right) & \text{if } r \leq \frac{n-4}{2} \\ l+\frac{1}{2}(n-r)n & \text{for all } r \end{cases}$$

If $r > \frac{n-4}{2}$ then $\frac{1}{2}(n-r)n < \frac{1}{4}n^2+n$.

If $r \leq \frac{n-4}{2}$, the first bound is increasing with r (for $r \geq 0$) and the second is decreasing with r so the maximum is obtained where the two are equal, that is when $\left(\frac{n+2r}{2}+2\right)\left(\frac{n+2r}{2}+1\right) = (n-r)n$. Rearranging gives $r^2+(2n+3)r-(\frac{3}{4}n^2-\frac{3}{2}n-2)=0$. Solving for r, we get that the maximum is obtained when

$$r = \frac{-(2n+3)+\sqrt{7n^2+6n+1}}{2}$$

Thus we have that the maximum is

$$\begin{aligned}\frac{(n-r)n}{2} &= \frac{\left(4n+3-\sqrt{7n^2+6n+1}\right)n}{4} \\ &\leq \frac{\left(4n+3-\sqrt{7}\left(n+\frac{1}{\sqrt{7}}\right)\right)n}{4} \\ &= \frac{4-\sqrt{7}}{4}n^2+\frac{1}{2}n\end{aligned}$$

Putting this together with the bound on $\mathrm{rdv}(4,n)$ from Theorem 2.3 we get the final bound.

□

It is clear that we could continue applying this method in the way we have here to obtain upper bounds on $\mathrm{rdv}(k,n)$ for larger k. However, as it stands the method does not give an improvement on the bound $\mathrm{rdv}((k,n) < \left\lfloor \frac{k-1}{2} \right\rfloor \frac{n^2}{2}$ given by Theorem 2.2 for larger k. We remain hopeful that the method could be improved upon to give results for larger k. One approach might be to alter Lemma 2.6 to allow one to go directly from a result about $\mathrm{rdv}(k,n)$ to a result about $\mathrm{rdv}(k+c,n)$ for c larger than 1.

2.2.1 The Error in Gonze and Jungers

Gonze and Jungers [15] claim a proof that the triple-rendezvous time $r(3,n) < \frac{n^2}{4}$. However, there is an error in the proof of Theorem 3.8.

To identify the error we will first need to explain some of the notation used. Fix an automaton Ω on n states and let T_3 be the minimum weight of a triple. If $t < T_3$ the sets of weight $\leq t$ will only be singletons and pairs. We let G_t be the graph on n vertices with edge-set all pairs of weight $\leq t$.

Let $A(t)$ be a matrix with rows indexed by $[n]$ and columns indexed by the sets in Ω of weight $\leq t$. A column corresponding to the set S will have a 1 in rows indexed by elements of S and a 0 in all other rows. For example, $A(0)$ will be the identity matrix

since a set has weight 0 is and only if it is a singleton.

Define $Prog_{A(t)}$ to be the linear program

$$\begin{aligned} &\max_{q,k} k \\ \text{s.t.} &A(t)\mathbf{q}^T \geq k\mathbf{e}^T \\ &\mathbf{e}\mathbf{q}^T = 1 \\ &\mathbf{q} \geq 0 \end{aligned}$$

where $\mathbf{e}$ is the all ones vector. Let $k(t)$ be the maximum value attained by $Prog_{A(t)}$ and let P_t be the set of optimal solutions $\mathbf{q}$ to $Prog_{A(t)}$.

The linear program $Prog_{A(t)}$ can be thought of in terms of assigning weights to G_t. The vector $\mathbf{q}$ assigns a weight to each vertex and edge of G_t such that the sum of all the weights is one. The condition $A(t)\mathbf{q}^T \geq k\mathbf{e}^T$ means that for each vertex v the sum of the weight of v and the weights of all edges incident to v is at least k. Then $k(t)$ is maximal subject to a weighting of G_t existing that satisfies these conditions. Figure 2.3 demonstrates this for the Černý automaton on 4 vertices.

It is interesting to note that if we take the dual of this linear program and rescale by $\frac{1}{k}$ then we get the fractional independence number of G_t. A weighting of the vertices of G_t is a *fractional independent set* if each vertex has weight in $[0, 1]$ and for each edge of G the weights of the endpoints of the edge sum to at most one. The *fractional independence number* is the maximum total weight (that is, sum of all vertex weights) of a fractional independent set.

Note also that the complement of a fractional independent set (that is, where the new weight of a vertex is one minus its previous weight) is a *fractional vertex cover*, where each vertex has weight in $[0, 1]$ and for each edge of G the weights of the endpoints of the edge sum to at least one. The *fractional vertex cover number* is the minimum total

weight of a fractional vertex cover. In particular, the fractional independence number is n minus the fractional vertex cover number, and the vertex cover number is the form in which it is more commonly studied.

Lemma 3.6 of [15] proves that if $t < T_3$, there exists a collection C_t of pairs of weight $\leq t$ such that (a) the graph with edge-set C_t is composed of disjoint singletons, pairs, and odd cycles, and (b) if $A_c(t)$ has columns indexed by C_t then the linear program $Prog_{A_c(t)}$ has the same maximum, $k(t)$, as $Prog_{A(t)}$ does. Let R_t be the set of optimal solutions to $Prog_{A_c(t)}$.

Note that $k(t) \leq k(t+1)$, since G_t is a subgraph of G_{t+1}. Note also that $2/k(t)$ must be an integer – Gonze and Jungers show this using C_t but it can also be deduced from the fact that the fractional vertex cover number of any graph is half an integer.

The strategy of Gonze and Jungers' proof is to show that either $k(t+1) < k(t)$, or $k(t+1) = k(t)$ and $\dim(P_{t+1}) > \dim(P_t)$. Since $2/k(t)$ is an integer bounded between 1 and $2n$ and the dimension of P_t is also bounded, this would give a bound on the maximal value of t. This claim (the basis of Theorem 3.8 of [15]) could potentially hold, as we have found no counterexample.

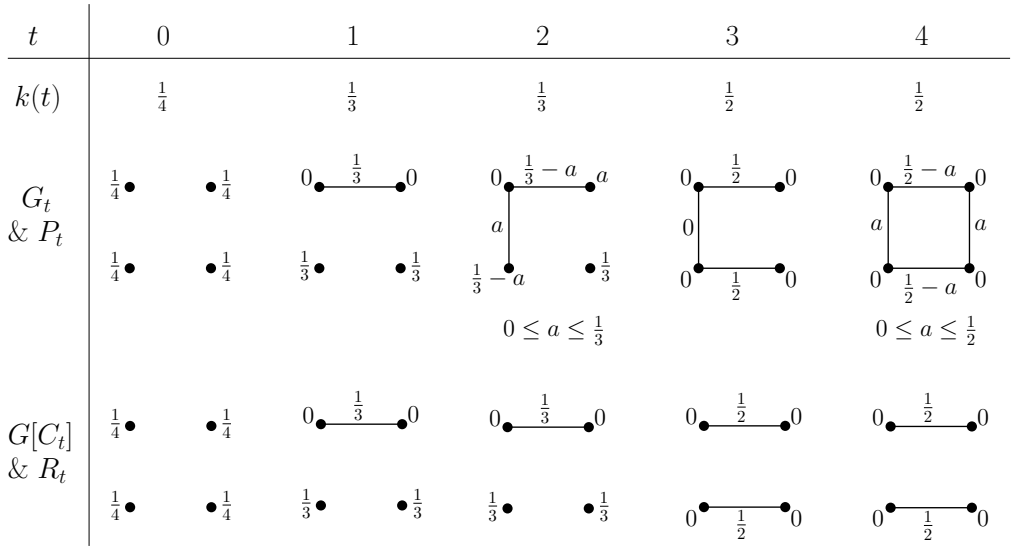


Figure 2.3: The graphs G_t and $G[C_t]$ for the Černý automaton on 4 states, with the sets of optimal solutions P_t and R_t indicated.

However, the authors' attempt to prove this claim by passing to R_t and this is where the problem lies. They claim that the same holds for R_t, that is if $k(t+1) = k(t)$ then $\dim(R_{t+1}) > \dim(R_t)$. This is not true in general. For example, it does not hold for the Černý automaton on 4 states where we have $G[C_1] = G[C_2]$ and $G[C_3] = G[C_4]$ and so the optimal solution sets $R_1 = R_2$ and $R_3 = R_4$. See Figure 2.3 for a diagram.

2.3 Non-sychronizing Automata with Large Rendezvous Time

We will prove a lower bound on $\mathrm{rdv}^*(k, n)$ via a construction of a suitable automaton. To introduce the main idea of the construction we give the simpler $k = 3$ case first.

Theorem 2.8. *For n sufficiently large,* $\mathrm{rdv}^*(3, n) > \frac{n^2}{8}$.

Proof. We will construct an automaton on $[n]$ where the minimal weight of a k-set is greater than $\frac{n^2}{8}$.

Partition $[n]$ into A and X, where $|A| = \lfloor \frac{n}{4} \rfloor$.

Label the vertices of A by $a_1, a_2, \ldots, a_{|A|}$ and label the vertices of X by $x_1, x_2, \ldots, x_{|X|}$.

Take two functions f and g as follows, as shown in figure 2.4.

$$f(x_t) = x_{(t+1 \mod |X|)}$$
$$f(a_{|A|}) = a_1$$
$$f(a_j) = x_j \text{ for } j \neq |A|$$

$$g(x_t) = \begin{cases} x_{t+1} & \text{if } 1 \leq t \leq |A| - 1 \\ x_{t-|A|+1} & \text{if } t = |A| \\ x_t & \text{otherwise} \end{cases}$$

$$g(a_j) = a_{(j+1 \mod |A|)}$$

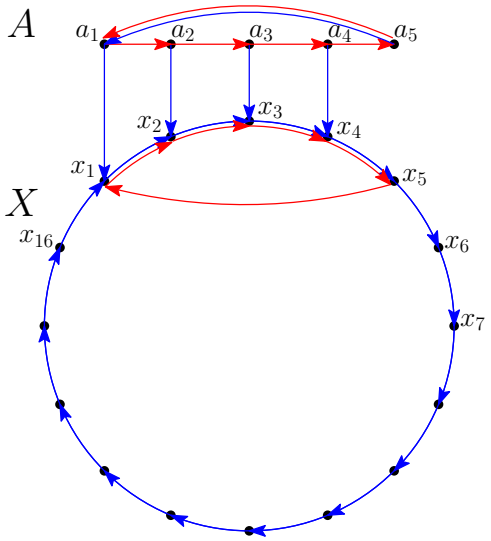


Figure 2.4: An example of the automaton for $k = 3$ and $n = 21$

Note that f and g restricted to X are permutations on X and so any set containing more than one vertex in X cannot be synchronized. Moreover, any set containing three vertices in A cannot be synchronized: the image of such a set under g still has three vertices in A, and the image under f contains two vertices in X.

It follows that a synchronizable triple must contain two vertices in A and one vertex

in X. Fix such a triple S and consider a word that synchronizes this set acting on it. We will obtain that the triple of minimal weight is in fact $\{x_{|X|}, a_1, a_{|A|}\}$.

Note that for a shortest word from a triple to a singleton the first step must map a triple to a pair. In particular, the first map of the shortest word must be f, as g is a permutation. The triple S must contain two points in A, one of which must be $a_{|A|}$ else applying f gives two points in X. Let the other be a_t, where $1 \leq t \leq |A| - 1$. After applying f, we have the points a_1 and x_t, which must be the only point in X.

Note that

$$fg^{l-1}(a_1) = \begin{cases} a_1 & \text{if } l \equiv 0 \pmod{|A|} \\ x_{(l \mod |A|)} & \text{otherwise} \end{cases}$$

and for $1 \leq t \leq |A| - 1$,

$$fg^{l-1}(x_t) = \begin{cases} x_{|A|+1} & \text{if } t + l - 1 \equiv 0 \pmod{|A|} \\ x_{(t+l \mod |A|)} & \text{otherwise} \end{cases}.$$

This means that applying fg^{l-1} gives two points in X for any $l \not\equiv 0 \pmod{|A|}$. Thus the next step must be to apply fg^{l-1} where l is some multiple of $|A|$. This sends a_1 and x_t to themselves unless $t = 1$, in which case x_t is sent to $x_{|A|+1}$.

To further reduce the size of the set, we must map $x_{|A|+1}$ and a_1 to the same point. To do this, we must move the vertex in position $x_{|A|+1}$ round through $x_{|A|+2}, x_{|A|+3}, \ldots$ until we reach $x_{|X|}$, without moving the second vertex that is currently in A into X as we do so.

Suppose we have just applied f, and we now want to move x_s to x_{s+1} without adding any extra vertices into X (where s is some value not in $\{|X|, 1, 2, 3 \ldots, |A|\}$). Since we have just applied f, the vertex in A must be at position a_1 (having just come from position $a_{|A|}$). We need to apply f to move x_s, but we can only apply f when the vertex in A is at position $a_{|A|}$ and so we must first apply $g^{|A|-1}$ to move the vertex at a_1 to

be at $a_{|A|}$. Only then can we apply f, and so the shortest word moving x_s to x_{s+1} is $fg^{|A|-1}$.

Repeatedly applying this, we have that the shortest word squashing a triple to a singleton is $f\left(fg^{|A|-1}\right)^{(|X|-(|A|+1))} fg^{|A|-1}f$ which has length

$$1 + (|X| - |A|)|A| + 1 = \left(n - 2\left\lfloor\frac{n}{4}\right\rfloor\right)\left\lfloor\frac{n}{4}\right\rfloor + 2 > \frac{n^2}{8}.$$

□

The general case extends the construction given in Theorem 2.8. We still have two transition functions and a set of states X on which both transition functions act as permutations, meaning that any synchronizable set has at most one vertex in X. Rather than having a single gadget A we will need $k-2$ gadgets A_0, A_1, A_{k-3}, each with the same structure as A but of coprime sizes.

To synchronize a k-set we will need to apply a transition function f to move a vertex around X. As before, we will not be able to apply f without first applying the other transition function g several times to move the vertex in in each A_i from $a_1^{(i)}$ to $a_0^{(i)}$. Because we chose the A_i to have coprime sizes, each such move will neccessitate many applications of g.

Theorem 2.9. *Let $k \geq 3$. For n sufficiently large,* $\mathrm{rdv}^*(k,n) \geq \frac{4}{3}\left(\frac{n}{4k}\right)^{k-1}$.

Proof. Fix the integer k. We will construct an automaton on $[n]$ where the minimal weight of a k-set is $\frac{4}{3}\left(\frac{n}{4k}\right)^{k-1}$.

Partition $[n]$ into $A_0, A_1, A_2, \ldots, A_{k-3}$ and X, where $\frac{n}{4k} \leq |A_i| \leq \frac{n}{3k}$ and $\gcd\{A_0, A_1, A_2, \ldots, A_{k-3}\} = 1$. This is possible for n sufficiently large, for example by the prime number theorem.

Label the vertices in each A_i by $a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, \ldots$ and label the vertices of X by $x_1, x_2, x_3 \ldots$. Let $q = \left\lfloor\frac{2n}{3k}\right\rfloor$.

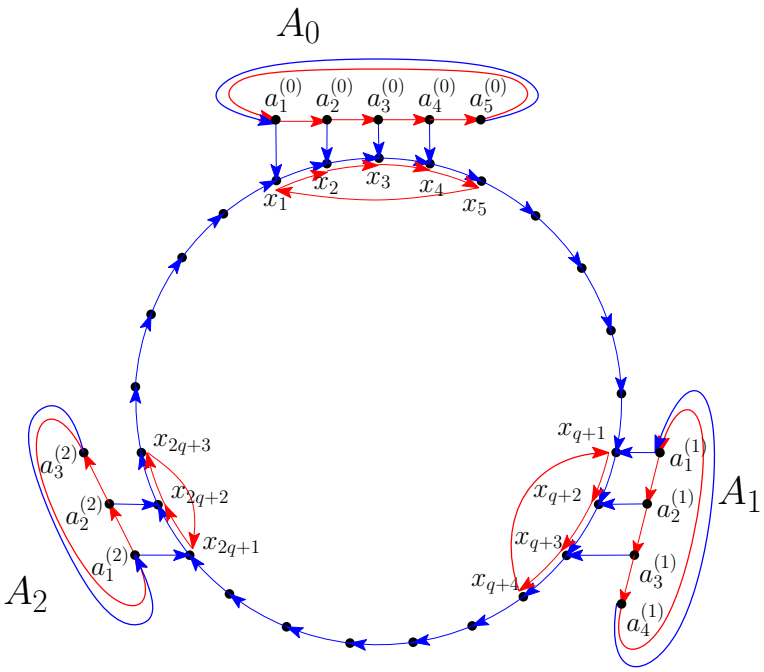


Figure 2.5: An example of the automaton for $k = 5$

Take two functions f and g as follows, as shown in figure 2.5.

$$f(x_t) = x_{(t+1 \mod |X|)}$$

$$f\left(a_j^{(i)}\right) = \begin{cases} a_1^{(i)} & \text{if } j = |A_i| \\ x_{iq+j} & \text{otherwise} \end{cases}$$

$$g(x_t) = \begin{cases} x_{t+1} & \text{if } iq+1 \le t \le iq+|A_i|-1 \text{ for some } i \\ x_{t-|A_i|+1} & \text{if } t = iq+|A_i| \text{ for some } i \\ x_t & \text{otherwise} \end{cases}$$

$$g(a_j^{(i)}) = a_{(j+1 \mod |A_i|)}^{(i)}$$

Note that f and g restricted to X are permutations on X and so any set containing more than one vertex in X cannot be synchronized.

Moreover, any set containing three vertices in some A_i cannot be synchronized: the image of such a set under g still has three vertices in A_i, and the image under f contains two vertices in X. Similarly, any set containing two vertices in A_i and two vertices in A_j for some distinct i and j also cannot be synchronized.

It follows that a synchronizable set of size k must contain two vertices in some A_i, one vertex in every other A_j and one vertex in X. Fix such a set S and consider a word that synchronizes this set acting on it.

For a shortest word from a triple to a singleton the first step must map a triple to a pair and so the first map must be f. The set S contains two points in A_i, one of which must be $a^{(i)}_{|A_i|}$ else applying f gives two points in X. Let the other be $a^{(i)}_t$, where $1 \leq t \leq |A_i| - 1$. After applying f, we have the points $a^{(i)}_1$ and x_{iq+t}, which must be the only point in X.

Note that

$$fg^{l-1}\left(a^{(i)}_1\right) = \begin{cases} a^{(i)}_1 & \text{if } l \equiv 0 \pmod{|A_i|} \\ x_{iq+(l \mod |A_i|)} & \text{otherwise} \end{cases}$$

and

$$fg^{l-1}\left(x_{iq+t}\right) = \begin{cases} x_{iq+|A_i|+1} & \text{if } t + l - 1 \equiv 0 \pmod{|A_i|} \\ x_{iq+(t+l \mod |A_i|)} & \text{otherwise} \end{cases}.$$

Since $1 \leq t \leq |A_i| - 1$ this means that applying fg^{l-1} gives two points in X for any $l \not\equiv 0 \pmod{|A_i|}$. Thus the next step must be to apply fg^{l-1} where l is some multiple of $|A_i|$. This sends $a^{(i)}_1$ and x_{iq+t} to themselves unless $t = 1$, in which case x_{iq+t} is sent to $x_{iq+|A_i|+1}$.

To further reduce the size of the set, we must map the vertex in X and some vertex in some A_j to the same point. To do this, we must move the vertex $x_{iq+|A_i|+1}$ in X round to be in $\{x_{jq}, x_{jq+1}, \ldots, x_{jq+|A_j|-2}\}$, without adding extra vertices to X as we do

so.

Suppose we have just applied f, and we now want to move x_s to x_{s+1} without adding any extra vertices into X (where s is some value not in $\{x_{jq+1}, x_{jq+2}, \ldots, x_{jq+|A_j|-1}\}$ for any j). Since we have just applied f, the vertex in each A_j must be at position $a_1^{(j)}$ (having just come from position $a_{|A_j|}^{(j)}$). We must apply f to move x_s, but we can only apply f when for each A_j, the vertex in A_j is at position $a_{|A_j|}^{(j)}$. Thus we must use g to move the vertex at $a_1^{(j)}$ to be at $a_{|A_j|}^{(j)}$ for each j.

The number of times g is applied must be congruent to -1 modulo $|A_j|$ for all j. Since $|A_0|, |A_1|, \ldots, |A_{k-3}|$ are coprime, the smallest such number is $\prod_{j=0}^{k-3} |A_j| - 1$. This is followed by an application of f and so it takes at least $\prod_{j=0}^{k-3} |A_j|$ steps to move x_s to x_{s+1}.

Applying this repeatedly, we see that the length of a word taking the vertex in X from $x_{iq+|A_i|+1}$ to some vertex of the form $\{x_{jq}, x_{jq+1}, \ldots, x_{jq+|A_j|-2}\}$ without introducing a second vertex to X must be at least

$$(q - (|A_i| - 1)) \prod_{j=0}^{k-3} |A_j| \geq \left(\frac{2n}{3k} - \frac{n}{3k}\right) \left(\frac{n}{4k}\right)^{k-2} = \frac{4}{3} \left(\frac{n}{4k}\right)^{k-1}.$$

□

We have shown that for fixed k we have $\mathrm{rdv}^*(k, n) = \Theta\left(n^{k-1}\right)$. A natural question to ask is what are the correct asymptotics for $\mathrm{rdv}^*(k, n)$? In the case $k = 3$ we have $\frac{n^2}{8} \leq \mathrm{rdv}^*(k, n) \leq \frac{n^2-n-1}{2}$.

Question 2.6. *Is there an automaton which attains* $\mathrm{rdv}^*(3, n) = (\frac{1}{2} + o(1))n^2$*?*

An upper bound on the minimum weight of a triple $\mathrm{rdv}^*(3, n)$ is the total number of synchronizable pairs plus one. To get a minimum weight triple of weight $(\frac{1}{2} + o(1))n^2$ we would need the automaton to be almost synchronizing in the sense that all but an arbitrarily small proportion of pairs are synchronizable.

Consider the construction given in the proof of Theorem 2.8. We know that a pair of vertices both in X is not synchronizable. In fact, it is straightforward to check that only pairs of the following forms are synchronizable:

- $\{a_i, x_s\}$ for $i \in \{1, 2, 3, \ldots, |A|\}$ and $s \notin \{1, 2, 3, \ldots, |A|\}$,
- $\{a_i, x_i\}$ for $i \in \{1, 2, 3, \ldots, |A|\}$,
- $\{a_1, x_{|A|}\}$ and (a_i, x_{i-1}) for $i \in \{2, 3, \ldots, |A|\}$, and
- $\{a_i, a_{(i+1 \mod |A|)}\}$ for $i \in \{1, 2, 3, \ldots, |A|\}$.

In particular, the automaton has $|A|(|X| - |A|) + 3|A| = \frac{n^2}{8} + O(n)$ synchronizable pairs. We have that number of synchronizable pairs and the minimum weight of a triple are asymptotically equal in this example. Is it possible to construct an automaton with this same property where a larger proportion of pairs are synchronizable?

2.3.1 An Alternative Construction

We found an alternative construction of an automaton with the minimal weight of a triple being approximately $2n$. This is twice the maximum that has been found for any synchronizing automaton but not as good as the value of $\frac{n^2}{8}$ obtained in Theorem 2.8. We mention it here because it is interesting that it works in a genuinely different way to the above construction. It also has much in common with the construction by Gonze and Jungers [15] of a synchronizing automaton with minimal weight in L_3 of $n + 3$.

The construction is as follows. See figure 2.6 for an example. Fix $l, m \geq 1$ where $l \nmid m$ and $l + m = n$. Label the vertices $x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m$. We define three

maps a, b and c.

$$\begin{aligned} a(y_1) &= x_1 \\ a(v) &= v \qquad \text{if } v \neq y_1 \end{aligned}$$

$$\begin{aligned} b(x_{2k}) &= x_{2k+1} \quad \text{if } 2 \leq 2k \leq l-1 \\ b(x_{2k+1}) &= x_{2k} \quad \text{if } 2 \leq 2k \leq l-1 \\ b(y_{2k}) &= y_{2k+1} \quad \text{if } 2 \leq 2k \leq m-1 \\ b(y_{2k+1}) &= y_{2k} \quad \text{if } 2 \leq 2k \leq m-1 \\ b(v) &= v \qquad \text{if } v = x_1, y_1, x_l \text{ if } l \text{ even}, y_m \text{ if m even} \end{aligned}$$

$$\begin{aligned} c(x_{2k-1}) &= x_{2k} \quad \text{if } 2k \leq l \\ c(x_{2k}) &= x_{2k-1} \quad \text{if } 2k \leq l \\ c(y_{2k-1}) &= y_{2k} \quad \text{if } 2k \leq m \\ c(y_{2k}) &= y_{2k-1} \quad \text{if } 2k \leq m \\ c(v) &= v \qquad \text{if } v = x_1, y_1, x_l \text{ if } l \text{ odd}, y_m \text{ if m odd} \end{aligned}$$

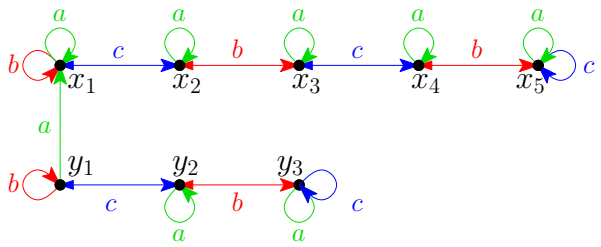


Figure 2.6: The alternative automaton for $l = 5, m = 3$

The function a synchronizes the pair $\{x_1, y_1\}$ and this is the only weight 1 pair. To find a synchroizing triple of minimal weight we consider preimages of x_1, y_1. Since b^2 and c^2 are the identity and a^{-1} acts as the identity on any pair not containing x_1, we consider

pairs of the form $\left(c^{-1}b^{-1}\right)^t(\{x_1, y_1\})$ and $b^{-1}\left(c^{-1}b^{-1}\right)^t(\{x_1, y_1\})$. We have that

$$\left(c^{-1}b^{-1}\right)^t(x_1) = \begin{cases} x_{1+2t} & \text{if } 2t \leq l-1 \\ x_{2(l-t)} & \text{if } l \leq 2t \leq 2l \end{cases}$$

$$b^{-1}\left(c^{-1}b^{-1}\right)^t(x_1) = \begin{cases} x_{2+2t} & \text{if } 2t \leq l-1 \\ x_{2(l-t)-1} & \text{if } l \leq 2t < 2l-2 \end{cases}$$

It is easy to check that $b^{-1}\left(c^{-1}b^{-1}\right)^{l-1}(y_1) \neq y_1$ so long as $m \nmid l$. We obtain that other than $\{x_1, y_1\}$, the pair containing x_1 of minimum weight is $\{x_1, b^{-1}\left(c^{-1}b^{-1}\right)^{l-1}(y_1)\}$ of weight $2l$.

Thus the minimum weight triple is $\{x_1, y_1, a^{-1}b^{-1}\left(c^{-1}b^{-1}\right)^{l-1}(y_1)\}$ of weight $2l+1$. This is maximised when m is the smallest positive integer not dividing n, in which case the minimum weight triple is $2(n-m)+1$.

Chapter 3

Semi-perfect 1-Factorizations of the Hypercube

3.1 Introduction

A *1-factorization* of a graph H is a partition of the edges of H into disjoint perfect matchings $\{M_1, M_2, \ldots, M_d\}$. The perfect matchings are also known as 1-factors, so-called because every vertex has degree 1.

Note that for a graph G to have a 1-factorization it is necessary that G is regular and has an even number of vertices. For a d-regular graph a 1-factorization can be thought of as a proper edge colouring using d colours. Figure 3.1 shows examples of 1-factorizations for the complete graph K_6 and the 3-dimensional cube Q_3.

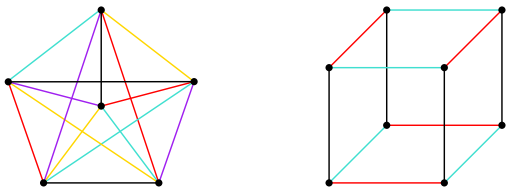


Figure 3.1: Examples of 1-factorizations of K_6 and Q_3

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_d\}$ be a 1-factorization of a graph G. The union $M_i \cup M_j$

of any two 1-factors must be a spanning 2-regular subgraph, known as a 2-factor. Note that a 2-factor must be a disjoint union of cycles.

We say that $\mathcal{M}$ is a *perfect 1-factorization* if the union $M_i \cup M_j$ of any pair of distinct 1-factors forms a Hamilton cycle in G. For example, in Figure 3.1 the 1-factorization of K_6 is perfect but the 1-factorization of Q_3 is not as the union of the red and blue 1-factors is not a Hamilton cycle.

The biggest open problem in the field of 1-factorizations is a conjecture due to Kotzig [18].

Conjecture 3.1 (Kotzig)**.** *For all n the complete graph K_{2n} on an even number of vertices has a perfect 1-factorization.*

It is easy to find a (not necessarily perfect)1-factorization of the complete graph K_{2n}. The following construction was known in the 1890s and is attributed to Walecki [21]. Arrange $2n-1$ vertices $x_0, x_1, x_2, \ldots, x_{2n-2}$ equally around a circle and put one vertex y in the centre. For $0 \leq i \leq 2n-2$ let the 1-factor M_i consist of the edge yx_i and all edges perpendicular to it.

$$M_i = yx_i \cup \bigcup_{t=1}^{n-1} x_t x_{(2i-t \mod 2n-1)}.$$

The 1-factorization of K_6 in Figure 3.1 is an example of this construction.

When $2n-1$ is prime, this 1-factorization of K_{2n} is perfect. This is straightforward to check, given the observation that by rotational symmetry it is sufficient to show that $M_0 \cup M_i$ is a Hamilton cycle for all i. In particular, we have that Kotzig's conjecture holds when $2n = p+1$ for some odd prime p.

When $2n-1$ is not prime this 1-factorization is not perfect: if $2n-1 = ab$ then it is straightforward to check that $M_0 \cup M_a$ is not a Hamilton cycle. Figure 3.2 gives a demonstration of how this fails for the 1-factorization of K_{10}.

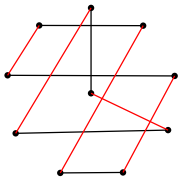


Figure 3.2: Part of a 1-factorization of K_{10} that is not perfect.

Anderson [1] and Nakamura [22] independently found a perfect 1-factorization of the complete graph K_{2p} on $2p$ vertices where p is an odd prime. Split the vertices onto two classes labelled $a_1, a_2, \ldots, a_p$ and $b_1, b_2, \ldots, b_p$. We take two kinds of 1-factor, M_i and N_j.

$$M_i = a_i b_i \cup \bigcup_{t=1}^{\frac{p-1}{2}} \left(a_t a_{(2i-t \mod p)} \cup b_t b_{(2i-t \mod p)} \right) \qquad \text{for } 1 \leq i \leq p$$

$$N_j = \bigcup_{t=1}^{p} a_t b_{(t+j \mod p)} \qquad \text{for } 1 \leq j \leq p-1$$

Figure 3.3 shows an example of the 1-factors $M_1, \ldots, M_5$ for K_{10}.

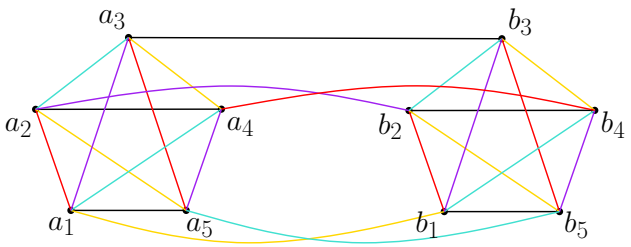


Figure 3.3: A partial example of the Anderson-Nakamura construction for K_{10}.

Using symmetry, checking that this 1-factorization is perfect can be reduced to checking that $M_0 \cup M_i$, $M_0 \cup N_j$ and $N_1 \cup N_j$ are Hamilton cycles for all i and j. This is not hard to check.

These two infinite families (i.e. $2n = p+1$ and $2n = 2p$ where p is prime) are the only infinite families of values for which Kotzig's conjecture is known to hold. The conjecture has also been checked for certain other sporadic values of $2n$ — see [23] for references and the most recently settled case of K_{56} . The smallest case for which the answer is

unknown is K_{64}.

The existence or non-existence of perfect (or 'close to perfect') 1-factorizations has been studied for various other families of regular graphs on an even number of vertices. Perhaps the most natural next families to consider are the complete bipartite graphs $K_{n,n}$ and the hypercubes Q_d for $d \geq 2$.

Let us briefly consider the complete bipartite graphs $K_{n,n}$. There is a direct relation to Kotzig's conjecture: if K_{n+1} has a perfect 1-factorization $\{M_1, M_2, \ldots, M_n\}$ then so does $K_{n,n}$. Label the vertices of K_{n+1} by $y, x_1, x_2, \ldots, x_n$ and the vertices of $K_{n,n}$ by $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$. Let M_i be a 1-factor of K_{n+1} and define M_i' a 1-factor of $K_{n,n}$ where for each edge $x_r x_s \in M$ we have the edges $a_r b_s$ and $a_s b_r$ in M_i', and the edge $y x_t \in M_i$ gives the edge $a_t b_t$ in M_i'. It is not hard to check that $\{M_1', M_2', \ldots, M_n'\}$ is a perfect 1-factorization of $K_{n,n}$. Figure 3.4 gives an example where $n = 5$.

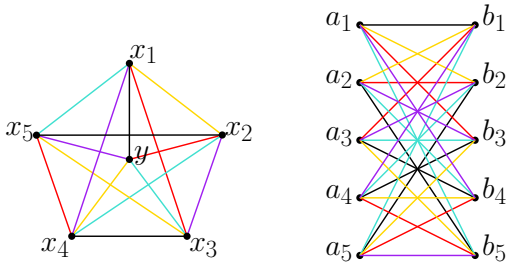


Figure 3.4: From a 1-factorization of K_6 (left) to a 1-factorization of $K_{5,5}$ (right).

This reduces the case where n is odd to Kotzig's conjecture. Clearly $K_{2,2}$ has a perfect 1-factorization. For even $n > 2$, the graph $K_{n,n}$ cannot have a perfect 1-factorization, a generalisation of which will be discussed later and which is a specific case of Theorem 3.1. In particular, a proof of Kotzig's conjecture would also completely answer the question of which complete bipartite graphs $K_{n,n}$ have a perfect 1-factorization.

The converse, that if $K_{n,n}$ has a perfect 1-factorization then so does K_{n+1}, may not be true. Wanless [34] showed that there are 37 non-isomrphic perfect 1-factorizations of $K_{9,9}$ but only one perfect 1-factorization of K_{10}, which is evidence against a direct construction. Bryant, Maenhaut and Wanless [6] found a construction of a perfect 1-

factorization of K_{p^2,p^2} where p is an odd prime. This does not come from K_{n+1} (indeed, infinitely many p^2 are not of the form q or $2q-1$ for some prime q) but rather uses Latin squares to construct 1-factorizations.

From now on we shall focus on 1-factorizations of the hypercube. For $d \geq 1$ the d-dimensional hypercube graph Q_d has vertices the subsets of $\{1, 2, \ldots, d\}$ and two vertices joined by an edge if they differ in a single element. The hypercube Q_d can also be equivalently defined with vertex set $\{0,1\}^d$ and two vertices joined by an edge if they are at Hamming distance 1 (that is, if they differ in exactly one coordinate).

We say a vertex of Q_d is even if the set contains an even number of elements, and odd if not. Note that every edge of Q_d goes from an odd vertex to an even vertex and so Q_d is bipartite with one vertex class of odd vertices and one vertex class of even vertices, each of size 2^{d-1}.

We say an edge is in direction i if its two endpoints differ in element i. This allows us to define some natural 1-factors of Q_d, called the *directional matchings*: for each direction $i = 1, \ldots, d$ let D_i be all edges in direction i. The collection of all directional matchings is a 1-factorization of Q_d, and note that the union of any pair $D_i \cup D_j$, with i, j distinct, is a disjoint union of 4-cycles. Thus the factorization into directional matchings is very far from perfect.

Perfect 1-factorizations are generally hard to find (we shall see later that they do not exist for Q_d with $d > 2$) and so we consider a weaker notion. We call a 1-factorization $\mathcal{M} = \{M_1, M_2, \ldots, M_d\}$ *semi-perfect* if there is a specified 1-factor, say M_1, such that $M_1 \cup M_i$ is a Hamilton cycle for all $i \neq 1$. We call a 1-factorization *k-semi-perfect* if there is a set of k 1-factors $M_1, \ldots, M_k$ such that $M_i \cup M_j$ is a Hamilton cycle for all $1 \leq i \leq k$ and all $k < j \leq d$. .

Craft [2] conjectured that for every integer $d \geq 2$ there is a semi-perfect 1-factorization of Q_d. This was proved independently by Gochev and Gotchev [14] and by Královič and Královič [19] in the case where d is odd, and settled for d even by Chitra and Muthusamy

[8].

Gochev and Gotchev in fact went further and defined $\mathcal{M}$ to be k-semi-perfect if $M_i \cup M_j$ forms a Hamilton cycle for every $1 \leq i \leq k$ and $k+1 \leq j \leq d$. They proved that there is a k-semi-perfect factorization of Q_d whenever k and d are both even with $k < d$.

This leads us to wonder how close to a perfect factorization we can get. Is there a k-semi-perfect factorization of Q_d for all $k < d$? Is there a perfect factorization of Q_d? If not, what is the maximal number of pairs of 1-factors whose union is a Hamilton cycle? Let us introduce some definitions.

It is convenient to introduce an auxiliary graph to express these (and similar) properties of 1-factorizations. For a 1-factorization $\mathcal{M} = \{M_1, M_2, \ldots, M_d\}$ of H, we define an auxiliary graph $G[\mathcal{M}]$ with vertices labelled $M_1, \ldots, M_d$ and an edge between M_i and M_j if $M_i \cup M_j$ is a Hamilton cycle on H. Note that the definitions above can be easily restated using $G[\mathcal{M}]$: $\mathcal{M}$ is perfect if $G[\mathcal{M}]$ is complete, $\mathcal{M}$ is semi-perfect if $G[\mathcal{M}]$ contains $K_{1,d-1}$ as a subgraph, and $\mathcal{M}$ is k semi-perfect if $G[\mathcal{M}]$ contains $K_{k,d-k}$ as a subgraph.

With this new notation, we can rephrase our questions and ask what is the maximal number of edges that $G[\mathcal{M}]$ can contain if $\mathcal{M}$ is a 1-factorization of Q_d? Which graphs can $G[\mathcal{M}]$ be isomorphic to? It is in fact not possible for $G[\mathcal{M}]$ to be complete when $d > 2$ (i.e. $\mathcal{M}$ cannot be perfect). More than this, we can show that $G[\mathcal{M}]$ must be bipartite.

Theorem 3.1 ([20])**.** *Let H be a bipartite graph on two vertex classes each of size n, where n is even. Let $\mathcal{M}$ be a partition of H into perfect matchings. Then $G[\mathcal{M}]$ must be bipartite.*

A version of Theorem 3.1 with the weaker conclusion that $G[\mathcal{M}]$ is not complete has been, according to Bryant, Maenhaut and Wanless [6] proved many times, including by

Laufer in 1980 [20]. We re-prove it here for a few reasons, the main one being that we extend the argument slightly to show that $G[\mathcal{M}]$ is bipartite. The proof also introduces ideas that we will be using later (in Theorem 3.8).

In addition, it is hard to find the theorem and its proof in the literature — in particular, when making the conjecture that there is a semi-perfect 1-factorization of Q_d, Craft also asked whether a *perfect* 1-factorization of Q_d could be found. Theorem 3.1 is not mentioned in any of the papers that proved Craft's semi-perfect conjecture.

Proof. Let X and Y be the vertex classes of H. A perfect matching M naturally induces a function $M : X \to Y$, where $(x, M(x))$ is an edge of M.

For two perfect matchings M_i and M_j, let $\pi_{j,i}$ be the permutation $M_j^{-1}M_i$ on X. Note that $\pi_{i,i} = id$, $\pi_{i,j} = \pi_{j,i}^{-1}$ and $\pi_{k,j}\pi_{j,i} = \pi_{k,i}$. Note further that if M_iM_j is an edge of $G[\mathcal{M}]$ then $M_i \cup M_j$ is a Hamilton cycle and so $\pi_{j,i}$ is a cycle of length n on X.

Suppose for a contradiction that $G[\mathcal{M}]$ contains an odd cycle and let $M_{i_1}, M_{i_2}, \ldots, M_{i_k}, M_{i_1}$ be such a cycle. The permutations $\pi_{i_2,i_1}, \pi_{i_3,i_2}, \ldots, \pi_{i_k,i_{k-1}}, \pi_{i_1,i_k}$ are all cycles of length n. Since n is even, all of these are odd permutations. Now,

$$\begin{aligned} 1 = \operatorname{sgn}(\pi_{i_1,i_1}) &= \operatorname{sgn}(\pi_{i_1,i_k}\pi_{i_k,i_{k-1}}\pi_{i_{k-1},i_{k-2}}\ldots\pi_{i_3,i_2}\pi_{i_2,i_1}) \\ &= \operatorname{sgn}(\pi_{i_1,i_k})\operatorname{sgn}(\pi_{i_k,i_{k-1}})\ldots\operatorname{sgn}(\pi_{i_3,i_2})\operatorname{sgn}(\pi_{i_2,i_1}) \\ &= (-1)^k = -1 \end{aligned}$$

We have a contradiction, hence $G[\mathcal{M}]$ contains no odd cycles. □

In the light of Theorem 3.1, the only remaining question is whether for any k, d there is a 1-factorization $\mathcal{M}$ of Q_d such that $G[\mathcal{M}]$ is isomorphic to the complete bipartite graph $K_{k,d-k}$. (Equivalently, whether there is a k-semi-perfect 1-factorization of Q_d for every k and d, in the language of Gochev and Gotchev.) We almost fully resolve this problem, with the one exception being for whether $G[\mathcal{M}]$ can be isomorphic to $K_{3,3}$.

Theorem 3.2. *For $k, l \in \mathbb{N}$ not both equal to 3, there is a 1-factorization $\mathcal{M}$ of the hypercube Q_{k+l} such that $G[\mathcal{M}]$ is isomorphic to the complete bipartite graph $K_{k,l}$.*

We also explain, in section 3.3, why the $K_{3,3}$ case cannot be resolved with our methods. In particular, the 1-factorizations we construct in the proof of the main theorem have a direction respecting property. We show that any 1-factorization $\mathcal{M}$ of Q_6 satisfying this direction respecting property cannot have $G[\mathcal{M}]$ is isomorphic to $K_{3,3}$.

We finish with some open questions.

3.2 Main Theorem

To prove the theorem, we will use the following result due to Stong, which concerns the symmetric directed hypercube $\overleftrightarrow{Q_d}$, obtained from Q_d by replacing each edge with two directed edges, one in each direction.

Theorem 3.3 ([28])**.** *For $d \neq 3$, the symmetric directed hypercube $\overleftrightarrow{Q_d}$ can be partitioned into d directed Hamilton cycles.*

The result of Stong's Theorem is false when $d = 3$, which is easy but unenlightening to check. As a result, when proving Theorem 3.2 we will have to deal separately with the case when one of k or l is equal to 3. This is also the reason why we have been unable to resolve the $k = l = 3$ case.

Stong's result applies to directed cubes, but the following corollary allows us to use it for undirected cubes.

Corollary 3.4. *For $d \neq 3$, the cube Q_d can be partitioned into 1-factors $A_1, A_2, \ldots, A_d$ and also partitioned into 1-factors $B_1, B_2, \ldots, B_d$ such that $A_i \cup B_i$ is a Hamilton cycle for all $i = 1, 2, \ldots, d$.*

Proof. Using Theorem 3.3, partition $\overleftrightarrow{Q_d}$ into directed Hamilton cycles H_1, H_2, …, H_d. Let E be the even vertices of $\overleftrightarrow{Q_d}$ and O the odd vertices, so that $\overleftrightarrow{Q_d}$ is bipartite with

respect to the vertex classes E and O. For each H_i, we define A_i to be the edges of H_i that go from E to O, and B_i to be the edges that go from O to E.

Since $H_1, H_2, \ldots, H_d$ partition $\overleftrightarrow{Q_d}$, every edge from E to O is in a unique A_i and every edge from O to E is in a unique B_j. If we now ignore the directions on the edges, every edge of Q_d is in a unique A_i and a unique B_j. It is clear that A_i and B_i are perfect matchings and $A_i \cup B_i$ is a Hamilton cycle by construction. □

Note that we have slightly abused notation in the case $d = 1$, since $A_1 = B_1 = Q_1$ and so $A_1 \cup B_1$ is a single edge rather than a cycle. This will not matter in the cases $k \neq 3, l = 1$, and we will consider the case $k = 3, l = 1$ separately.

Corollary 3.4 together with a theorem of Gochev and Gotchev [14, Theorem 3.1] is enough to show that it is possible to have $G[\mathcal{M}]$ isomorphic to $K_{k,n-k}$ for all $k \neq 3$ and all even $n - k$. We will improve on their arguments to deal with all but one of the remaining cases.

We will split the theorem for three different cases and prove each separately. Before we do so, let us outline the ideas involved.

We can view the hypercube Q_{k+l} as a k-dimensional hypercube whose 'vertices' are copies of Q_l (i.e. as the Cartesian product of Q_k and Q_l). It may be helpful to recall the alternative definition of Q_{k+l} as having vertex set $\{0,1\}^{k+l}$ and two vertices joined by an edge if they are at Hamming distance 1. Using this definition, we think of Q_k as 'the first k coordinates' and Q_l as 'the last l coordinates'.

Let us formalise this idea: Label the vertices of Q_k as subsets of $\{1, 2, \ldots, k\}$ in the usual way. For each vertex u of Q_k, we define a different copy of Q_l within Q_{l+k}: let Q_l^u be the induced subgraph of Q_{k+l} on all vertices w where $w \cap \{1, 2, \ldots, k\} = u$.

Conversely, we can view Q_{k+l} as a l-dimensional hypercube whose 'vertices' are copies of Q_k. This time, label the vertices of Q_l as subsets of $\{k+1, k+2, \ldots, k+l\}$ in the

natural way. For each vertex v of Q_l, we define a different copy of Q_k within Q_{l+k}: let Q_k^v be the induced subgraph of Q_{k+l} on all vertices x with $x \cap \{k+1, k+2, \ldots k+l\} = v$.

The most straightforward case of the theorem is when neither k nor l is equal to 3, proved in Proposition 3.5. To prove this we use a generalisation of Gochev and Gotchev's construction [14].

The idea of the proof is as follows: first, we construct k disjoint matchings that use only edges in directions $1, \ldots, k$. The matchings used within the Q_k^vs are those obtained from applying Corollary 3.4 to Q_k. Next we construct l disjoint matchings that use only edges in directions $k+1, \ldots, k+l$. Similarly, the matchings used within the Q_l^us are those obtained from applying Corollary 3.4 to Q_l. We then prove that taking the union of a matching of the first kind and a matching of the second kind gives a Hamilton cycle.

The second case of the theorem is when $k = 3$ and l is not equal to 1 or 3, proved in Proposition 3.6. We use a similar construction to the first case, the only difference being that while we can use Corollary 3.4 on Q_l, we cannot apply it to Q_3. We will instead take directional matchings on the copies of Q_3; it turns out this can be made to work here.

Finally, we are left with two cases: $(k, l) = (3, 1)$ and $(k, l) = (3, 3)$. The first of these is proved in Proposition 3.7 by means of an explicit example. The case $(k, l) = (3, 3)$ is left unsolved. The difficulty of these final two cases is discussed in Section 3.3.

The following useful notation is common to the proofs of propositions 3.5 and 3.6. For a perfect matching M and a vertex v, we define $M(v)$ to be the other endpoint of the edge containing v in M. (Note that this clashes slightly with our notation in Theorem 3.1: by that notation we are here conflating M and M^{-1}.)

Proposition 3.5. *When neither k nor l is equal to 3, there is a 1-factorization $\mathcal{M}$ of the hypercube Q_{k+l} such that $G[\mathcal{M}]$ is isomorphic to the complete bipartite graph $K_{k,l}$.*

Proof. Using Corollary 3.4 partition Q_k into matchings $A_1, A_2, \ldots, A_k$ and matchings $B_1, B_2, \ldots, B_k$ such that $A_i \cup B_i$ is a Hamilton cycle for all i.

For $i = 1, 2, \ldots, k$ define M_i to be the matching on Q_{k+l} defined by taking the following edges:

$$\begin{cases} A_i & \text{on } Q_k^{\emptyset} \\ B_i & \text{on } Q_k^v \text{ for } v \neq \emptyset \end{cases}$$

Note that the M_i are all disjoint, and they only use edges in directions $1, 2, \ldots, k$.

Also partition Q_l into matchings X_1, $X_2, \ldots$, X_l and matchings Y_1, $Y_2, \ldots$, Y_l such that $X_j \cup Y_j$ is a Hamilton cycle for all j.

For $j = 1, 2, \ldots, l$ define N_j to be the matching on Q_{k+l} defined by taking the following edges:

$$\begin{cases} X_j & \text{on } Q_l^u \text{ for } u \text{ even} \\ Y_j & \text{on } Q_l^u \text{ for } u \text{ odd} \end{cases}$$

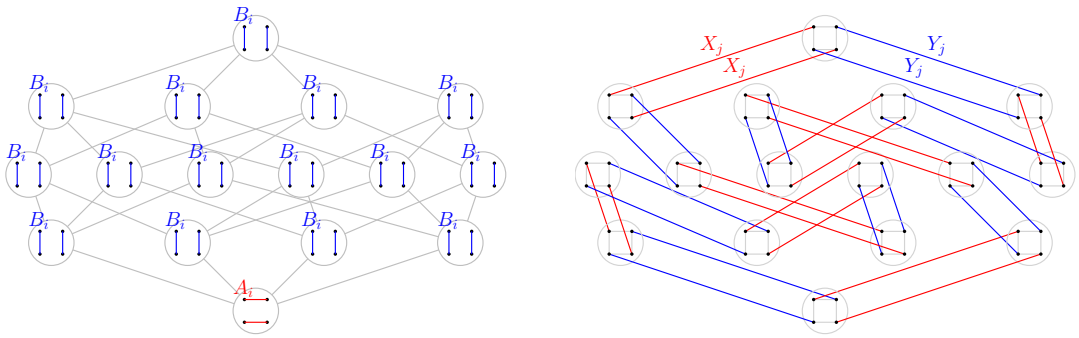
Another way to think of N_j is as containing edges between copies of Q_k^v. From an even vertex in Q_k^v we add an edge to the corresponding vertex in $Q_k^{X_j(v)}$, and from an odd vertex in Q_k^v we add an edge to the corresponding vertex in $Q_k^{Y_j(v)}$.

Note that the N_j are all disjoint, and they only use edges in directions $k+1, k+2, \ldots, k+l$. Thus the matchings $\{M_i\}_{i=1}^k \cup \{N_j\}_{j=1}^l$ are all disjoint and form a 1-factorization of Q_{k+l} (see Figure 3.5).

Note that since the auxiliary graph $G[\mathcal{M}]$ must be bipartite, to show that $G[\mathcal{M}]$ is isomorphic to $K_{k,l}$ to finish the proof it is sufficient to show that $M_i \cup N_j$ is a Hamilton cycle for all i, j.

Consider following the cycle starting at a vertex u that lies in $Q_k^{\emptyset}$ and alternating between edges first in N_j and then in M_i.

Every time we travel along an edge in M_i the parity of the vertex in Q_k^v switches,



(a) M_i (b) N_j

Figure 3.5: An example when $k = 2$ and $l = 4$

and so we will alternate using edges from X_j and edges from Y_j in N_j. As $X_j \cup Y_j$ is a Hamilton cycle, the first time the cycle returns to $Q_k^\emptyset$ we will have travelled through each other Q_k^v exactly once.

Each time we travel through a different Q_k^v we use an edge from B_i within it. After passing through $2^l - 1$ copies of Q_k^v we will have bounced between u and $B_i(u)$ an odd number of times, so the first vertex we encounter in our return to $Q_k^\emptyset$ is $B_i(u)$. The next vertex would then be $A_i(B_i(u))$.

After passing through $2(2^l)$ distinct vertices (two in each Q_k^v) we have moved from u to $A_i(B_i(u))$, i.e. made two steps of the Hamilton cycle $A_i \cup B_i$ within $Q_k^\emptyset$. Thus the first time we will return to $u \cup \emptyset$ is after passing through $2^k 2^l$ vertices, which is the total number of vertices in the graph. Hence we have a Hamilton cycle. □

Proposition 3.6. *For l not equal to 1 or 3, there is a 1-factorization $\mathcal{M}$ of the hypercube Q_{3+l} such that $G[\mathcal{M}]$ is isomorphic to the complete bipartite graph $K_{3,l}$.*

Proof. Using Corollary 3.4, partition Q_l into matchings $A_1, A_2, \ldots, A_l$ and $B_1, B_2, \ldots, B_l$ such that $A_i \cup B_i$ is a Hamilton cycle for all j. Let X_1, X_2 and X_3 be the three directional matchings of Q_3 — that is, X_j contains all edges in direction j.

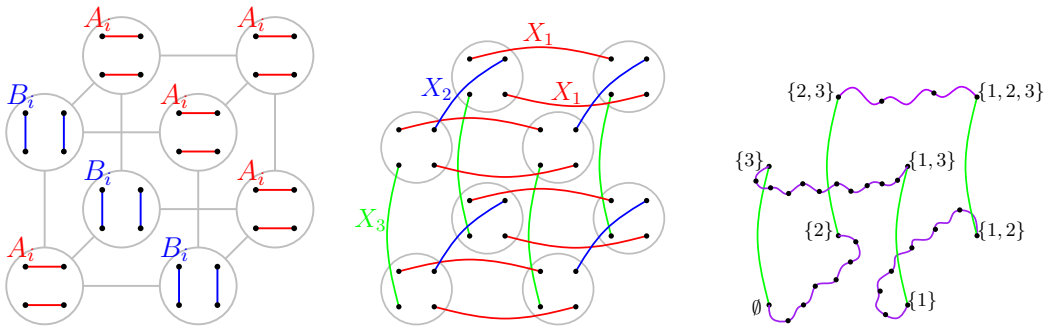
For $i = 1, 2, \ldots, l$ define M_i to be the matching on Q_{3+l} defined by taking the following

edges:

$$\begin{cases} A_i & \text{on } Q_l^{\emptyset}, Q_l^{\{1,2\}}, Q_l^{\{1,3\}}, Q_l^{\{2,3\}} \text{ and } Q_l^{\{1,2,3\}} \\ B_i & \text{on } Q_l^{\{1\}}, Q_l^{\{2\}} \text{ and } Q_l^{\{3\}} \end{cases}$$

For $j = 1, 2, 3$ define N_j to be the matching on Q_{3+l} defined by taking the following edges, where the subscripts for the Xs are taken modulo 3:

$$\begin{cases} X_j & \text{on } Q_3^v \text{ for } v \text{ odd} \\ X_{j+1} & \text{on } Q_3^v \text{ for } v \text{ even and } v \neq \emptyset \\ X_{j+2} & \text{on } Q_3^{\emptyset} \end{cases}$$



(a) M_i (b) N_1 (c) Sketch of cycle $M_i \cup N_1$

Figure 3.6: An example when $l = 2$

Now $\{M_i\}_{i=1}^{l} \cup \{N_j\}_{j=1}^{3}$ is a set of $3 + l$ disjoint perfect matchings. As $G[\mathcal{M}]$ must be bipartite, to complete the proof it only remains to show that $M_i \cup N_j$ is a Hamilton cycle for any i and j.

Note that $\{M_i\}$ is invariant under the permutation that cycles directions 1,2 and 3. Since N_2 and N_3 are obtained from N_1 by such cyclic permutations, we can without loss of generality assume that $j = 1$.

Consider $M_i \cup N_1$ with the edges in $Q_3^{\emptyset}$ removed; that is, the edges $\emptyset\{3\}$, $\{1\}\{1,3\}$, $\{2\}\{2,3\}$ and $\{1,2\}\{1,2,3\}$. We will show that the resulting graph comprises four paths, from $\emptyset$ to $\{2\}$, from $\{2,3\}$ to $\{1,2,3\}$, from $\{1,2\}$ to $\{1\}$ and from $\{1,3\}$ to $\{3\}$. Thus

when we add back the four edges in direction 3, we get a Hamilton cycle. See figure 3.6c for an example.

View Q_{3+l} as an l-dimensional hypercube whose 'vertices' are copies of Q_3. Starting at a vertex in $Q_3^{\emptyset}$ and following the path from it, we will not return to $Q_3^{\emptyset}$ until we have made 2^l steps around $A_i \cup B_i$.

A path starting at $\emptyset$ will move in directions according to A_i then X_1 then B_i then X_2 and then repeat this pattern. It will return to $Q_3^{\emptyset}$ after 2^l moves from $A_i \cup B_i$ and $2^l - 1$ moves from $X_1 \cup X_2$. Since $l \geq 2$, this means we end at the vertex $\{2\}$, and the path contains $2(2^l)$ vertices.

The same argument works to show that there is a path from $\{1, 2\}$ to $\{1\}$ containing $2(2^l)$ vertices.

A path starting at $\{2, 3\}$ will move in directions according to A_i then X_1 then A_i, ending at the vertex $\{1, 2, 3\}$ and containing 4 vertices.

A path starting at $\{1, 3\}$ will move in directions according to A_i, X_1, B_i, X_2, A_i, X_1, A_i, X_2, and then repeat this pattern. It will return to $Q_3^{\emptyset}$ after $2(2^l) - 2$ moves from $A_i \cup B_i$ and $2(2^l) - 3$ moves from $X_1 \cup X_2$. Thus we end at the vertex $\{3\}$, and the path contains $4(2^l) - 4$ vertices.

The sum of the lengths of these paths is $8(2^l)$, and so every vertex is contained in one of these paths. □

Proposition 3.7. *There is a 1-factorization $\mathcal{M}$ of the hypercube Q_4 such that $G[\mathcal{M}]$ is isomorphic to the complete bipartite graph $K_{3,1}$.*

Proof. The four matchings are shown in figure 3.7. It is easy to check that the top matching forms a Hamilton cycle with any of the three bottom matchings (in fact, by symmetry you need only check one pair). □

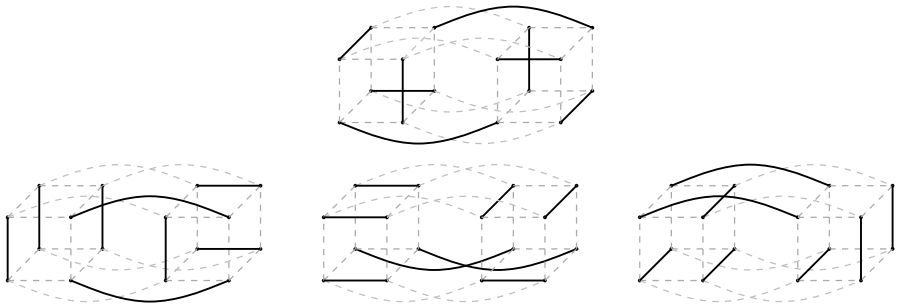


Figure 3.7: The matchings for $k = 1$ and $l = 3$

Proof of Theorem 3.2. Combine the results of Propositions 3.5, 3.6 and 3.7. □

3.3 Direction Respecting 1-Factorizations

The only case not covered by Theorem 3.2 is whether $G[\mathcal{M}]$ can be isomorphic to $K_{3,3}$. This case cannot be resolved with our methods alone. To explain why, we will introduce a notion of direction respecting 1-factorizations.

Fix k and l and let $\mathcal{M} = M_1, M_2, \ldots M_{k+l}$ be a 1-factorization of Q_{k+l}. We call the 1-factorization $\mathcal{M}$ *direction respecting* if $M_1, M_2, \ldots, M_k$ only use edges in directions $1, \ldots, k$ and $M_{k+1}, M_{k+2}, \ldots M_{k+l}$ only use edges in directions $k+1, \ldots k+l$.

Note that the matchings constructed in Propositions 3.5 and 3.6 were direction respecting for the appropriate k and l. However, the 1-factorization given in the proof of proposition 3.7 was not direction respecting. We shall prove that there is no direction respecting 1-factorization $\mathcal{M}$ with $G[\mathcal{M}]$ isomorphic to $K_{3,3}$ or $K_{3,1}$.

Theorem 3.8. *There is a direction respecting 1-factorization $\mathcal{M}$ of Q_{k+l} with $G[\mathcal{M}] = K_{k,l}$ if and only if (k, l) are not $(3, 1)$, $(1, 3)$ or $(3, 3)$.*

Proof. First note that the proof of Theorem 3.2 shows that such a direction respecting 1-factorization exists when (k, l) are not $(3, 1)$, $(1, 3)$ or $(3, 3)$.

Let $d = 3 + l$ where l is 3 or 1. For M a perfect matching on Q_d, think of M as a bijection from the odd vertices of Q_d to the even vertices (as in Theorem 3.1). If M and

N are perfect matchings then MN^{-1} is a permutation on the even vertices of Q_d. We define the *sign* of a 1-factorization $\{M_i\}_{i=1}^d$ of Q_d to be the product of the signs of the permutations $M_iM_j^{-1}$ for all $i < j$. That is,

$$\text{sgn}(\mathcal{M}) = \prod_{i<j} \text{sgn}\left(M_iM_j^{-1}\right).$$

Let $\mathcal{D}^{(d)} = \{D_i^{(d)}\}_{i=1}^d$ be the directional matchings of Q_d, where $D_i^{(d)}$ contains all edges in direction i. For $i \neq j$ the permutation $D_i^{(d)}\left(D_j^{(d)}\right)^{-1}$ consists of 2^{d-2} disjoint 4-cycles. Thus $\text{sgn}\left(D_i^{(d)}\left(D_j^{(d)}\right)^{-1}\right) = (-1)^{2^{d-2}} = 1$ for all i, j, and so $\text{sgn}(\mathcal{D}^{(d)}) = 1$.

Suppose $\mathcal{M} = \{M_i\}_{i=1}^3 \cup \{N_j\}_{j=1}^l$ is a 1-factorization of Q_d where $M_i \cup N_j$ is a Hamilton cycle for all i, j. The permutation $M_iN_j^{-1}$ is a cycle of length 2^{d-1} and so has sign -1. Note that $\text{sgn}\left(M_iM_s^{-1}\right) = \text{sgn}\left(M_iN_j^{-1}\right)\left((M_sN_j^{-1})^{-1}\right) = (-1)(-1) = 1$, and similarly $\text{sgn}(N_jN_t^{-1}) = 1$. Thus $\text{sgn}(\mathcal{M}) = (-1)^{3l} = -1$ for any $\mathcal{M}$ with $G[\mathcal{M}] = K_{3,l}$.

We will define a switching operation on 1-factorizations that preserves their sign. We will further show that any direction respecting 1-factorization $\mathcal{M}$ can be obtained from $\mathcal{D}^{(d)}$ using a series of switches. Since the sign of $\mathcal{D}^{(d)}$ is 1, this is enough to show that $G[\mathcal{M}] \neq K_{3,l}$.

Let $\mathcal{M} = \{M_i\}_{i=1}^d$ be a 1-factorization of Q_d. Take a 4-cycle x, y, v, w in Q_d and suppose that the edges xy and vw are in matching M_s and vy and xw are in matching M_t. A switch on w, v, y, w replaces $\mathcal{M}$ by the 1-factorization $\mathcal{M}' = \{M_i'\}_{i=1}^n$ where $M_s' = M_s \cup \{vy, xw\} \setminus \{xy, vw\}$, $M_t' = M_t \cup \{xy, vw\} \setminus \{vy, xw\}$, and $M_i' = M_i$ for $i \neq s, t$.

Viewing the 1-factors M_i as bijections from the even vertices to the odd vertices, we have composed M_s and M_t with the function swapping x and v, where x and v are the even vertices of x, y, v, w. Therefore the permutations $M_sM_i^{-1}$ and $M_s'M_i^{-1}$, where $i \neq s, t$, differ from each other by the transposition (x, v) and so have opposite sign. Similarly $M_tM_i^{-1}$ and $M_t'M_i^{-1}$ have opposite sign.

From this second interpretation of the switch it is clear that:

$$\begin{aligned}\operatorname{sgn}(\mathcal{M}') &= \prod_{i<j} \operatorname{sgn}\left(M_i'(M_j')^{-1}\right) \\ &= \prod_{\substack{\text{exactly one of} \\ i,j \text{ is } s \text{ or } t}} -\operatorname{sgn}\left(M_i M_j^{-1}\right) \prod_{\substack{\text{neither or both of} \\ i,j \text{ is } s \text{ or } t}} \operatorname{sgn}\left(M_i M_j^{-1}\right) \\ &= (-1)^{2(d-2)} \operatorname{sgn}(\mathcal{M}) = \operatorname{sgn}(\mathcal{M})\end{aligned}$$

All that is left to show is that a 1-factorization satisfying the conditions of the theorem can be obtained from $\mathcal{D}^{(d)}$ by a series of switches. We will use the following proposition.

Proposition 3.9. *Let $D_1^{(3)}, D_2^{(3)}, D_3^{(3)}$ be the directional matchings on Q_3 and let A_1, A_2, A_3 be another 1-factorization of Q_3. Then there are a series of switches that transform $D_1^{(3)}$, $D_2^{(3)}$, $D_3^{(3)}$ into A_1, A_2, A_3, respecting the ordering.*

Proof of Proposition. It is easy to check that there are only 4 ways to partition Q_3 into perfect matchings, up to ordering — one way uses three directional matchings and the other three ways each use one directional matching. Without loss of generality say that A_1 is a directional matching.

Note that we can use switches to re-order $D_1^{(3)}, D_2^{(3)}, D_3^{(3)}$. To swap $D_i^{(3)}$ and $D_j^{(3)}$ switch on $\emptyset, \{i\}, \{j\}, \{i,j\}$ and on $\{k\}, \{i,k\}, \{j,k\}, \{i,j,k\}$, where i,j,k is $1,2,3$ in some order. Thus we can assume without loss of generality that $A_1 = D_1^{(3)}$.

If A_2 and A_3 are also directional matchings then we are done. If not, then we can switch on $\emptyset, \{2\}, \{2,3\}, \{3\}$ to make them both directional matchings. □

Let $\mathcal{M} = \{M_i\}_{i=1}^3 \cup \{N_j\}_{j=1}^l$ be a 1-factorization of Q_d satisfying the conditions of the theorem.

As in Theorem 3.2, we can view Q_{3+l} as an l dimensional hypercube whose 'vertices' are copies of Q_3. For $v \subset \{3+1, \ldots, 3+l\}$ let Q_3^v be the induced subgraph of Q_{3+l} on

vertices of the form $u \cup v$ for all $u \subset \{1, 2, 3\}$. For each v in turn, apply the claim to Q_3^v and M_1, M_2, M_3 restricted to Q_3^v. In this way we obtain a series of switches that turn $D_1^{(d)}, D_2^{(d)}, D_3^{(d)}$ into M_1, M_2, M_3.

If $l = 1$, $N_1 = D_4^{(n)}$ and we are done. If $l = 3$, apply an analogous process to above to find switches that turn $D_4^{(d)}, D_5^{(d)}, D_6^{(d)}$ into N_1, N_2, N_3. Note that these switches will be only on edges in directions 4,5,6 and so will not interfere with M_1, M_2, M_3 in any way. □

3.4 Computer Experiments

Given that there is only one case missing from Theorem 3.2, we wrote a computer program to find k-semi-perfect 1-factorizations of the hypercube. Unfortunately, the running time grows very large with the dimension and so we were unable to find a 3-semi-perfect 1-factorization of Q_6. We did, however, find various 1-factorizations for cubes in fewer than 6 dimensions.

For $d \leq 4$, we can exhaustively find all 1-factorizations of Q_d for $d \leq 4$. In particular, we can find all k-semi-perfect 1-factorizations of Q_4 for $k < 4$.

We know that there is a 1-semi-perfect 1-factorization of Q_4 that is not direction respecting, as Figure 3.7 shows. One notable discovery of our computer experiments is that there also exist 2-semi-perfect 1-factorizations of Q_4 that are not direction respecting. An example is given in Table 3-A, where the rows represent the vertices of Q_4 (written in binary), the columns represent the four 1-factors, and the entries of the table give the direction of the edge adjacent to that vertex in the corresponding 1-factor.

It is clear to see that this is not a direction respecting 1-factorization as the 1-factors M_2 and M_3 both have at least one edge in every direction. We can check that $M_i \cup M_j$ is a Hamilton cycle for $0 \leq i \leq 1$ and $2 \leq j \leq 3$ (so this is 2-semi-perfect), and for this particular example we have in addition that $M_i \cup M_j$ is only two disjoint cycles for any other pair of distinct M_i, M_j.

	Edge Directions			
Vertices	M_0	M_1	M_2	M_3
0000	1	2	3	4
0001	1	3	4	2
0010	4	2	1	3
0011	4	3	1	2
0100	4	2	3	1
0101	4	3	2	1
0110	1	2	4	3
0111	1	3	2	4
1000	1	3	2	4
1001	1	2	4	3
1010	4	3	2	1
1011	4	2	3	1
1100	4	3	1	2
1101	4	2	1	3
1110	1	3	4	2
1111	1	2	3	4

Table 3-A: A 2-semi-perfect 1-factorization of Q_4 that is not direction respecting.

When the dimension is 5 the running time becomes too long to perform an exhaustive search. We were able to use the program to find an example of a 1-factorization of Q_5 where the union $M_i \cup M_j$ of any pair of distinct 1-factors forms exactly two disjoint cycles. This is shown in Table 3-B.

How does the program work? First, we store the hypercube Q_d as a list of vertices from 0 to $2^d - 1$, a list of edges as ordered pairs (i, j) with $i < j$, and a dictionary indexed by vertices where the entry for a vertex is a list of all adjacent edges.

To find all 1-factorizations, start with $i = 0$ and then apply the following algorithm:

1. Set P to be a list of all edges that are not in any M_j for $0 \leq j < i$. Think of P as the potential edges.

2. Put $(0, 2^i)$ into the 1-factor M_i and remove all edges adjacent to $(0, 2^i)$ from P.

3. If P is not empty, put the first edge e of P into M_i and remove all edges adjacent to e from P.

	Edge Directions				
Vertices	M_0	M_1	M_2	M_3	M_4
00000	1	2	3	4	5
00001	1	2	4	5	3
00010	1	2	4	3	5
00011	1	2	3	5	4
00100	1	2	3	5	4
00101	1	4	5	2	3
00110	1	2	4	3	5
00111	1	5	3	2	4
01000	1	2	3	4	5
01001	1	5	4	2	3
01010	1	2	4	5	3
01011	1	3	5	2	4
01100	1	5	3	2	4
01101	1	4	2	5	3
01110	1	5	4	2	3
01111	1	3	2	5	4
10000	1	4	3	2	5
10001	1	3	2	5	4
10010	1	4	3	2	5
10011	1	4	2	5	3
10100	1	2	3	5	4
10101	1	3	5	2	4
10110	1	2	3	4	5
10111	1	5	4	2	3
11000	2	4	1	3	5
11001	3	5	1	2	4
11010	2	4	3	5	1
11011	3	4	5	2	1
11100	2	5	1	3	4
11101	3	2	1	5	4
11110	2	5	3	4	1
11111	3	2	4	5	1

Table 3-B: A 1-factorization of Q_5 where the union of any pair of 1-factors is two cycles.

4. Repeat step 3 until P is empty.

   - If M_i is a 1-factor $^{(1)}$ and $i \neq d-1$, increase i by one and go back to step 1.
   - If M_i is a 1-factor $^{(1)}$ and $i = d-1$, then we have a 1-factorization. Store it $^{(2)}$. Then, reverse to the last point where P contained more than one edge at

the end of step 3. Remove the first edge from P and apply step 3 (so we will now be taking what was the next edge in P).

- Otherwise M_i is not a 1-factor. Reverse to the last point where P contained more than one edge at the end of step 3. Remove the first edge from P and apply step 3.

5. Continue until all possibilities have been exhausted.

Note that by ensuring that the edge $(0, 2^i)$ is in M_i, we have ruled out counting as different 1-factorizations where the 1-factors are simply reordered.

This algorithm can be adapted to find 1-factorizations with certain properties by adding in a check that M_i has the desired property at the places indicated by (1). For example, to find all k-semi-perfect 1-factorizations we will check that if $i \geq k$ then $M_i \cup M_j$ is a Hamilton cycle for all $0 \leq j < k$. To find all 1-factorizations where the union of any pair of 1-factors is at most two cycles, we add a check that $M_i \cup M_j$ is at most two cycles for all $j < i$. Checking as we go along rather than at the end saves exploring fruitless avenues.

The algorithm can also be adapted is to find a single example of a 1-factorization with a property rather than exhaustively searching for all of them. This can be done by simply stopping the algorithm at the place indicated by (2). This is particularly useful when dealing with dimensions greater than 4 where exhaustive searches are impractical.

The actual computer code that was written to find these examples can be found in Appendix A.

3.5 Open Questions

The most obvious question is the missing case from Theorem 3.2.

Question 3.2. *Is it possible to find a 1-factorization $\mathcal{M}$ of Q_6 such that $G[\mathcal{M}] = K_{3,3}$?*

Theorem 3.8 and its proof show that any such matching $\mathcal{M}$ cannot be obtained from applying a series of switches to the directional matchings. However, the answer could still be 'yes'. We found by hand a 1-semi-perfect factorization of Q_4 that wasn't obtained from applying a series of switches to the directional matchings. Computer checking shows that there in addition are other 1- and 2-semi-perfect 1-factorizations of Q_4 that are not direction respecting, of which the 1-factorisation in Table 3-A is an example.

We know from Theorem 3.1 that we cannot have a perfect 1-factorization of Q_d for $d > 2$. In fact, the maximum possible number of pairs of 1-factors whose union forms a Hamilton cycle is $\left\lfloor \frac{d^2}{4} \right\rfloor$, obtained when $G[\mathcal{M}] = K_{\lfloor d/2 \rfloor, \lceil d/2 \rceil}$. What can be said about the other pairs — can their union be close to a Hamilton cycle in some way?

Question 3.3. *Let $\mathcal{M} = \{M_i\}_{i=1}^{d}$ be a 1-factorization of Q_d. Is it possible for $M_i \cup M_j$ to contain a cycle of length $(1 - o(1))2^d$ for every $i \neq j$?*

Question 3.4. *Let $\mathcal{M} = \{M_i\}_{i=1}^{d}$ be a 1-factorization of Q_d. Is it possible for $M_i \cup M_j$ to consist of at most 2 cycles for every $i \neq j$?*

In any direction-respecting k-semi-perfect 1-factorization, we have that $M_i \cup M_j$ is contained solely within the copies of Q_k and so contains cycles of length at most 2^k and at least 2^l components (and similarly for $N_s \cup N_t$ with k and l swapped). Assuming without loss of generality that $k \leq l$, the 2-factor $M_i \cup M_j$ pretty far from a Hamilton cycle. In particular, a positive answer to Questions 3.3 and 3.4 would have to use a 1-factorization that is far from direction respecting.

The examples in Figure 3.7 or Table 3-A (for Q_4) and Table 3-B (for Q_5) show that for $d \leq 5$ the answer to question 3.4 is 'yes'.

One could also pose more general versions of these questions.

Question 3.5. *What is the largest $c = c(d)$ for which there exists a 1-factorization of Q_d such that for every $i \neq j$, $M_i \cup M_j$ contains a cycle of length at least c?*

Question 3.6. *What is the smallest $t = t(d)$ for which there exists a 1-factorization of Q_d such that for every $i \neq j$, $M_i \cup M_j$ consists of at most t components?*

We suspect that $c(d)$ grows exponentially in d, but are only able to prove that $c(d)$ is non-decreasing with d. Let $\mathcal{M} = \{M_1, \ldots, M_d\}$ be a 1-factorization of Q_d where every pair $M_i \cup M_j$ contains a cycle of length at least c. Write $Q_{d+1} = Q_d \times \{0, 1\}$. For $1 \leq i \leq d$ let the 1-factor M'_i be the same as M_i on $Q_d \times \{0\}$ and the same as $M_{(i+1 \mod d)}$ on $Q_d \times \{1\}$. Let M'_{d+1} be all edges in the $(d+1)$th direction. It is not hard to see that $\{M'_1, \ldots, M'_{d+1}\}$ is a 1-factorization of Q_{d+1} where any pair $M'_i \cup M'_j$ contains a cycle of length at least c.

The same construction shows that $t(d+1) \leq 2t(d)$. Again, we suspect that this is far from optimal and that, as per Question 3.4, it could be that $t(d)$ is bounded by a constant.

A different way of thinking of Hamilton cycles is as connected 2-factors. Thus a different generalisation of the problem would be to ask about the connectivity of other r-factors. For example,

Question 3.7. *For each d, let $r = r(d)$ be minimal subject to there existing a 1-factorization $\mathcal{M}$ of Q_d where the union of any r distinct 1-factors is connected. What is the value of $r(d)$?*

Theorem 3.1 shows that $r(d)$ is greater than 2 for $d > 2$. The 1-factorization given by Theorem 3.2 in the case $k = \lfloor \frac{d}{2} \rfloor$ and $l = \lceil \frac{d}{2} \rceil$ has the property that the union of any $\left(\lceil \frac{d}{2} \rceil + 1\right)$ 1-factors is connected, hence $r(d) \leq \lceil \frac{d}{2} \rceil + 1$ for $d \neq 6$. It seems possible that r is constant and it could be even as small as 3.

Chapter 4

Connectivity of High Dimension k-Nearest-Neighbour Graphs

4.1 Introduction

Suppose that you had n radios lying in a plane and each radio can communicate with its k nearest neighbours. How large does k need to be, in relation to n and d, for every radio to be able to communicate (maybe indirectly) with any other?

Let us formalise this notion. First, for each n, let S_n be the square of area n. Define an undirected random geometric graph $G = G(n, k)$ as follows: Let $\mathcal{P}$ be a Poisson process of density 1 in S_n. Join every point of $\mathcal{P}$ with its k nearest neighbours (in the Euclidean metric) by an undirected edge.

Note that we could have instead chosen a square of side-length 1 and a Poisson process of density n, which would simply be a rescaling. Note also that the expected number of points in the square is n, and if we condition on the number of points being exactly n we have a uniform distribution.

Throughout this document we say that G has a certain property *with high probability*

if the probability G has the property tends to 1 as n tends to infinity.

This random geometric graph has been well-studied. Xue and Kumar [35] proved that in this case the threshold for connectivity is $\Theta(\log n)$, showing in particular that if $k > 5.1774 \log n$ then G is connected with high probability and if $k < 0.074 \log n$ then G is disconnected with high probability. Balister, Bollobás, Sarkar and Walters [3] improved the upper and lower bounds to $0.5139 \log n$ and $0.3043 \log n$ respectively, and Walters [33] went on to further improve the upper bound to $0.4125 \log n$.

Choosing the k-nearest neighbours of each vertex gives the edges a natural orientation. With this in mind, define the directed random geometric graph $\overrightarrow{G} = \overrightarrow{G}(n, k)$ where for every point of a Poisson process of density 1 in S_n we add directed edges pointing out towards each of its k nearest neighbours. We say that this directed graph is connected if it is strongly connected, that is, if for every pair of points u, v there is a directed path from u to v and vice versa.

The directed random graph was also shown to have threshold for connectivity is $\Theta(\log n)$ by Balister, Bollobás, Sarkar and Walters [3], who proved upper and lower bounds of $0.9967 \log n$ and $0.7209 \log n$ respectively.

A natural question to ask is what is the connectivity threshold if you construct these graphs in higher-dimensional spaces? Let $\gamma_{d,n}$ be the d-dimensional cube of volume n (so the sidelengths of $\gamma_{d,n}$ are $n^{\frac{1}{d}}$) and let $\mathcal{P}$ be a Poisson process of density 1 in $\gamma_{d,n}$. Define the undirected random graph $G = G(d, n, k)$ and directed random graph $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ as in the 2-dimensional case.

Using very simple generalisations of the arguments in the 2-dimensional setting [3] it is easy to show that for fixed d, the threshold for connectivity is still $\Theta(\log n)$.These arguments give weak bounds on how the coefficient of $\log n$ depends on d: if $k = \Omega\left(\frac{1}{\log d} \log n\right)$ then G is connected with high probability and if $k = O\left(\frac{1}{e^d} \log n\right)$ then G is disconnected with high probability. Given the difference in terms of d between these bounds, one natural question is to ask how the threshold for connectivity depends on the dimension

d.

The main result of this chapter is to improve these bounds substantially.

Theorem 4.1. *Let $d \geq 2$. If $k \geq \frac{2.467}{d} \log n$ then $G(d, n, k)$ is connected with high probability. If $k \leq \frac{0.102}{d \log d} \log n$ then $G(d, n, k)$ is disconnected with high probability.*

We also establish bounds on the threshold for connectivity of the directed graph $\overrightarrow{G}$.

Theorem 4.2. *Let $d \geq 2$. If $k > \frac{2^d}{d} \log n$ then $\overrightarrow{G}(d, n, k)$ is connected with high probability. If $k < \frac{0.079}{\log d} \log n$ then $\overrightarrow{G}(d, n, k)$ is disconnected with high probability.*

In the course of each proof we reduce to considering only components of small diameter, where 'small' means less than $c(\log n)^{\frac{1}{d}}$ for some large constant c. In the directed case we must consider separately both components with no edges out and components with no edges in. The table below summarises the findings.

	Small component without adjacent		
	edges	out-edges	in-edges
No such component if $\frac{k}{\log n} >$	$\frac{2.467}{d}$	$\frac{1}{\log d}$	$\frac{1}{d} 2^d$
Exists such component if $\frac{k}{\log n} <$	$\frac{0.102}{d \log d}$	$\frac{0.721}{d}$	$\frac{0.079}{\log d}$

Table 4-A: Bounds on the thresholds for the existence of small diameter components.

Note that in the directed case the barrier to connectivity is the existence of small components with no in-edges. This is due to the way we construct the graph: each vertex is joined outwards to k nearest neighbours and so in the extreme case of a single vertex component there will always be out-edges, though there may well be no in-edges.

The exponential upper bound in the in-edge case comes from an argument involving vertices near the boundary of the cube. If we choose points at random in a d-dimensional torus rather than a d-dimensional cube we no longer need to worry about what happens at the boundaries. This allows us to improve the upper bound in the directed case.

Theorem 4.3. *Let $d \geq 2$ and let $\overrightarrow{G}_{tor}$ be the directed k-nearest-neighbour graph on a d-dimensional torus. If $k > 1.443 \log n$ then $\overrightarrow{G}_{tor}$ is connected with high probability. If $k < \frac{0.792}{\log d} \log n$ then $\overrightarrow{G}_{tor}$ is disconnected with high probability.*

We prove the upper bound on the threshold for connectivity for the undirected graph G in section 4.2. The upper bound for the directed graph $\overrightarrow{G}$ uses an adjustment of this argument and is proved in section 4.3, which also includes a proof of the improved upper bound when we instead consider the graph on a torus.

The lower bounds on the thresholds for connectivity for G and $\overrightarrow{G}$ are proved in sections 4.4 and 4.5 respectively. We end with some open questions in section 4.6.

4.2 An Upper Bound for the Undirected Graph

We will prove the following theorem.

Theorem 4.4. *Let $d \geq 2$. If $k > \frac{2.467 \log n}{d}$, then as $n \to \infty$ the undirected graph $G(d, n, k)$ is connected with high probability.*

Fix d and assume that $k = \lceil c \log n \rceil$ (where c might depend on d). Before we get to the proof of Theorem 4.4 we will prove two useful lemmas, Lemma 4.5 and 4.6, which will allow us to approximate the structure of $G = G(d, n, k)$.

Lemma 4.5. *There exist constants c_1, c_2 and c_3, depending on d and k but not n, such that with high probability $G = G(d, n, k)$ has the following properties:*

1. *All points distance $\leq c_1 (\log n)^{\frac{1}{d}}$ apart are joined by a (bidirectional) edge;*
2. *All points distance $\geq c_2 (\log n)^{\frac{1}{d}}$ apart are not joined by an edge;*
3. *Any half-ball of radius $c_3 (\log n)^{\frac{1}{d}}$ based at a point of G that is contained entirely within $\gamma_{d,n}$ contains at least one other point of G.*

Proof. This lemma follows from simple properties of the Poisson distribution.

1. Fix a vertex P and let $B = B(P, c_1(\log n)^{\frac{1}{d}})$ be the ball around P of radius $c_1(\log n)^{\frac{1}{d}}$. Suppose that P is not joined to every vertex of $B \cap \gamma_{d,n}$. Then $B \cap \gamma_{d,n}$ has volume at most $v = c_1^d \log n V_d$, where V_d is the volume of a d-dimensional unit ball, and it contains at least k additional vertices of G.

   The probability of this happening can be bounded as follows

$$\begin{aligned} p = e^{-v} \sum_{l=k}^{\infty} \frac{v^l}{l!} &< e^{-v} \sum_{l=k}^{\infty} \left(\frac{v}{k}\right)^{l-k} \frac{v^k}{k!} \\ &< e^{-v} \frac{k}{k-v} \frac{v^k}{k!} && \text{if } v < k \\ &< e^{-v} \frac{k}{k-v} \left(\frac{ve}{k}\right)^k \\ &\leq \frac{c}{c - c_1^d V_d} n^{-c_1^d V_d + c\left(1 + \log\left(\frac{c_1^d V_d}{c}\right)\right)}. \end{aligned}$$

   This is $o\left(n^{-1}\right)$ if both $c_1^d V_d < c$ and $-c_1^d V_d + c\left(1 + \log\left(\frac{c_1^d V_d}{c}\right)\right) < -1$. This works if, for example, we take c_1 such that $c_1^d V_d = c e^{-1-\frac{1}{c}}$.

   The expected number of vertices is n. Thus the expected number of vertices P such that the ball around P of radius $c_1(\log n)^{\frac{1}{d}}$ contains at least k additional vertices is $o(1)$. Hence the probability there is such a vertex is $o(1)$ and with high probability there is no such vertex P.

2. Fix a vertex P and let $B = B(P, c_2(\log n)^{\frac{1}{d}})$ be the ball around P of radius $c_2(\log n)^{\frac{1}{d}}$. Suppose that fewer than k of P's nearest neighbours are in $B \cap \gamma_{d,n}$. Then $B \cap \gamma_{d,n}$ contains at most k vertices of G. The volume of $B \cap \gamma_{d,n}$ is at least $v = c_2^d \log n V_d / 2^d$.

The probability of this happening can be bounded as follows

$$\begin{aligned} p = e^{-v}\sum_{l=0}^{k}\frac{v^l}{l!} &< e^{-v}\sum_{l=0}^{k}\left(\frac{k}{v}\right)^{k-l}\frac{v^k}{k!} \\ &< e^{-v}\frac{v}{v-k}\frac{v^k}{k!} && \text{if } v > k \\ &< e^{-v}\frac{v}{v-k}\left(\frac{ve}{k}\right)^k \\ &\leq \left(\frac{\left(\frac{c_2}{2}\right)^d V_d}{\left(\frac{c_2}{2}\right)^d V_d - c}\right) n^{-\left(\frac{c_2}{2}\right)^d V_d + c\left(1+\log\left(\frac{\left(\frac{c_2}{2}\right)^d V_d}{c}\right)\right)} \end{aligned}$$

This is $o\left(n^{-1}\right)$ if both $\left(\frac{c_2}{2}\right)^d V_d > c$ and $-\left(\frac{c_2}{2}\right)^d V_d + c\left(1+\log\left(\frac{\left(\frac{c_2}{2}\right)^d V_d}{c}\right)\right) < -1$. This works if, for example, we take c_2 such that $\left(\frac{c_2}{2}\right)^d V_d = e(1+c)$ (using that $\log\left(\frac{1+c}{c}\right) \leq \frac{1}{c}$).

The expected number of vertices is n and so the expected number of vertices P such that the ball around P of radius $c_2(\log n)^{\frac{1}{d}}$ contains fewer than k additional vertices is $o(1)$. Hence the probability there is such a vertex is $o(1)$ and with high probability there is no such vertex P.

3. Fix a vertex P and let $B = B(P, c_3(\log n)^{\frac{1}{d}})$ be the ball around P of radius $c_3(\log n)^{\frac{1}{d}}$. Split B into 2^d regions using hyperplanes through P parallel to the faces of $\gamma_{d,n}$. We shall show that any one of these regions that is fully contained within $\gamma_{d,n}$ is non-empty with high probability. Since any half-ball based at P must fully contain one of the 2^d regions, this will suffice to prove the result.

   Let B' be one of the 2^d regions and suppose that it is empty and contained entirely within $\gamma_{d,n}$. B' has volume $v = c_3^d \log n V_d/2^d$. The probability that it is empty is

$$p = e^{-v} = n^{-\left(\frac{c_3}{2}\right)^d V_d},$$

   which is $o\left(n^{-1}\right)$ if we take c_3 such that $\left(\frac{c_3}{2}\right)^d V_d > 1$.

   The expected number of vertices is n and so the expected number of empty regions

based at a vertex of G is $n2^d p = o(1)$. Hence with high probability no such region is empty.

□

Lemma 4.6. *There exists a constant c_4 depending on d and k but not n such that with high probability $G = G(d, n, k)$ has at most one component of large diameter $\geq c_4(\log n)^{\frac{1}{d}}$.*

The proof of Lemma 4.6 follows the same approach as the proof of Lemma 12 in [3]. First, we show that two components cannot be too close together.

Lemma 4.7. *With high probability the distance between any two edges belonging to different components of G is at least $c_1(\log n)^{\frac{1}{d}}/2$.*

Proof. By Lemma 4.5 we may assume that any vertices that do not have an edge between them are distance $> c_1(\log n)^{\frac{1}{d}}$ apart for some c_1. Let $\gamma = c_1(\log n)^{\frac{1}{d}}$.

Let xy be an edge in one component of G and wz be an edge in a different component of G. Let u be the point on the line segment xy that is closest to wz, let v be the point on wz that is closest to xy, and let $h = d(u, v)$ be the distance between xy and wz. We need to show that $h > \gamma/2$.

Suppose first that u is an endpoint of xy and v is an endpoint of wz. Without loss of generality $u = x$ and $v = w$. Since there is no edge between x and w, $h = d(x, w) > \gamma$ and we are done.

Suppose next that exactly one of u and v is an endpoint of their respective edges, say $u = x$. Without loss of generality let z be one of the k nearest neghbours of w. Since x is not one of the k nearest neighbours of w, we must have $d(w, x) \geq d(w, z)$, By the triangle inequality, we have $d(v, x) > d(z, x) - d(z, v)$ and $d(v, x) > d(w, x) - d(w, v) > d(w, z) - d(w, v)$. Summing these, we have that

$$2d(v, x) > d(z, x) + d(w, z) - d(z, v) - d(w, v) = d(z, x) > \gamma$$

and so $h = d(v, x) > \gamma/2$ as required.

Finally, suppose that neither u nor v are endpoints of their respective edges. Then the line segment uv is orthogonal to both xy and wz. Without loss of generality suppose that y is one of the k nearest neighbours of x and z is one of the k nearest neighbours of w. We observe that the following inequalities must hold:

$$d(w, y) > d(w, z) \tag{4.1}$$

$$d(w, x) > d(w, z) \tag{4.2}$$

$$d(x, z) > d(x, y) \tag{4.3}$$

$$d(x, w) > d(x, y) \tag{4.4}$$

$$d(x, z) > \gamma \tag{4.5}$$

$$d(w, y) > \gamma \tag{4.6}$$

$$d(y, z) > \gamma \tag{4.7}$$

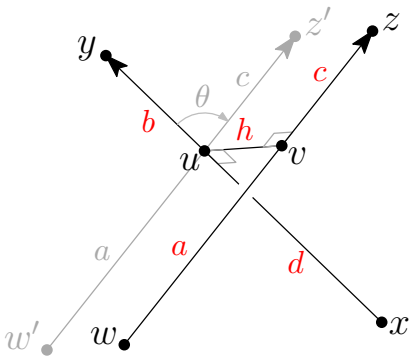


Figure 4.1: The edges xy, wz and $w'z'$ with lengths a, b, c, d, h labelled.

Let $a = d(w, v)$, $b = d(y, u)$, $c = d(z, v)$ and $d = d(x, u)$, so that $d(x, y) = b + d$ and $d(w, z) = a + c$. Let $w' = w + \overrightarrow{vu}$ and $z' = z + \overrightarrow{vu}$, so that $\overrightarrow{w'z'}$ and $\overrightarrow{xy}$ lie in the same plane with normal $\overrightarrow{uv}$. Then let θ be the angle between the vectors $\overrightarrow{xy}$ and $\overrightarrow{w'z'}$. See figure 4.1 for a diagram.

Using the cosine rule in triangle $w'uy$ and Pythagoras's Theorem on triangle $ww'y$

we can calculate that

$$d(w,y)^2 = \left(a^2 + b^2 + 2ab\cos\theta\right) + h^2.$$

By a similar argument,

$$\begin{aligned} d(x,z)^2 &= \left(c^2 + d^2 + 2cd\cos\theta\right) + h^2, \\ d(w,x)^2 &= \left(a^2 + d^2 - 2ad\cos\theta\right) + h^2, \text{ and} \\ d(y,z)^2 &= \left(b^2 + c^2 - 2bc\cos\theta\right) + h^2. \end{aligned}$$

Substituting these into the squares of the first set of inequalities and rearranging gives

$$h^2 > c^2 - b^2 + 2ac - 2ab\cos\theta \tag{4.1$'$}$$

$$h^2 > c^2 - d^2 + 2ac + 2ad\cos\theta \tag{4.2$'$}$$

$$h^2 > b^2 - c^2 + 2bd - 2cd\cos\theta \tag{4.3$'$}$$

$$h^2 > b^2 - a^2 + 2bd + 2ad\cos\theta \tag{4.4$'$}$$

$$h^2 > \gamma^2 - c^2 - d^2 - 2cd\cos\theta \tag{4.5$'$}$$

$$h^2 > \gamma^2 - a^2 - b^2 - 2ab\cos\theta \tag{4.6$'$}$$

$$h^2 > \gamma^2 - b^2 - c^2 + 2bc\cos\theta \tag{4.7$'$}$$

Consider the following linear combination of these inequalities:

$$\frac{1}{b+d}\left(d(4.1') + b(4.2') + b(4.5') + d(4.7')\right) + \frac{1}{a+c}\left(a(4.3') + c(4.4') + c(4.6') + a(4.7')\right).$$

Cancelling like terms, this gives that

$$4h^2 > \frac{1}{b+d}\left(b\gamma^2 + d\gamma^2 - 2b^2d - 2bd^2 + 2acd + 2abc\right) + \frac{1}{a+c}\left(c\gamma^2 + a\gamma^2 - 2a^2c - 2ac^2 + 2abd + 2bcd\right)$$

which further simplifies to $4h^2 > 2\gamma^2$ and so $h > \frac{\gamma}{\sqrt{2}}$. This completes the proof. □

We use this result to prove the lemma.

Proof of Lemma 4.6. By Lemma 4.5, we may assume that any two vertices within distance $c_1(\log n)^{\frac{1}{d}}$ are connected. Suppose there exist two components G_1 and G_2 each of diameter at least $D = c_4(\log n)^{\frac{1}{d}}$.

Tile the cube $\gamma_{d,n}$ with tiles of side-length $s = \frac{c_1(\log n)^{\frac{1}{d}}}{4\sqrt{d}}$. Colour a tile red if it contains a vertex of G_1 or intersects an edge of G_1. Colour a tile blue if it contains a vertex of G_2 or intersects an edge of G_2. Colour a tile black if it is neither red or blue but contains a vertex, and colour a tile white if it contains no vertices. This is well-defined by Lemma 4.7. Furthermore, the distance between two points in two tiles that meet (even if only at a corner) is $< 2s\sqrt{d} = c_1(\log n)^{\frac{1}{d}}/2$, and so by Lemma 4.7 a red tile can only touch red or white tiles, and a blue tile can only touch blue or white tiles.

Let $l = \frac{n^{\frac{1}{d}}}{s}$ (assuming for convenience that this is an integer) and identify the tiling with the d-dimensional grid graph $[l]^d$. The set of red tiles R and the set of blue tiles B form connected components. Since G_1 and G_2 have diameter at least D and each tile has diameter $s\sqrt{d}$, the number of red tiles is at least $\frac{D}{s\sqrt{d}} = \frac{4c_4}{c_1}$ and the number of blue tiles is at least $\frac{4c_4}{c_1}$ too.

The complement of B splits into components $C_1, C_2, \ldots, C_t$. Since R is connected it is contained entirely within one of these components, say C_1. Now consider C_1 and its complement $C_1^c = B \cup \bigcup_{i=2}^{t} C_i$. Both C_1 and C_1^c are connected and since C_1 contains R and C_1^c contains B they both contain at least $\frac{4c_4}{c_1}$ tiles.

Let ∂C_1^c be the set of all tiles that touch C_1^c but are not contained in C_1^c. Since $C_1^c = B \cup \bigcup_{i=2}^{t} C_i$, by our earlier observation all tiles in ∂C_1^c must be white. Also note that ∂C_1^c is connected in the grid $[l]^d$ and it contains the usual vertex boundary of C_1^c.

By the vertex isoperimetric inequality in the grid [5], $|\partial C_1^c|$ is bounded below by

the size of the vertex boundary of a simplex of volume $\frac{4c_4}{c_1}$. This is some constant c' depending on d and c_4 but independent of n.

Hence we have $|\partial C_1^c|$ is a connected component of size $> c'$ containing only empty tiles. We now just need to bound the probability of the existence of such a set. The probability that a set containing c' tiles is empty is $e^{-c's^d}$.

For any graph with maximum degree Δ, the number of connected subsets of size t containing a particular vertex is at most $(e\Delta)^t$. Thus the number of connected components of $[l]^d$ of size c' containing a particular tile is at most $(e2d)^{c'}$ and there are at most $l^d(e2d)^{c'}$ such components in total.

Putting all of this together, the probability that there is a connected set of c' empty tiles is

$$p = l^d(e2d)^{c'}e^{-c's^d} = \frac{\left(\frac{4\sqrt{d}}{c_1}\right)^d(e2d)^{c'}}{\log n}n^{1-c'\left(\frac{c_1}{4\sqrt{d}}\right)^d}$$

Recall that c' is the size of the vertex boundary of a simplex of volume $\frac{4c_4}{c_1}$. Thus by choosing c_4 large enough, thus ensuring that c' is large enough, we can obtain $p = o(1)$.

□

4.2.1 Proof of Theorem 4.4

We need only consider graphs G that have the four properties given by Lemmas 4.5 and 4.6, since these properties occur with high probability. Let c_1, c_2, c_3, c_4 be the constants defined in these two lemmas. We will use these properties to obtain a granular model $\widehat{G}$ of the graph G.

Tile the cube $\gamma_{d,n}$ with tiles of small side-length $s < c_5(\log n)^{\frac{1}{d}}$, where $c_5 = c_5(d)$ is a small constant independent of n. (Note that this may be a different side-length s to the one given in the proof of Lemma 4.6.) Any two points in the same tile are distance

$\leq c_5\sqrt{d}$ apart, so we insist that $c_5 < \frac{c_1}{\sqrt{d}}$ to guarantee that any pair of points that are in the same tile are connected.

Let $\widehat{G}$ be the graph where the vertices are the tiles and two tiles are joined if the distance between their centres is less than $c_2(\log n)^{\frac{1}{d}} + \sqrt{d}s$. If two points are connected then the distance between them is $< c_2(\log n)^{\frac{1}{d}}$, and so their respective tiles are connected.

Note that $\widehat{G}$ has bounded degree independent of n: each tile is connected to fewer than $\left(2\left\lceil\frac{c_2(\log n)^{\frac{1}{d}}+\sqrt{d}s}{s}\right\rceil\right)^d = \left(2\left\lceil\frac{c_2}{c_5}+\sqrt{d}\right\rceil\right)^d$ other tiles.

Suppose that G is not connected. Then it contains a component C of small diameter $< c_4(\log n)^{\frac{1}{d}}$. Let $\widehat{C}$ be the collection of tiles containing points of C. Then the following hold:

- if a point is in a tile in $\widehat{C}$ then it is in C (since any two points in the same tile are connected);
- the tiles in $\widehat{C}$ form a connected subgraph in $\widehat{G}$ (since if two points are connected then so are the tiles containing them);
- The number of tiles in $\widehat{C}$ is bounded by a constant independent of n. In particular, the number of tiles is less than $\left(2\left\lceil\frac{c_4(\log n)^{\frac{1}{d}}+\sqrt{d}s}{s}\right\rceil\right)^d = \left(2\left\lceil\frac{c_4}{c_5}+\sqrt{d}\right\rceil\right)^d$.

Since $\widehat{G}$ has bounded degree, the number of connected, bounded-size subgraphs of $\widehat{G}$ containing any given tile is bounded by a constant. Summing over all tiles containing a point, in total there are $O(n)$ possibilities for $\widehat{C}$.

Our aim will be to show that the probability of a particular connected subset $\widehat{C}$ actually containing a connected component C of G (i.e. having no in-edges or out-edges) is $o\left(\frac{1}{n}\right)$ when $k = \frac{2\log n}{d}$.

First some notation: for a set S let $|S|$ denote the volume of S and let $\#S$ denote the number of points of G contained in S. We will use in our proof the following simple technical lemma due to Walters [33].

Lemma 4.8. *Suppose A, B and C are three sets in $\gamma_{d,n}$ with $|A| \leq |C|$ and $|B| \leq |C|$. Then*

$$\mathbb{P}(\#A \geq k, \#B \geq k, \#(A \cap B) = 0 \text{ and } \#C = 0) \leq \left(\frac{4|A||B|}{(|A| + |B| + |C|)^2}\right)^k.$$

Proof. See Lemma 6 of [33]; the proof given is independent of the number of dimensions. □

4.2.1.1 If C is not near the boundary

For ease of explanation, we shall first run the argument ignoring any issues that arise due to the boundary of $\gamma_{d,n}$. We assume that no point of C is within distance $c_3(\log n)^{\frac{1}{d}}$ of the boundary. We use two constructions to bound the probability and combine them.

First, let us use that there are no out-edges. Consider taking the smallest box X that contains the component C. On each face of the box there is a point x_i of the component: let $B_1, \ldots, B_{2d}$ be the k-nearest-neighbour balls of each of these x_i. Pick j such that $B_j \cap X$ has the smallest volume, and let $A' = B_j \cap \widehat{C}$. For each i let A_i be the region $B_i \cap X$ reflected in the relevant face of X. Note that all these regions are disjoint, A' has the smallest volume, and if the component has no edges out, there must be $\geq k$ vertices in A' and no vertices in each A_i. See figure 4.2 for an example in 2 dimensions.

We combine this with another construction. Let $P \in C$ and $Q \notin C$ be such that the distance between them, $r_0 = d(P, Q)$, is minimised. Note that since P and Q are not connected, we must have $r_0 \geq c_1(\log n)^{\frac{1}{d}}$. We also have that $r_0 \leq c_3(\log n)^{\frac{1}{d}}$, since the half-ball of radius $c_3(\log n)^{\frac{1}{d}}$ around the right-most point of C is not empty. Then let $B = B_{r_0}(Q) \setminus B_{r_0}(P)$ and note that B must contain Q's k nearest neighbours. Let $r = r_0 - s\sqrt{d}$. We have $r \leq c_3(\log n)^{\frac{1}{d}} - s\sqrt{d}$ and so $\widehat{C}_{(r)}$ is contained in $\gamma_{d,n}$.

See figure 4.3 for an example in 2 dimensions.

Let $x = \left(\frac{r^d V_d}{|A'|}\right)^{\frac{1}{d}}$ where V_d is the volume of a d-dimensional ball of radius 1. Note

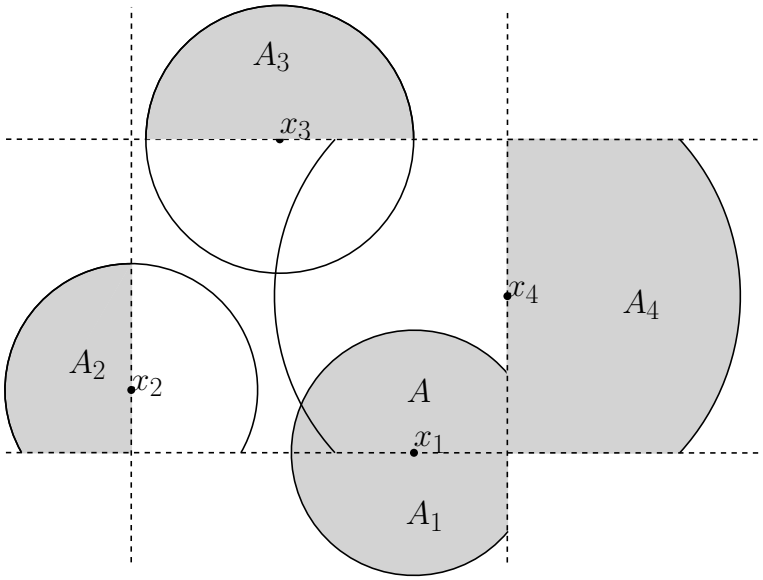


Figure 4.2: The first construction for the 2-dimensional case

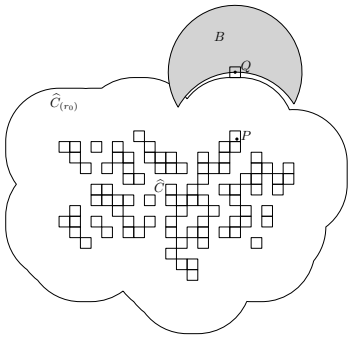


Figure 4.3: The second construction for the 2-dimensional case

that

$$\begin{aligned} x^d|A'| = r^d V_d = \left(\frac{r}{r_0}\right)^d r_0^d V_d &= \left(1 - \frac{s\sqrt{d}}{r_0}\right)^d r_0^d V_d \\ &\geq \left(1 - \frac{c_5(\log n)^{\frac{1}{d}}\sqrt{d}}{c_1(\log n)^{\frac{1}{d}}}\right)^d r_0^d V_d \quad \text{since } r_0 \geq c_1(\log n)^{\frac{1}{d}} \\ &= \left(1 - \frac{c_5\sqrt{d}}{c_1}\right)^d r_0^d V_d \end{aligned}$$

Note that B has volume strictly less than a ball of radius r_0 with a ball of radius $\frac{r_0}{2}$ removed, and so $|B| < r_0^d V_d - \left(\frac{r_0}{2}\right)^d V_d = \left(1 - \frac{1}{2^d}\right) r_0^d V_d$. Using this, we take $c_5 = c_5(d)$

sufficiently small to get that

$$x^d|A'| > |B|.$$

Let us now use these constructions to bound the probability of a small connected component C.

Start with a connected component $\widehat{C}$ of $\widehat{G}$ containing $< \left(2\left\lceil \frac{c_4}{c_5} + \sqrt{d}\right\rceil\right)^d$ tiles. Suppose that the points contained in the tiles of $\widehat{C}$ form a connected component C of the graph G. Then, as we've discussed, there exist regions $A', A_1, A_2 \ldots, A_{(2d)}$ determined by $2d$ points defining the faces of the convex hull and the kth nearest neighbour of each of these. There is also the region B determined by the points P and Q. All of these $4d+2$ points are in C, except for Q, which is the nearest point to C that is not itself in C.

Let Z be the event that, given $\widehat{C}$, there are $4d+2$ points of G defining regions as above, with $\#A' \geq k, \#B \geq k, \#\bigcup_{i=1}^{2d} A_i = 0$ and $\#\widehat{C}_{(r)} \setminus \widehat{C} = 0$. Then the probability that the points contained in $\widehat{C}$ form a connected component is at most the probability of Z.

Fix a particular collection of $4d+2$ points of G and let Z' be the event that these points witness Z.

First, let us use the first construction and apply lemma 4.8 with $A = A', B = B$ and $C = \bigcup_{i=1}^{2d} A_i$. We need to bound the sizes of these regions. Clearly, we have $|A'| \leq 2d|A'| \leq |\bigcup_{i=1}^{2d} A_i|$. If $x \leq (2d)^{\frac{1}{d}}$ then $|B| \leq x^d|A'| \leq 2d|A'| \leq |\bigcup_{i=1}^{2d} A_i|$.

Thus if $x < (2d)^{\frac{1}{d}}$ then

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \#B \geq k, \#\bigcup_{i=1}^{2d} A_i = 0\right) &\leq \left(\frac{4|A'||B|}{(|A'|+|B|+|\bigcup_{i=1}^{2d} A_i|)^2}\right)^k. \\ &\leq \left(\frac{4\frac{|B|}{|A'|}}{\left(1+\frac{|B|}{|A'|}+2d\right)^2}\right)^k \\ &\leq \left(\frac{4x^d}{(1+x^d+2d)^2}\right)^k\end{aligned}$$

where it is easy to check that the last inequality holds given $\frac{|B|}{|A'|} \leq x^d \leq 2d$.

Next, let us use the second construction and apply lemma 4.8 with $A = A', B = B$ and $C = (\widehat{C}_{(r)} \setminus \widehat{C})$.

We need to bound $|\widehat{C}_{(r)} \setminus \widehat{C}|$. Let $|D|$ and $|D'|$ be balls of volume $|\widehat{C}_{(r)}|$ and $|A'|$ respectively. Note that $r = r_0 - s\sqrt{d} < c_3(\log n)^{\frac{1}{d}} - s\sqrt{d}$ and so $\widehat{C}_{(r)}$ doesn't intersect the boundary. Thus by the isoperimetric inequality,

$$|\widehat{C}_{(r)} \setminus \widehat{C}| \geq |D_{(r)} \setminus D| > |D'_{(r)} \setminus D'|$$

D' is a ball of radius $\left(\frac{|A'|}{V_d}\right)^{\frac{1}{d}}$, and so $D'_{(r)}$ has radius $\left(\frac{|A'|}{V_d}\right)^{\frac{1}{d}} + r = (1+x)\left(\frac{|A'|}{V_d}\right)^{\frac{1}{d}}$. Thus

$$|\widehat{C}_{(r)} \setminus \widehat{C}| \geq |D'_{(r)} \setminus D'| = \left((1+x)^d - 1\right)|A'|.$$

We have $|B| \leq x^d|A'| < \left((1+x)^d - 1\right)|A'| \leq |\widehat{C}_{(r)} \setminus \widehat{C}|$. If $x \geq 2^{\frac{1}{d}} - 1$ then $|A'| \leq \left((1+x)^d - 1\right)|A'| \leq |\widehat{C}_{(r)} \setminus \widehat{C}|$.

Thus if $x \geq 2^{\frac{1}{d}} - 1$ then

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \#B \geq k, \#\widehat{C}_{(r)} \setminus \widehat{C} = 0\right) &\leq \left(\frac{4|A'||B|}{(|A'| + |B| + |\widehat{C}_{(r)} \setminus \widehat{C}|)^2}\right)^k. \\ &\leq \left(\frac{4\frac{|B|}{|A'|}}{\left(1 + \frac{|B|}{|A'|} + (1+x)^d - 1\right)^2}\right)^k \\ &\leq \left(\frac{4x^d}{((x^d + (1+x)^d)^2}\right)^k\end{aligned}$$

where again it is easy to check that the last inequality holds.

Combining these two bounds we get that

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4x^d}{\left(2d+1+x^d\right)^2}\right)^k & \text{if } x \leq (2d)^{\frac{1}{d}} \\ \left(\frac{4x^d}{\left((x+1)^d+x^d\right)^2}\right)^k & \text{if } x \geq 2^{\frac{1}{d}} - 1 \end{cases}$$

It is straightforward to see that the former is smaller when $(x+1)^d \leq 2d+1$ and the latter is smaller otherwise. Since $2^{\frac{1}{d}} - 1 < (2d+1)^{\frac{1}{d}} - 1 < (2d)^{\frac{1}{d}}$, we can re-write this as

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4x^d}{\left(2d+1+x^d\right)^2}\right)^k & \text{if } x \leq (2d+1)^{\frac{1}{d}} - 1 \\ \left(\frac{4x^d}{\left((x+1)^d+x^d\right)^2}\right)^k & \text{if } x \geq (2d+1)^{\frac{1}{d}} - 1 \end{cases}$$

One can check that the first expression is increasing for $0 < x < (2d+1)^{\frac{1}{d}}$, and so it is maximised at $x = (2d+1)^{\frac{1}{d}} - 1$, where it is equal to the second expression. Thus we only need to bound the second expression for $x \geq (2d+1)^{\frac{1}{d}} - 1$.

We have that

$$\frac{4x^d}{((x+1)^d + x^d)^2} \leq \frac{4x^d}{((x+1)^d)^2}$$

and by differentiating, we find that this is maximised when $x = 1$. In particular,

$$\mathbb{P}(Z') \leq \left(\frac{4}{2^{2d}}\right)^k = 4^{k(1-d)}.$$

Now we can calculate the probability of Z. Recall that we chose $4d+2$ points within $c_2 (\log n)^{\frac{1}{d}}$ of $\widehat{C}$, so the number of tiles we could choose from was bounded by a constant. Thus we picked the $4d + 2$ points within an area of size $O(\log n)$, giving with high probability $O(\log n)^{4d+2}$ ways of choosing the points. The expected number of sets of points that witness Z is therefore $O\left((\log n)^{4d+2}\right)\mathbb{P}(Z')$. So we have

$$\mathbb{P}(Z) = O\left((\log n)^{4d+2} 4^{k(1-d)}\right)$$

We conclude that if $\frac{k}{\log n} > \frac{1.443}{d} > \frac{2}{d \log 4} \geq \frac{1}{(d-1)\log 4}$ then $\mathbb{P}(Z) = o(n^{-1})$.

Finally, we can bound the probability that G is connected. We already observed that the number of possible connected bounded-size subgraphs $\widehat{C}$ is $O(n)$. Thus the probability that there is a component in G of small diameter $< c_4 \log n^{\frac{1}{d}}$ is $O(n)\mathbb{P}(Z) = o(1)$ if $\frac{k}{\log n} > \frac{1.443}{d}$. In particular, if $\frac{k}{\log n} > \frac{1.443}{d}$ then with high probability there are no components in G of diameter $< c_4 \log n^{\frac{1}{d}}$. By Lemma 4.6, G has only one component with high probability.

4.2.1.2 If C is near the boundary

Let f be the number of faces of the cube $\gamma_{d,n}$ that are within distance $c_3(\log n)^{\frac{1}{d}}$ of $\widehat{C}$.

As before, construct the smallest box X containing the component C, where X has faces parallel to the faces of the cube $\gamma_{d,n}$, and take x_i to be the points of C lying on each face of the box. For each x_i construct B_i, the k-nearest-neighbour ball around x_i intersected with $\gamma_{d,n}$, and construct A_i, the region $B_i \cap \widehat{C}$ reflected in the relevant face of X, again intersected with $\gamma_{d,n}$. Pick j such that $B_j \cap X$ has the smallest volume and let $A' = B_j \cap \widehat{C}$.

Up to f of the A_i might intersect the boundary of $\gamma_{d,n}$. Without loss of generality, suppose that $A_1, \ldots, A_{2d-f}$ do not intersect the boundary. We will ignore these regions. It is the still the case that A has the smallest volume out of $A, A_1, \ldots, A_{2d-f}$.

Let Z be the event that, given $\widehat{C}$, there are $4d+2$ points of G defining regions as above, with $\#A' \geq k, \#B \geq k, \#\bigcup_{i=1}^{2d-f} A_i = 0$ and $\#\left(\widehat{C}_{(r)} \setminus \widehat{C}\right) \cap \gamma_{d,n} = 0$. Then the probability that the points contained in $\widehat{C}$ form a connected component is at most the probability of Z.

Fix a particular collection of $4d+2$ points of G and let Z' be the event that these points witness Z.

Let us use the first construction and apply lemma 4.8 with $A = A', B = B$ and $C = \bigcup_{i=1}^{2d-f} A_i$. Clearly, we have $|A'| \leq (2d-f)|A'| \leq |\bigcup_{i=1}^{2d-f} A_i|$. Let $x = \left(\frac{r^d V_d}{|A'|}\right)^{\frac{1}{d}}$ as before, and note that we still have $|B| < x^d|A'|$. If $x \leq (2d-f)^{\frac{1}{d}}$ then $|B| \leq x^d|A'| \leq (2d-f)|A'| \leq |\bigcup_{i=1}^{2d-f} A_i|$.

Thus if $x < (2d-f)^{\frac{1}{d}}$ then

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \#B \geq k, \#\bigcup_{i=1}^{2d-f} A_i = 0\right) &\leq \left(\frac{4|A'||B|}{(|A'|+|B|+|\bigcup_{i=1}^{2d-f} A_i|)^2}\right)^k \\ &\leq \left(\frac{4x^d}{(1+x^d+2d-f)^2}\right)^k\end{aligned}$$

Now consider the second construction. Let r_0 be the minimal distance between a vertex in C and a vertex not in C and let $r = r_0 - s\sqrt{d}$. Note that since C does not intersect the boundary and $\widehat{C}_{(r)}$ intersects at most f faces, we have that $|(\widehat{C}_{(r)} \setminus \widehat{C}) \cap \gamma_{d,n}| \geq \left(\frac{1}{2}\right)^f |(\widehat{C}_{(r)} \setminus \widehat{C})|$. Using the bound we calculated in Section 4.2.1.1 we have $|(\widehat{C}_{(r)} \setminus \widehat{C}) \cap \gamma_{d,n}| \geq \left(\frac{1}{2}\right)^f \left((1+x)^d - 1\right)|A'|$.

We will split into two cases. First, suppose $x \geq 1$. Then

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \#\left((\widehat{C}_{(r)} \setminus \widehat{C}) \cap \gamma_{d,n}\right) = 0\right) &\leq \left(\frac{|A'|}{|A' \cup (\widehat{C}_{(r)} \setminus \widehat{C}) \cap \gamma_{d,n}|}\right)^k. \\ &\leq \left(\frac{1}{1 + \left(\frac{1}{2}\right)^f ((1+x)^d - 1)}\right)^k \\ &\leq \left(\frac{2^f}{(1+x)^d}\right)^k\end{aligned}$$

Next, suppose $x \leq 1$. Let us try to apply lemma 4.8 with $A = A', B = B$ and $C = (\widehat{C}_{(r)} \setminus \widehat{C}) \cap \gamma_{d,n}$. We have that $|B| \leq x^d |A'| \leq |A'|$. If we also have that $1 \leq \left(\frac{1}{2}\right)^f \left((1+x)^d - 1\right)$ then $|A'| \leq \left(\frac{1}{2}\right)^f \left((1+x)^d - 1\right)|A'| \leq |\widehat{C}_{(r)} \setminus \widehat{C}|$.

Thus if $(2^f + 1)^{\frac{1}{d}} - 1 \leq x \leq 1$ we can apply the lemma to get

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \#B \geq k, \#\widehat{C}_{(r)} \setminus \widehat{C} = 0\right) &\leq \left(\frac{4|A'||B|}{(|A'| + |B| + |\widehat{C}_{(r)} \setminus \widehat{C}|)^2}\right)^k. \\ &\leq \left(\frac{4\frac{|B|}{|A'|}}{\left(1 + \frac{|B|}{|A'|} + \left(\frac{1}{2}\right)^f ((1+x)^d - 1)\right)^2}\right)^k \\ &\leq \left(\frac{4x^d}{(x^d + 2^{-f}(1+x)^d)^2}\right)^k\end{aligned}$$

Combining these calculations for each construction, we have that

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4x^d}{\left(2d - f + 1 + x^d\right)^2}\right)^k & \text{if } x \leq (2d-f)^{\frac{1}{d}} \\ \left(\frac{4x^d}{\left(2^{-f}(x+1)^d + x^d\right)^2}\right)^k & \text{if } (2^f+1)^{\frac{1}{d}} - 1 \leq x \leq 1 \\ \left(\frac{2^f}{(1+x)^d}\right)^k & \text{if } 1 \leq x \end{cases}$$

We will split into cases depending on the value of f.

Case 1: $f = d$. We have that

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4x^d}{\left(d+1+x^d\right)^2}\right)^k & \text{if } x \leq d^{\frac{1}{d}} \\ \left(\frac{2^d}{(1+x)^d}\right)^k & \text{if } d^{\frac{1}{d}} \leq x \end{cases}$$

The former bound is increasing with x on its domain, while the latter is decreasing with x on its domain. Thus

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4d}{(2d+1)^2}\right)^k & \text{if } x \leq d^{\frac{1}{d}} \\ \left(\frac{2}{1+d^{\frac{1}{d}}}\right)^{dk} & \text{if } d^{\frac{1}{d}} \leq x \end{cases}$$

In particular, we have $\mathbb{P}(Z') \leq (c')^k$, where $c' = \max\left(\frac{4d}{(2d+1)^2}, \left(\frac{2}{1+d^{\frac{1}{d}}}\right)^d\right) < 1$ is some constant depending only on d.

Case 2: $f < d-1$. We have $(2^f+1)^{\frac{1}{d}} - 1 \leq 1 \leq (2d-f)^{\frac{1}{d}}$ and so we can weaken the bounds to get

$$\mathbb{P}(Z') \leq \begin{cases} \left(\frac{4x^d}{\left(2d-f+1+x^d\right)^2}\right)^k & \text{if } x \leq (2^f+1)^{\frac{1}{d}} - 1 \\ \left(\frac{4x^d}{\left(2^{-f}(x+1)^d+x^d\right)^2}\right)^k & \text{if } (2^f+1)^{\frac{1}{d}} - 1 \leq x \leq 1 \\ \left(\frac{2^f}{(1+x)^d}\right)^k & \text{if } 1 \leq x \end{cases}$$

Note that the final constraint is decreasing with x, so is maximised at $x = 1$, giving $\mathbb{P}(Z') \leq \frac{1}{2^{k(d-f)}}$ when $x \geq 1$.

The first constraint is increasing with x on its domain, so is maximised at $x = (2^f+1)^{\frac{1}{d}} - 1$. Note that for this value of x we have $2^{-f}(x+1)^d = 2^{-f}(2^f+1) < 2d-f$, which means that the first constraint is strictly smaller than the second constraint. Thus we only need to bound the second expression. We have

$$\frac{4x^d}{\left(2^{-f}(x+1)^d + x^d\right)^2} \leq \frac{4x^d}{\left(2^{-f}(x+1)^d\right)^2} \leq \frac{4}{4^{d-f}}$$

where the second inequality comes from the fact that the expression is maximised at $x = 1$. Putting this together with the $x \geq 1$ case, we have that when $f \leq d-2$ the probability satisfies

$$\mathbb{P}(Z') \leq \max\left(\left(\frac{1}{2^{d-f}}\right)^k, \left(\frac{4}{4^{d-f}}\right)^k\right) = \left(\frac{1}{2^{d-f}}\right)^k = 2^{-k(d-f)}$$

Case 3: $f = d-1$. The calculations for the $f < d-1$ case still hold in this case, but the final bound is not quite strong enough when $f = d-1$. In this case, we use that $(x+1)^2 \geq 4x$ and $(x+1)^d > 2^f = 2^{d-1}$ to get the following improved bound:

$$\begin{aligned}\frac{4x^d}{\left(2^{-(d-1)}(x+1)^d + x^d\right)^2} &\leq \frac{4x^d}{4^{1-d}\left((x+1)^2\right)^d + 2\left(2^{1-d}\right)x^d(x+1)^d} \\ &\leq \frac{4x^d}{4^{1-d}(4x)^d + 2\left(2^{1-d}\right)x^d 2^{d-1}} = \frac{2}{3}\end{aligned}$$

Again, we combine this with the bound when $x \geq 1$ to get that when $f = d-1$ the probability satisfies

$$\mathbb{P}(Z') \leq \max\left(\left(\frac{1}{2}\right)^k, \left(\frac{2}{3}\right)^k\right) = \left(\frac{2}{3}\right)^k$$

Because $\widehat{C}$ is within distance $c_3 \log n^{\frac{1}{d}}$ of f faces and has diameter $< c_4 \log n^{\frac{1}{d}}$, we have that $\widehat{C}$ must contain a vertex that is within distance $(c_3 + c_4)\log n^{\frac{1}{d}}$ of f faces of the cube $\gamma_{d,n}$. Call such a vertex x.

The volume of the region that is within distance $(c_3 + c_4)(\log n)^{\frac{1}{d}}$ of f faces of the cube $\gamma_{d,n}$ is $\leq \binom{d}{f} 2^f \left((c_3 + c_4)\log n^{\frac{1}{d}}\right)^f \left(n^{\frac{1}{d}}\right)^{(d-f)}$. Thus with high probability there are $O\left(n^{\frac{d-f}{d}+o(1)}\right)$ choices for x, giving that with high probability the number of possible connected bounded-size components $\widehat{C}$ within distance $c_3(\log n)^{\frac{1}{d}}$ of f faces of $\gamma_{d,n}$ is also $O\left(n^{\frac{d-f}{d}+o(1)}\right)$.

As before, given $\widehat{C}$ there are $O\left((\log n)^{4d+2}\right)$ ways of choosing the $4d+2$ points that define Z, and so putting everything together we have that the probability that there is a component of diameter $< c_4(\log n)^{\frac{1}{d}}$ is

$$\begin{cases} O\left(n^{\frac{d-f}{d}+o(1)}2^{-k(d-f)}\right) & \text{if } f < d-1 \\ O\left(n^{\frac{1}{d}+o(1)}\left(\frac{2}{3}\right)^k\right) & \text{if } f = d-1 \\ O\left(n^{o(1)}\left(c'\right)^k\right) & \text{if } f = d \end{cases}$$

It is now clear to see that taking $\frac{k}{\log n} > \frac{2.467}{d} > \frac{1}{\left(\log\frac{3}{2}\right)d}$ gives that the expected number of small components is $o(1)$ (using that $c' < 1$). Thus when $k > \frac{2.467\log n}{d}$, with high probability there are no small diameter components.

4.3 An Upper Bound for the Directed Graph

Theorem 4.9. *If $\frac{k}{\log n} > \frac{2^d}{d}$, then as $n \to \infty$ the directed graph $\overrightarrow{G} = \overrightarrow{G}(d,n,k)$ is connected with high probability.*

The proof of Theorem 4.9 will follow a similar approach to the proof of Theorem 4.4. Lemma 4.5 still holds in the directed case by the same proof. We will need something analogous to Lemma 4.6.

Fix d and assume that $k = \lceil c\log n\rceil$ (where c might depend on d).

Call a set C an *in-component* if it has no adjacent edges coming in to it, and an *out-component* if it has no adjacent edges going out of it. Note that any two disjoint in-components have no edges between them and any two disjoint out-components have no edges between them. Thus by Lemma 4.6, with high probability we do not have two disjoint in-components of diameter $> c_4(\log n)^{\frac{1}{d}}$ or two disjoint out-components of diameter $> c_4(\log n)^{\frac{1}{d}}$.

To deal with the possibility of both an in-component and an out-component of large

diameter, we prove the following lemma which is analogous to Lemma 4.6. The proof follows the same approach as the proof of Lemma 14 in [3].

Lemma 4.10. *There exists a constant c_4 (depending on d and k but not n) such that with high probability $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ does not contain an in-component and an out-component that are disjoint and both of diameter $\geq c_4(\log n)^{\frac{1}{d}}$.*

Proof. By Lemma 4.5, we may assume that any two points at distance $\leq c_1(\log n)^{\frac{1}{d}}$ are connected, and any two points at distance $\geq c_2(\log n)^{\frac{1}{d}}$ are not connected.

Let G_1 be an out-component and G_2 an in-component, both of diameter $\geq D = c_4(\log n)^{\frac{1}{d}}$. Unlike in the proof of Lemma 4.6, we do not necessarily have that edges of G_1 and G_2 are distance $> \frac{c_1(\log n)^{\frac{1}{d}}}{2}$ apart. However, using arguments from the proof of Lemma 4.7, we can say that a vertex x not in the out-component G_1 is distance at least $\frac{c_1(\log n)^{\frac{1}{d}}}{2}$ away from any edge wz of G_1.

As before, tile the cube $\gamma_{d,n}$ with tiles of side-length $s = \frac{c_1(\log n)^{\frac{1}{d}}}{4\sqrt{d}}$. Colour a tile red if it contains a vertex of G_1 or intersects an edge of G_1. Colour a tile blue if it contains a vertex of G_2 (unlike before, we do not colour blue tiles intersecting an edge of G_2). Colour a tile black if it is neither red or blue but contains a vertex, and colour a tile white if it contains no vertices. This is well-defined by our observation above, and furthermore, any tile touching a red tile must be either red or white.

Let $l = \frac{n^{\frac{1}{d}}}{s}$ and identify the tiling with the d-dimensional grid graph $[l]^d$, which we will call $\hat{G}$. Let R be the set of red tiles and B the set of blue tiles. Note that R forms a connected component. Since no points at distance $\geq c_2(\log n)^{\frac{1}{d}}$ are connected, there must be at least $\frac{D}{c_2(\log n)^{\frac{1}{d}}} = \frac{c_4}{c_2}$ red tiles and at least $\frac{c_4}{c_2}$ blue tiles.

The complement of R splits into components $C_1, C_2, \ldots, C_m$. Let $C_1, \ldots, C_p$ be all components that contain blue tiles, and let $V = \bigcup_{i=1}^{p} C_i$. Let $U = V^c$ and note that $U = R \cup \bigcup_{i=p+1}^{m} C_i$ is connected. Since all blue tiles are in V and all red tiles are in U, they must each contain $\geq \frac{c_4}{c_2}$ tiles.

Let ∂U be the set of tiles not in U but touching at least one tile in U. By our earlier observation all tiles in ∂U must be white. By the vertex isoperimetric inequality in the grid [5], $|\partial U|$ is bounded below by the size of the vertex boundary of a simplex of volume $\frac{c_4}{c_2}$. This is some constant c' depending on d and c_4 but independent of n.

Unlike in the proof of Lemma 4.6, we do not necessarily have that ∂U is connected, however we can show that it is connected in some power of the the grid graph $\hat{G} = [l]^d$.

Let $t = 2\frac{c_4(\log n)^{\frac{1}{d}}}{s} = \frac{8\sqrt{d}c_4}{c_1}$ so that the blue tiles are connected in $\hat{G}^{(t)}$. Suppose that ∂U is not connected in $\hat{G}^{(t)}$. Then we can partition ∂U into two non-empty sets A and B with no tile in A within distance t of any tile in B. For $1 \leq i \leq p$, let $\partial V_i = \partial U \cap C_i$. Since ∂V_i is connected in G, we must have A and B are unions of ∂V_is. Now, there must be a pair of $1 \leq i, j \leq p$ with $\partial V_i \subseteq A$, $\partial V_j \subseteq B$ and blue tiles $b_i \in V_i$, $b_j \in V_j$ at distance $< t$ apart. The shortest path from b_i to b_j passes through ∂V_i and ∂V_j and has length $< t$, giving that $d(\partial V_i, \partial V_j) < t$ which is a contradiction.

Thus ∂U is connected in $\hat{G}^{(t)}$. Recall that for any graph with maximum degree Δ, the number of connected subsets of size t containing a particular vertex is at most $(e\Delta)^t$. The maximum degree Δ of $\hat{G}^{(t)}$ is $\leq (2t)^d$ and so the number of connected components of $\hat{G}^{(t)}$ of size c' containing a particular tile is at most $\left(e(2t)^d\right)^{c'}$. Thus there are at most $l^d(2t)^{dc'}e^{c'}$ such components in total.

Putting all of this together, the probability that there is a connected set of c' empty tiles is

$$p = l^d(2t)^{dc'}e^{c'}e^{-c's^d} = \frac{\left(\frac{4\sqrt{d}}{c_1}\right)^d(2t)^{dc'}}{\log n}\exp\left(\log n + c' - c'\left(\frac{c_1}{4\sqrt{d}}\right)^d \log n\right)$$

Recall that c' is the size of the vertex boundary of a simplex of volume $\frac{4c_4}{c_1}$. Thus by choosing c_4 large enough, thus ensuring that c' is large enough, we can obtain $p = o(1)$. □

4.3.1 Proof of Theorem 4.9

We will prove two results, one to bound the probability of a small-diameter in-component and the other to bound the probability of a small-diameter out-component. These combined with Lemmas 4.6 and 4.10 give us Theorem 4.9.

Theorem 4.11. *If* $\frac{k}{\log n} > \frac{1}{\log d}$*, then as* $n \to \infty$ *the directed graph* $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ *contains no component of diameter* $< c_4(\log n)^{\frac{1}{d}}$ *with no adjacent out-edges.*

Theorem 4.12. *If* $\frac{k}{\log n} > \frac{2^d}{d}$*, then as* $n \to \infty$ *the directed graph* $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ *contains no component of diameter* $< c_4(\log n)^{\frac{1}{d}}$ *with no adjacent in-edges.*

Proof of Theorem 4.11. Suppose that C is an out-component of diameter $< c_4(\log n)^{\frac{1}{d}}$.

Tile the cube $\gamma_{d,n}$ as in the the proof of Theorem 4.4 and construct $\widehat{C}$, as well as the sets A' and $A_1, \ldots, A_{2d}$. Note that since there are no edges out of C, we have that $A_1, \ldots, A_{2d}$ are empty and A' contains at least k points.

Let f be the number of faces of the cube $\gamma_{d,n}$ that are within $c_3(\log n)^{\frac{1}{d}}$ of C, and without loss of generality suppose that $A_1, \ldots, A_{2d-f}$ do not intersect the boundary, so we have $(2d-f)|A'| \leq |\bigcup_{i=1}^{2d-f} A_i|$.

$$\begin{aligned}\mathbb{P}\left(\#A' \geq k, \# \bigcup_{i=1}^{2d-f} A_i = 0\right) &= \left(\frac{|A'|}{|A'| + \left|\bigcup_{i=1}^{2d-f} A_i\right|}\right)^k \\ &\leq \left(\frac{1}{1+2d-f}\right)^k\end{aligned}$$

If $\frac{k}{\log n} > \frac{1}{\log d+1}$ we have

$$\mathbb{P}\left(\#A' \geq k, \# \bigcup_{i=1}^{2d-f} A_i = 0\right) \leq n^{-\frac{2\log(2d-f+1)}{\log d}} = o\left(n^{-1}\right)$$

Now just as in the undirected case, the number of possible connected bounded-size subgraphs $\widehat{C}$ is $O(n)$, and given such a $\widehat{C}$ there are $O\left((\log n)^{4d}\right)$ ways of choosing the $4d$ points that define $A', A_1, \ldots, A_{2d}$. The probability that these define an out-component is certainly bounded by $\mathbb{P}\left(\#A' \geq k, \#\bigcup_{i=1}^{2d-f} A_i = 0\right)$. Putting this all together, the probability that there is a small-diameter out-components in G is $o\left(n(\log n)^{4d}n^{-1}\right) = o(1)$ when $\frac{k}{\log n} > \frac{1}{\log(d+1)}$. □

Proof of Theorem 4.12. Suppose that C is an in-component of diameter $< c_4(\log n)^{\frac{1}{d}}$. Tile the cube $\gamma_{d,n}$ as in the the proof of Theorem 4.4 and construct $\widehat{C}$. Let $P \in C$ and $Q \notin C$ be such that the distance between them $r_0 = d(P, Q)$ is minimised, and let $r = r_0 - s\sqrt{d}$. Construct the set B as before and note that since there are no edges into C the k nearest neighbours of Q must be in B.

Let $x^d = \frac{r^d V_d}{|\widehat{C}|}$. As before, we have $|B| < r^d V_d = x^d |\widehat{C}|$.

Let f be the number of $(d-1)$-dimensional faces of the cube $\gamma_{d,n}$ that are within distance $c_3(\log n)^{\frac{1}{d}}$ of our component.

We have that $|\widehat{C}_{(r)} \setminus \widehat{C} \cup \gamma_{d,n}| \geq 2^{-f}|\widehat{C}_{(r)} \setminus \widehat{C}| \geq 2^{-f}\left((1+x)^d - 1\right)$, using the same arguments as before.

Applying these bounds on $|B|$ and $|\widehat{C}_{(r)} \setminus \widehat{C}|$ we can bound the following probability:

$$\begin{aligned}
\mathbb{P}\left(\#B \geq k, \#\left(\widehat{C}_{(r)} \setminus \widehat{C} \cap \gamma_{d,n}\right)\right) &= \left(\frac{\left|B \setminus \left(\widehat{C}_{(r)} \setminus \widehat{C} \cap \gamma_{d,n}\right)\right|}{\left|B \cup \left(\widehat{C}_{(r)} \setminus \widehat{C} \cap \gamma_{d,n}\right)\right|}\right)^k \\
&\leq \left(\frac{|B|}{|B| + \left|\widehat{C}_{(r)} \setminus \widehat{C} \cap \gamma_{d,n}\right|}\right)^k \\
&\leq \left(\frac{x^d}{x^d + 2^{-f}\left((1+x)^d - 1\right)}\right)^k \\
&\leq \left(\frac{1}{1 + 2^{-f}}\right)^k
\end{aligned}$$

As in the undirected case, with high probability the number of possible connected bounded-size components $\widehat{C}$ within distance $c_3(\log n)^{\frac{1}{d}}$ of f faces of $\gamma_{d,n}$ is $O\left(n^{\frac{d-f}{d}+o(1)}\right)$.

There are $O\left((\log n)^2\right)$ ways of choosing the points P and Q that define B and $\widehat{C}$, and so putting everything together we have that the probability that there is an in-component of diameter $< c_4(\log n)^{\frac{1}{d}}$ is

$$O\left(\left(\frac{1}{1+2^{-f}}\right)^k n^{\frac{d-f}{d}+o(1)}\right).$$

In order for this to be $o(1)$ we need that $k > \frac{d-f}{d\log(1+2^{-f})}\log n$ for all $0 \leq f \leq d$. Using the inequality $\log x \geq 1 - \frac{1}{x}$, we obtain that $\frac{d-f}{d\log(1+2^{-f})} \leq \frac{(2^f+1)(d-f)}{d} \leq \frac{2^d}{d}$, so it suffices to take $\frac{k}{\log n} > \frac{2^d}{d}$.

□

4.3.2 An Upper Bound for the Directed Graph on a Torus

Suppose that we instead construct a k-nearest neighbour graph on a torus. Let the torus $T_{d,n}$ be formed from the cube $\gamma_{d,n}$ with opposite faces identified and let $\mathcal{P}_{tor}$ be a Poisson process of density 1 on $T_{d,n}$. Let $\overrightarrow{G}_{tor} = \overrightarrow{G}_{tor}(d,n,k)$ have vertices given by $\mathcal{P}_{tor}$ and each vertex joined by a directed edge out to its k nearest neighbours under the Euclidean metric.

The directed graph $\overrightarrow{G}_{tor}$ is like the directed graph $\overrightarrow{G}$ on the cube $\gamma_{d,n}$ but with any restrictions arising due to the boundary of the cube removed. In particular, we can obtain a better upper bound on the connectivity threshold for $\overrightarrow{G}_{tor}$ than for $\overrightarrow{G}$.

Theorem 4.13. *If $\frac{k}{\log n} > 1.443$, then as $n \to \infty$ the directed graph $\overrightarrow{G}_{tor} = \overrightarrow{G}_{tor}(d,n,k)$ is connected with high probability.*

Lemmas 4.5, 4.6 and 4.10 and Theorem 4.11 still hold for $\overrightarrow{G}_{tor}$ as for $\overrightarrow{G}$. If we can improve the bound on the threshold for a large in-component we will be done.

Theorem 4.14. *If* $\frac{k}{\log n} > 1.443$*, then as* $n \to \infty$ *the directed graph* $\overrightarrow{G}_{tor} = \overrightarrow{G}_{tor}(d, n, k)$ *contains no component of diameter* $< c_4(\log n)^{\frac{1}{d}}$ *with no adjacent in-edges.*

Proof. The proof is identical to that of Theorem 4.12, except that there are no boundary faces to consider so we can substitute in $f = 0$ throughout. This gives that the probability that there is an in-component of diameter $< c_4(\log n)^{\frac{1}{d}}$ is

$$O\left(\left(\frac{1}{1+2^0}\right)^k n^{\frac{d}{d}+o(1)}\right) = O\left(2^{-k} n^{1+o(1)}\right).$$

If $\frac{k}{\log n} > 1.443 > \frac{1}{\log 2}$ then this probabiltiy is $o(1)$ and there are with high probability no small diameter in-components. □

It seems likely that the existence of the boundary in the cube compared to the torus really does change the threshold for the existence of a small diameter in-component. Suppose there is a point x of $\overrightarrow{G}$ close to several faces the cube $\gamma_{d,n}$. A ball around the point has will be sliced by the boundary faces and have much reduced volume, and so it is much more likely that it contains no points nearby. In particular, there is a high probability that no points will have x as one of their nearest neighbours, making x an in-component of its own. This is not an obstacle in the undirected case or for out-components as the point x will always send out edges to at least k other points.

4.4 A Lower Bound for the Undirected Graph

Theorem 4.15. *If* $\frac{k}{\log n} < \frac{0.102}{d \log d}$*, then as* $n \to \infty$ *the undirected graph* $G(d, n, k)$ *is disconnected with high probability.*

The idea of the proof is to define a certain density distribution on a ball D such that

- if a region had this density distribution it would necessarily contain a small connected component,
- the probability that a Poisson process on D gives the specified density distribution

is 'large', and

- we can fit many balls of size $|D|$ into the cube $\gamma_{d,n}$.

The latter two points will give that the expected number of copies of D with the specified density distribution is at least 1. By the first point this copy of D will thus contain a small connected component.

Proof. Let D be a ball of radius $(2-\alpha)r_0$ for some α and r_0 around a point (called the origin). We will name some of the regions of D. Let X be the ball of radius αr_0 about the origin; let Y be the annulus from radius αr_0 to r_0; let Z_1 be the annulus from radius r_0 to $(1+\delta)r_0$; and let Z_2 by the annulus from radius $(1+\delta)r_0$ to $(2-\alpha)r_0$. See Figure 4.4 for a diagram of the 2-dimensional case.

We will specify that Y is empty. We want X to contain an isolated component, which requires two things. Firstly, that X contains $\geq k+1$ points, so that every point in X has its k nearest neighbours in X. Secondly, that for any vertex x at radius $r > r_0$, the number of vertices in $B_{(r-\alpha r_0)}(x)$ is at least $k+1$: this ensures that the k nearest neighbours of x lie outside of X.

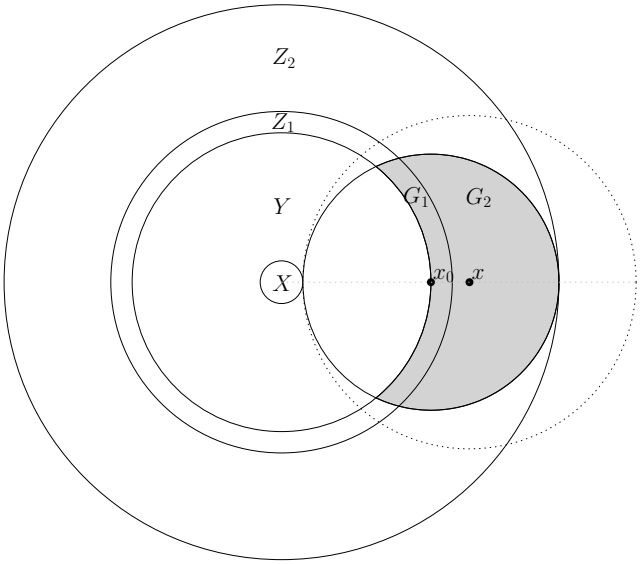


Figure 4.4: The regions in the 2-dimensional case

Note that $B_{(r-\alpha r_0)}(x) \supset B_{(r_0-\alpha r_0)}(x_0)$ where x_0 is on the same line from the origin as x, but at radius $r_0 < r$. Thus it is sufficient to stipulate that $B_{(r_0-\alpha r_0)}(x_0)$ contains $k+1$ vertices for all x_0 at radius r_0: as such, fix a particular such x_0 and define $G = B_{(r_0-\alpha r_0)}(x_0) \setminus (X \cup Y)$. Let $G_1 = G \cap Z_1$ and $G_2 = G \cap Z_2$ so that $G = G_1 \cup G_2$.

Now we will specify the density function ρ, as a function of the distance r from the origin.

$$\rho(r) = \begin{cases} \dfrac{\rho_1|G_1| + |G_2|}{|X|} & \text{if} \quad 0 \le r \le \alpha r_0 \quad (X) \\ 0 & \text{if} \quad \alpha r_0 < r \le r_0 \quad (Y) \\ \rho_1 & \text{if} \quad r_0 < r \le (1+\delta)\, r_0 \quad (Z_1) \\ 1 & \text{if} \quad (1+\delta)\, r_0 < r \le (2-\alpha)\, r_0 \quad (Z_2) \end{cases}$$

where

$$\rho_1 = \frac{|D| - |G_2| - |Z_2|}{|G_1| + |Z_1|}$$

is independent of r_0.

By the choice of ρ_1, we have that the total density is 1:

$$\int_D \rho \; dV = (\rho_1|G_1| + |G_2|) + \rho_1|Z_1| + |Z_2| = |D|$$

We also have that

$$\int_X \rho \; dV = \rho_1|G_1| + |G_2| = \int_G \rho \; dV.$$

and so by adjusting r_0 we can ensure that, as was required, we have

$$k+1 = \int_X \rho \; dV = \int_G \rho \; dV.$$

To approximate the probability of the density being close to this, we will make use of the following lemma which can be found in [3] (stated for the 2-dimensional case, but the dimension is actually irrelevant to the proof).

Lemma 4.16. *Let $A_1, \dots A_t$ be disjoint volumes, and $\rho_1, \dots, \rho_t \geq 0$ be real numbers such that $\rho_i |A_i| \in \mathbb{Z}$. Then the probability that a Poisson process with intensity 1 has precisely ρ_i points in each region A_i is*

$$\exp\left(\sum_{i=1}^{t}(\rho_i - 1 - \rho_i \log \rho_i)|A_i| + O\left(t \log_+ \sum \rho_i |A_i|\right)\right)$$

with the convention that $0 \log 0 = 0$ and $\log_+ x = \max(\log x, 1)$.

Note that the exact probability is easy to write down and Lemma 4.16 just gives us a simple usable approximation.

If we divide $Z_1 \cup Z_2$ into very small tiles A_i, we can apply this lemma to the regions X, Y and the A_i, and the densities they should have according to the function ρ. Note that the number t of tiles used may depend on d, ρ_1, δ and α but does not depend on n or k. Let p be the probability that we are very close to the specified density, in the sense that each of the regions X, Y and the A_i contain the same number of points that they would under the density ρ. We have that p satisfies

$$\begin{aligned} -\log p &= \int_D \rho \log \rho \; dV - \int_D \rho \; dV + \int_D \; dV + O(t \log \int_D \rho \; dV) \\ &= \int_D \rho \log \rho \; dV + O(\log |D|) \\ &= \frac{(k+1)}{\int_X \rho \; dV} \int_D \rho \log \rho \; dV + O(\log |D|) \end{aligned}$$

and so we want to bound

$$\frac{\int_D \rho \log \rho \; dV}{\int_X \rho \; dV} = \log\left(\frac{\rho_1 |G_1| + |G_2|}{|X|}\right) + \log(\rho_1) \frac{\rho_1 |Z_1|}{\rho_1 |G_1| + |G_2|}.$$

Since $|D| > |Z_1| + |Z_2| + |G_1| + |G_2|$, we have that

$$\rho_1 = \frac{|D| - |G_2| - |Z_2|}{|G_1| + |Z_1|} > 1.$$

We also have that

$$\rho_1 = \frac{|D| - |G_2| - |Z_2|}{|G_1| + |Z_1|} \leq \frac{|D| - |Z_2|}{|Z_1|} = \frac{(1+\delta)^d}{(1+\delta)^d - 1} \leq 1 + \frac{1}{\delta}.$$

Now we can calculate

$$\begin{aligned}\frac{\rho_1|Z_1|}{\rho_1|G_1| + |G_2|} \leq \frac{\rho_1|Z_1|}{|G|} &\leq \frac{|D| - |Z_2|}{|G|} \\ &\leq \frac{2(1+\delta)^d}{(1-\alpha)^d} \qquad \text{since } G \text{ is at least half of } B_{(1-\alpha)r_0}(x)\end{aligned}$$

and

$$\begin{aligned}\frac{\rho_1|G_1| + |G_2|}{|X|} &\leq \frac{\rho_1|G|}{|X|} \\ &\leq \left(1 + \frac{1}{\delta}\right)\frac{(1-\alpha)^d}{\alpha^d} \qquad \text{since } G \text{ is smaller than } B_{(1-\alpha)r_0}(x).\end{aligned}$$

Thus

$$\frac{\int_D \rho \log \rho \, dV}{\int_X \rho \, dV} \leq \log\left(\left(1 + \frac{1}{\delta}\right)\left(\frac{1-\alpha}{\alpha}\right)^d\right) + 2\left(\frac{1+\delta}{1-\alpha}\right)^d \log\left(1 + \frac{1}{\delta}\right)$$

Set $\delta = \frac{1}{2d}$ and $\alpha = \frac{1}{2(d+1)}$ so that $\left(\frac{1+\delta}{1-\alpha}\right)^d = \left(1 + \frac{1}{d}\right)^d \leq e$. We get

$$\begin{aligned}\frac{\int_D \rho \log \rho \, dV}{\int_X \rho \, dV} &\leq \log\Big((2d+1)(2d+1)^d\Big) + 2e \log(2d+1) \\ &= (d + 1 + 2e)\log(2d+1) \\ &\leq \frac{(3+2e)\log 5}{2\log 2}(d \log d) \qquad \text{using that } d \geq 2.\end{aligned}$$

Letting $c = \frac{(3+2e)\log 5}{2\log 2}$, we have

$$-\log p \leq (k+1)cd\log d + O(\log |D|).$$

Note that

$$k+1 = \rho_1|G_1| + |G_2| > |G| > \frac{1}{2}\left(\frac{1-\alpha}{2-\alpha}\right)^d |D| \\ = \frac{1}{2}\left(\frac{2d-1}{4d-1}\right)^d |D| \geq \frac{1}{2}\left(\frac{3}{7}\right)^d |D|.$$

and so $\log|D| \leq \log k + 1 - d\log(7/3) + \log 2 = O(\log(k+1))$. We get that

$$p \geq \exp\left(-(k+1)cd\log d + O(\log(k+1))\right).$$

Now, we need to bound the number of expected number of copies of D. The number of balls of radius $(2-\alpha)r_0$ that fit inside the cube $\gamma_{d,n}$ is

$$\geq \left(\left\lfloor \frac{n^{\frac{1}{d}}}{(2-\alpha)r_0} \right\rfloor\right)^d = \Theta\left(\frac{n}{|D|}\right) = \Theta\left(\frac{n}{k+1}\right).$$

In particular, the expected number of copies of D is $\Theta\left(\frac{np}{k+1}\right)$ and so we can say that the graph is disconnected with high probability if

$$\lim_{n\to\infty} \frac{np}{k+1} = \frac{n\exp\left(-(k+1)cd\log d + O(\log(k+1))\right)}{k+1} > 1.$$

If

$$\frac{k}{\log n} < \frac{0.102}{d\log d} < \frac{1}{cd\log d}$$

then the graph is disconnected with high probability.

□

4.5 A Lower Bound for the Directed Graph

We have the following two theorems on the existence of small diameter in-components and small diameter out-components respectively.

Theorem 4.17. *If* $\frac{k}{\log n} < \frac{0.721}{d}$ *then with high probability the directed graph* $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ *contains an out-component of diameter* $< k^{\frac{1}{d}}$.

Theorem 4.18. *If* $\frac{k}{\log n} < \frac{0.079}{\log d}$ *then with high probability the directed graph* $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ *contains an in-component of diameter* $< k^{\frac{1}{d}}$.

The general theorem follows immediately from the second of these.

Theorem 4.19. *If* $\frac{k}{\log n} < \frac{0.079}{\log d}$ *then with high probability the directed graph* $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ *is not strongly connected.*

Similarly to the proof of Theorem 4.15 for the undirected graph, to prove each of Theorems 4.18 and 4.17 we will define a density distribution on a ball that guarantees an out-component or in-component respectively and then bound the probability of such a density distribution occuring.

Proof of Theorem 4.17. Let D be a ball of radius r_0 about a point (which we call the origin). Let X be the ball of radius $\frac{r_0}{4}$ about the origin and let Y be the annulus from $\frac{r_0}{4}$ to r_0. Specify a density function ρ as a function of the distance r from the origin.

$$\rho(r) = \begin{cases} 4^d & \text{if} \quad 0 \leq r \leq \dfrac{r_0}{4} \quad (X) \\ 0 & \text{if} \quad \dfrac{r_0}{4} < r \leq r_0 \quad (Y) \end{cases}$$

Note that the distance between two points in X is closer than the distance between a point in X and a point outside D. By adjusting r_0, we can ensure that we have

$$k + 1 = r_0^d V_d = |D| = \int_D \rho \, dV.$$

This means that the k nearest neighbours of any point in X are also in X, and so any region with this density must be an out-component.

Let p be the probability that region X contains $k + 1$ points and region Y is empty.

Then by Lemma 4.16 we have that p satisfies

$$\begin{aligned}
-\log p &= \int_D \rho \log \rho - \rho + 1 \; dV + O(2 \log \int_D \rho \; dV) \\
&= \left(4^d \log 4^d - 4^d + 1\right) \frac{r_0^d V_d}{4^d} + \left(1 - \frac{1}{4^d}\right) r_0^d V_d + O(2 \log (k+1)) \\
&= (k+1) d \log 4 + O(2 \log (k+1))
\end{aligned}$$

We bound the number of expected number of copies of D as before. The number of balls of radius r_0 that fit inside the cube $\gamma_{d,n}$ is

$$\geq \left(\left\lfloor \frac{n^{\frac{1}{d}}}{r_0} \right\rfloor \right)^d = \Theta\left(\frac{n}{(r_0)^d}\right) = \Theta\left(\frac{n}{k+1}\right).$$

Then the expected number of copies of D that exist is $\Theta(\frac{np}{k+1})$. Thus we can say that with high probability the graph contains an out-component of diameter $\frac{r_0}{2} < k^{\frac{1}{d}}$ if

$$\lim_{n\to\infty} \frac{np}{k+1} = \lim_{n\to\infty} \frac{ne^{-(k+1)d\log 4 + O(2\log k+1)}}{k+1} > 1.$$

which certainly holds when $\frac{k}{\log n} < \frac{0.721}{d} < \frac{1}{d \log 4}$. □

Proof of Theorem 4.18. Let D be a ball of radius $(2-\alpha)r_0$ about a point (which we call the origin), and define the regions X, Y, Z_1, Z_2, G, G_1 and G_2 as in the proof of Theorem 4.15 (see Figure 4.4).

We specify a density function ρ on D as a function of the distance r from the origin.

$$\rho(r) = \begin{cases}
\rho_1 & \text{if} \quad 0 \leq r \leq \alpha r_0 & (X) \\
0 & \text{if} \quad \alpha r_0 < r \leq r_0 & (Y) \\
\rho_2 & \text{if} \quad r_0 < r \leq (1+\delta) r_0 & (Z_1) \\
1 & \text{if} \quad (1+\delta) r_0 < r \leq (2-\alpha) r_0 & (Z_2)
\end{cases}$$

We need that $\int_D \rho \, dV = |D|$, so that the total density is correct; that $\int_X \rho \, dV = 1$, so that the component X is non-empty; and that $\int_G \rho \, dV = k+1$, so that there are no edges in to X. This last follows from the fact that for any point x at distance $r > r_0$ from the origin, the ball of radius $r - \alpha r_0$ around x contains a region G, as demonstrated in Figure 4.4.

Take ρ_1 and ρ_2 satisfying the simultaneous equations $(k+1)\rho_1|X| = \rho_2|G_1| + |G_2|$ and $\rho_1|X| + \rho_2|Z_1| = |D| - |Z_2|$. In particular,

$$\rho_2 = \frac{(k+1)\left(|D| - |Z_2|\right) - |G_2|}{(k+1)|Z_1| + |G_1|} \quad \text{and} \quad \rho_1 = \frac{\rho_2|G_1| + |G_2|}{(k+1)|X|}.$$

This gives that $\int_D \rho \, dV = \rho_1|X| + \rho_2|Z_1| + |Z_2| = |D|$. By scaling r_0 we can ensure that $\int_X \rho \, dV = \rho_1|X| = 1$ and $\int_G \rho \, dV = \rho_2|G_1 + |G_2| = k+1$.

We would like to apply Lemma 4.16. Divide $Z_1 \cup Z_2$ into small tiles $A_1, A_2, \ldots, A_t$ where the number t of tiles used may depend on $d, \rho_1, \rho_2, \delta$ and α but does not depend on n or k. Let p be the probability that that we are very close to the specified density on D, in the sense that the regions X, Y and $A_1, A_2, \ldots, A_t$ contain the same number of points that they would under the density ρ. We get that p satisfies the following equation:

$$\begin{aligned} -\log p &= \int_D \rho \log \rho \, dV - \int_D \rho \, dV + \int_D \, dV + O(t \log \int_D \rho \, dV) \\ &= \rho_1|X| log \rho_1 + \rho_2|Z_1| \log \rho_2 + O(\log |D|) \\ &= \log \rho_1 + (k+1)\frac{\rho_2|Z_1|}{\rho_2|G_1| + |G_2|} \log \rho_2 + O(\log |D|). \end{aligned}$$

Let us bound some of these quantities. We have that $|D| - |Z_1| - |Z_2| = |X| + |Y| > |G|$ and so

$$\rho_2 = \frac{(k+1)\left(|D| - |Z_2|\right) - |G_2|}{(k+1)|Z_1| + |G_1|} > \frac{|D| - |Z_2| - |G_2|}{|Z_1| + |G_1|} > 1.$$

We also have that

$$\rho_2 < \frac{|D| - |Z_2|}{|Z_1|} = \frac{(1+\delta)^d}{(1+\delta)^d - 1} = 1 + \frac{1}{(1+\delta)^d - 1} < 1 + \frac{1}{\delta}.$$

Using these, we can calculate

$$\frac{\rho_2|Z_1|}{\rho_2|G_1| + |G_2|} < \frac{\rho_2|Z_1|}{|G|} \leq \frac{|D| - |Z_2|}{|G|} < \frac{2(1+\delta)^d}{(1-\alpha)^d}.$$

Finally, we can bound

$$\rho_1 = \frac{\rho_2|G_1| + |G_2|}{(k+1)|X|} < \frac{\left(1 + \frac{1}{\delta}\right)|G|}{(k+1)|X|} < \frac{\left(1 + \frac{1}{\delta}\right)(1-\alpha)^d}{(k+1)\alpha^d}.$$

Putting these together, we get that

$$-\log p < \log\left(\frac{\left(1 + \frac{1}{\delta}\right)(1-\alpha)^d}{(k+1)\alpha^d}\right) + (k+1)\frac{2(1+\delta)^d}{(1-\alpha)^d}\log\left(1 + \frac{1}{\delta}\right) + O(\log|D|).$$

Set $\delta = \frac{1}{2d}$ and $\alpha = \frac{1}{2d+2}$ so that $\left(\frac{1+\delta}{1-\alpha}\right)^d = \left(1 + \frac{1}{d}\right)^d < e$. We get

$$\begin{aligned} -\log p &< \log\left((2d+1)^{d+1}\right) - \log(k+1) + (k+1)2e\log(2d+1) + O(\log|D|) \\ &= (k+1)2e\log(2d+1) - \log(k+1) + O(\log|D|). \end{aligned}$$

Note that

$$k + 1 = \rho_1|G_1| + |G_2| \geq |G| \geq \frac{((1-\alpha)r_0)^d V_d}{2} \geq \frac{\left(\frac{(2-\alpha)}{3}r_0\right)^d V_d}{2} = \frac{|D|}{2 \cdot 3^d}.$$

In particular, $\log|D| = \log(k+1) + d\log 3 + \log 2$, which gives

$$\begin{aligned} -\log p &= (k+1)2e\log(2d+1) - \log(k+1) + O(\log(k+1)) \\ &= (k+1)2e\log(2d+1) + O(\log(k+1)). \end{aligned}$$

The number of balls of radius $(2-\alpha)r_0$ we can fit in the cube $\gamma_{d,n}$ is at least

$$\left\lfloor \frac{n^{\frac{1}{d}}}{2(2-\alpha)r_0} \right\rfloor^d = \left\lfloor \frac{n^{\frac{1}{d}}}{6\left(\frac{2(k+1)}{V_d}\right)^{\frac{1}{d}}} \right\rfloor^d = \Theta\left(\frac{n}{k+1}\right).$$

Then the expected number of copies of D that exist is $\Theta(\frac{np}{k+1})$. We can say that with high probability the graph contains an in-component of diameter $(2-\alpha)r_0$ if

$$\lim_{n\to\infty} \frac{np}{k+1} = \lim_{n\to\infty} \frac{n\exp\left[-(k+1)2e\log\left(2d+1\right)+O(\log\left(k+1\right))\right]}{k+1} > 1.$$

In particular, the graph contains an in-component of diameter $(2-\alpha)r_0$ with high probability if

$$\frac{k}{\log n} < \frac{0.079}{\log d} < \frac{\log 2}{2e\log 5\log d} < \frac{1}{2e\log 2d+1}.$$

□

The above proof of the lower bound on the connectivity threshold for $\overrightarrow{G}$ does not make use of the boundary of the cube $\gamma_{d,n}$. This means that it also gives a lower bound on the connectivity threshold for the directed graph on a torus $\overrightarrow{G}_{tor}$. We know from Theorem 4.13 that the connectivity threshold for $\overrightarrow{G}_{tor}$ is bounded above by a constant times $\log n$ and so without considering the boundary of the cube we cannot possibly improve this lower bound by more than a $\log d$ factor.

As discussed in Section 4.3.2, it seems likely that the threshold for connectivity for the $\overrightarrow{G}$ and the directed graph on a torus $\overrightarrow{G}_{tor}$ are different and the crucial difference lies in the existence of small diameter in-components. Instead of the proof of Theorem 4.18 based on a density distribution on a ball one could try looking at some density distribution defined on some region lying on the boundary of the cube $\gamma_{d,n}$.

Unfortunately it doesn't seem possible to improve the bound by just looking instead at a half-ball or quarter-ball and so on at the boundary of the cube and keeping the

density distribution uniform in shells. Suppose x lies on the intersection of f faces of $\gamma_{d,n}$. Suppose we want to define a density distribution in shells on a $\frac{1}{2^f}$-ball D' around x (that is, D' is a segment of a ball bisected by f orthogonal planes). Consider a point y lying on the same f faces as x at distance r from x. If we want x to be in a small diameter in-component not containing y then we need that a $\frac{1}{2^f}$-ball G' of radius r around y contains at least k points. In particular, this tells us that the density distribution in the shells of D' woud have to be equivalent to those in the shells of the ball D in the above proof of Theorem 4.18. These have the same probability of occuring.

One could potentially get around this either by considering a more complicated shape that 'flattens out' near the boundary rather than a segment of a ball, or by allowing the density to increase nearer to the boundary. The trade-off is that it becomes much more difficult to bound the probability of these more complicated densities occuring.

4.6 Open Questions

The bounds we've obtained for the threshold for connectivity for the graph $G = G(d, n, k)$ do not match, differing by a $\log d$ factor. We conjecture that the true threshold is $\Theta\left(\frac{\log n}{d}\right)$: it seems likely that the existence of small in-components in the directed setting is the barrier for existence of small components in the undirected setting. If this is the case then the threshold for no small component and the threshold for no small in-component would both have to have $k = \Theta\left(\frac{\log n}{d}\right)$.

Conjecture 4.1. *There exist constants c_1 and c_2 such that for all $d \geq 2$ we have*

- *if $k > \frac{c_1 \log n}{d}$ then the graph $G(d, n, k)$ is connected with high probability and*
- *if $k < \frac{c_2 \log n}{d}$ then the graph $G(d, n, k)$ is disconnected with high probability.*

For the directed graph $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ we have a much greater discrepancy in the upper and lower bounds on the connectivity threshold, with a difference of an exponential factor. However, we have that if you ignore boundary effects by considering the graph on

a torus we get a much improved upper bound with only a $\log d$ gap between the bounds. The question of what happens on the torus is is an interesting one in its own right.

Question 4.2. *Let $\overrightarrow{G}_{tor}(d, n, k)$ be the directed k-nearest neighbour graph defined on a d-dimensional torus (rather than a cube). What is the threshold on k, in terms of d and n, for $\overrightarrow{G}_{tor}(d, n, k)$ to be connected?*

Theorem 4.3 gives upper and lower bounds on the answer to this question.

Let us return to $\overrightarrow{G}(d, n, k)$ defined on the cube $\gamma_{d,n}$. Considering the boundary effects of the cube increases the upper bound on the connectivity threshold from constant to exponential. This is perhaps surprising as in the undirected graph $\overrightarrow{G} = \overrightarrow{G}(d, n, k)$ considering the boundary did not fundamentally change the upper bound, except to increase the constant.

It seems that the boundary is crucial in the directed case because connectivity is driven by the existence of small in-components. A vertex x lying near the boundary of the cube could form a small out-component of its own if there are no points nearby which have x as one of their k-nearest neighbours.

The lower bound on the connectivity threshold calculated in Section 4.5 ignores any boundary effects. As discussed after the proof, we believe that it should be possible to increase the lower bound using by defining an appropriate density function for a region on the boundary of the cube $\gamma_{d,n}$ and finding its probability of occuring. We conjecture that this will give an exponential bound to match the upper bound.

Conjecture 4.3. *There exist constants c_1 and c_2 such that for all $d \geq 2$ we have*

- *if $k > c_1^d \log n$ then the graph $\overrightarrow{G}(d, n, k)$ is connected with high probability, and*
- *if $k < c_2^d \log n$ then the graph $\overrightarrow{G}(d, n, k)$ is disconnected with high probability.*

References

[1] B. A. Anderson. Finite topologies and Hamiltonian paths. *J. Combinatorial Theory Ser. B*, 14:87–93, 1973.

[2] Dan Archdeacon. Problems in topological graph theory. `http://www.cems.uvm.edu/TopologicalGraphTheoryProblems/perfectq.htm`, 1995. [Online; accessed 27-July-2018].

[3] Paul Balister, Béla Bollobás, Amites Sarkar, and Mark Walters. Connectivity of random k-nearest-neighbour graphs. *Adv. in Appl. Probab.*, 37(1):1–24, 2005.

[4] B. Bollobás. On generalized graphs. *Acta Math. Acad. Sci. Hungar.*, 16:447–452, 1965.

[5] Béla Bollobás and Imre Leader. Edge-isoperimetric inequalities in the grid. *Combinatorica*, 11(4):299–314, 1991.

[6] Darryn Bryant, Barbara M. Maenhaut, and Ian M. Wanless. A family of perfect factorisations of complete bipartite graphs. *J. Combin. Theory Ser. A*, 98(2):328–342, 2002.

[7] Debsoumya Chakraborti and Po-Shen Loh. Minimizing the numbers of cliques and cycles of fixed size in an f-saturated graph. `https://arxiv.org/abs/1907.01603`, 2019. (preprint).

[8] Vaithiyalingam Chitra and Appu Muthusamy. A note on semi-perfect 1-factorization and Craft's conjecture. *Graph Theory Notes N. Y.*, 64:58–62, 2013.

[9] Louis Dubuc. Sur les automates circulaires et la conjecture de černỳ. *RAIRO-Theoretical Informatics and Applications*, 32(1-3):21–34, 1998.

[10] David Eppstein. Reset sequences for monotonic automata. *SIAM Journal on Computing*, 19(3):500–510, 1990.

[11] Paul Erdos and Arthur H Stone. On the structure of linear graphs. *Bull. Amer. Math. Soc*, 52(1087-1091):1, 1946.

[12] Jill R Faudree, Ralph J Faudree, and John R Schmitt. A survey of minimum saturated graphs. *The Electronic Journal of Combinatorics*, 1000:DS19–Jul, 2011.

[13] Peter Frankl. An extremal problem for two families of sets. *European Journal of*

Combinatorics, 3(2):125–127, 1982.

[14] Vasil S. Gochev and Ivan S. Gotchev. On k-semiperfect 1-factorizations of Q_n and Craft's conjecture. *Graph Theory Notes N. Y.*, 58:36–41, 2010.

[15] François Gonze and Raphaël M. Jungers. On the synchronizing probability function and the triple rendezvous time for synchronizing automata. *SIAM J. Discrete Math.*, 30(2):995–1014, 2016.

[16] Jarkko Kari. Synchronizing finite automata on eulerian digraphs. *Theoretical Computer Science*, 295(1-3):223–232, 2003.

[17] L. Kászonyi and Zs. Tuza. Saturated graphs with minimal number of edges. *J. Graph Theory*, 10(2):203–210, 1986.

[18] A. Kotzig. Hamilton graphs and Hamilton circuits. In *Theory of Graphs and its Applications (Proc. Sympos. Smolenice, 1963)*, pages 63–82. Publ. House Czechoslovak Acad. Sci., Prague, 1964.

[19] Rastislav Královič and Richard Královič. On semi-perfect 1-factorizations. In *Structural information and communication complexity*, volume 3499 of *Lecture Notes in Comput. Sci.*, pages 216–230. Springer, Berlin, 2005.

[20] P. J. Laufer. On strongly Hamiltonian complete bipartite graphs. *Ars Combin.*, 9:43–46, 1980.

[21] Edouard Lucas. *Les jeux de demoiselles*. 1883.

[22] G. Nakamura. Dudeney's round table problem for the cases of $n = p+1$ and $n = 2p$. *Sugaku Sem.*, 159:24–29, 1975. [in Japanese].

[23] David A Pike. A perfect one-factorisation of k_{56}. `https://arxiv.org/abs/1810.08734`, 2018. (preprint).

[24] Oleg Pikhurko. The minimum size of saturated hypergraphs. *Combinatorics, Probability and Computing*, 8(5):483–492, 1999.

[25] Oleg Pikhurko. Results and open problems on minimum saturated hypergraphs. *Ars Combin.*, 72:111–127, 2004.

[26] Jean-Eric Pin. On two combinatorial problems arising from automata theory. In *North-Holland Mathematics Studies*, volume 75, pages 535–548. Elsevier, 1983.

[27] Yaroslav Shitov. An improvement to a recent upper bound for synchronizing words

of finite automata. *arXiv preprint arXiv:1901.06542*, 2019.

[28] Richard Stong. Hamilton decompositions of directed cubes and products. *Discrete Math.*, 306(18):2186–2204, 2006.

[29] Marek Szykuła. Improving the upper bound and the length of the shortest reset words. In *35th Symposium on Theoretical Aspects of Computer Science*, volume 96 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 56, 13. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

[30] Paul Turán. On an extremal problem in graph theory. *Mat. Fiz. Lapok*, 48:436–452, 1941. [In Hungarian].

[31] Zsolt Tuza. Extremal problems on saturated graphs and hypergraphs. *Ars Combin*, 25:105–113, 1988.

[32] Mikhail V. Volkov. Synchronizing automata and the Černý conjecture. In *Language and automata theory and applications*, volume 5196 of *Lecture Notes in Comput. Sci.*, pages 11–27. Springer, Berlin, 2008.

[33] Mark Walters. Small components in k-nearest neighbour graphs. *Discrete Appl. Math.*, 160(13-14):2037–2047, 2012.

[34] I. M. Wanless. Perfect factorisations of bipartite graphs and Latin squares without proper subrectangles. *Electron. J. Combin.*, 6:Research Paper 9, 16, 1999.

[35] Feng Xue and Panganamala R Kumar. The number of neighbors needed for connectivity of wireless networks. *Wireless networks*, 10(2):169–181, 2004.

Appendix A

Finding 1-Factorisations of the Hypercube by Computer

The Python program below can be used to find all 1-factorisations of a d-dimensional hypercube Q_d for small d. It can also be used find all k-semi-perfect 1-factorisations of Q_d for any k, and a 1-factorisation of Q_d where the union of any pair of 1-factors is at most two cycles. More information on how the program works and some example outputs can be found in 3.4. The code was written in Python 3.

```
1  import numpy
2
3  class hypercube(object):
4      def __init__(self, n):
5          """
6          Initialises a hypercube object
7          n: dimension of the hypercube.
8          vertices: 0, ..., 2^n - 1.
9          edges: a list of all the edges, as ordered pairs of vertices
10         edgesFromVertices: a dictionary indexed by vertices of all
       edges adjacent to each vertex
11         """
```

```
12          self.n = n
13          self.vertices = range(1<<n)
14          self.edges = []
15          self.edgesFromVertices = {}
16
17          """
18          Initialise these to be empty until we generate them
19          """
20          self.factorizations = []
21          self.semiPerfectFactorizations = {}
22          self.goodFactorizations = False
23
24          def genEdges(self):
25              """
26              Generate the edges of the cube as ordered pairs of
    vertices (v,u) with v < u
27              """
28              for v in self.vertices:
29                  bit = 1
30                  for i in range(self.n):
31                      u = v^bit
32                      if v < u:
33                          self.edges.append([v,u])
34                      bit*=2
35          genEdges(self)
36
37          def edgesFromVertex(self,v):
38              """
39              Returns all edges adjacent to vertex v
40              """
41              edges = []
42              bit = 1
43              for i in range(self.n):
```

```
44                  u = v^bit
45                  if v < u:
46                      edges.append([v,u])
47                  else:
48                      edges.append([u,v])
49                  bit = bit<<1
50              return edges
51
52          def genEdgesFromVertices(self):
53              """
54              Generate a list where the vth entry is all edges adjacent
        to vertex v
55              """
56              for v in self.vertices:
57                  self.edgesFromVertices[v] = edgesFromVertex(self,v)
58
59          genEdgesFromVertices(self)
60
61
62  def findDisjointEdges(self, currentEdge, edgeList):
63      """
64      Returns a list of all of the edges in edgeList that are not
       adjacent to currentEdge
65      """
66      newEdgeList = list(edgeList)
67      for edge in self.edgesFromVertices[currentEdge[0]]:
68          if edge in newEdgeList:
69              newEdgeList.remove(edge)
70      for edge in self.edgesFromVertices[currentEdge[1]]:
71          if edge in newEdgeList:
72              newEdgeList.remove(edge)
73      return newEdgeList
74
```

```
75  def genFactorizations(self, matchings, unusedEdges, currentMatching,
        potentialEdges):
76      """
77      Generates recursively all 1-factorizations, in the form of a list
        of perfect matchings.
78      Matchings are stored as dictionaries indexed by vertices, where
        the entry for a vertex is its neighbour under the matching.
79
80      matchings: a list of 1-factors already included in the 1-
        factorization
81      currentMatching: the matching currently being generated
82      unusedEdges: edges of the cube not used in any matchings so far
83      potentialEdges: edges that we haven't ruled out from being added
        to currentMatching
84      """
85      matchingLength = 2**(self.n-1)
86      if len(currentMatching) == 2*matchingLength: #Current matching is
        perfect
87          newMatchings = list(matchings)
88          newMatchings.append(currentMatching)
89          if len(unusedEdges) > matchingLength: #Still more 1-factors
        to find
90              genFactorizations(self, newMatchings, unusedEdges, {},
        list(unusedEdges))
91          else: #This is a 1-factorization with the remaining edges
        forming the final 1-factor
92              newMatchings.append(matchingFromEdgeList(unusedEdges))
93              self.factorizations.append(newMatchings)
94
95      elif not currentMatching: #Current matching is empty, want to
        start with edge (0,i)
96          edge = unusedEdges.pop(0)
```

```
97          potentialEdges = findDisjointEdges(self, edge, potentialEdges
    [1:])
98          newMatching = {edge[0]:edge[1], edge[1]:edge[0]}
99          genFactorizations(self, matchings, unusedEdges, newMatching,
    potentialEdges)
100
101     elif potentialEdges: #Current matching is neither perfect nor
    empty and there are potential edges to add
102         for i in range(len(potentialEdges)-(matchingLength-(len(
    currentMatching)//2))+1): #Try adding each potential edge in turn
103             edge = potentialEdges[i]
104             newPotentialEdges = findDisjointEdges(self, edge,
    potentialEdges[i+1:])
105             newUnusedEdges = list(unusedEdges)
106             newUnusedEdges.remove(edge)
107             newMatching = dict(currentMatching)
108             newMatching[edge[0]] = edge[1]
109             newMatching[edge[1]] = edge[0]
110             genFactorizations(self, matchings, newUnusedEdges,
    newMatching, newPotentialEdges)
111
112
113 def genSemiPerfectFactorizations(self, k, matchings, unusedEdges,
    currentMatching, potentialEdges, display):
114     """
115     Finds recursively all k-semi-perfect 1-factorizations.
116     Similar to genFactorizations, but throwing out partial 1-
    factorizations that are not k-semi-perfect.
117
118     matchings: a list of perfect matchings already included in the 1-
    factorization
119     currentMatching: the matching currently being generated
120     unusedEdges: edges of the cube not used in any matchings so far
```

```
121     potentialEdges: edges that we haven't ruled out from being added
        to currentMatching
122
123     display: if true, the matchings are printed as we go along.
124     """
125     matchingLength = 2**(self.n-1)
126     if len(currentMatching) == 2*matchingLength: #Current matching is
         perfect
127         newMatchings = list(matchings)
128         newMatchings.append(currentMatching)
129         if lastMatchingSemiPerfect(self, k, newMatchings): #Check if
         the partial 1-factorization is k-semi-perfect
130             if len(unusedEdges) > matchingLength: #Still more 1-
        factors to find
131                 genSemiPerfectFactorizations(self, k, newMatchings,
        unusedEdges, {}, list(unusedEdges), display)
132             else: #This is a 1-factorization with the remaining
        edges forming the final 1-factor
133                 newMatchings.append(matchingFromEdgeList(unusedEdges)
        )
134                 if lastMatchingSemiPerfect(self, k, newMatchings): #
        The partial 1-factorization is k-semi-perfect
135                     self.semiPerfectFactorizations[k].append(
        newMatchings)
136                     if display: #Print the matchings
137                         prettyPrintFactorization(self, newMatchings)
138     elif not currentMatching: #Current matching is empty, want to
        start with edge (0,i)
139         edge = unusedEdges.pop(0)
140         potentialEdges = findDisjointEdges(self, edge, potentialEdges
        [1:])
141         newMatching = {edge[0]:edge[1], edge[1]:edge[0]}
```

```
142             genSemiPerfectFactorizations(self, k, matchings, unusedEdges,
        newMatching, potentialEdges, display)
143
144         elif potentialEdges: #Current matching is neither perfect nor
        empty and there are potential edges to add
145             for i in range(len(potentialEdges)-(matchingLength-(len(
        currentMatching)//2))+1): #try adding each potential edge in turn
146                 edge = potentialEdges[i]
147                 newPotentialEdges = findDisjointEdges(self, edge,
        potentialEdges[i+1:])
148                 newUnusedEdges = list(unusedEdges)
149                 newUnusedEdges.remove(edge)
150                 newMatching = dict(currentMatching)
151                 newMatching[edge[0]] = edge[1]
152                 newMatching[edge[1]] = edge[0]
153                 genSemiPerfectFactorizations(self, k, matchings,
        newUnusedEdges, newMatching, newPotentialEdges, display)
154
155 def lastMatchingSemiPerfect(self, k, fac):
156     """
157     Returns True if there are k or fewer 1-factors in the partial 1-
        factorization fac, or if the last 1-factor forms a Hamilton cycle
        with each of the first k 1-factors
158     """
159     if len(fac) <= k:
160         return True
161     semiperfect = True
162     i = 0
163     while semiperfect and i < k:
164         semiperfect = (numberOfCycles(self, fac[i],fac[-1]) == 1)
165         i+=1
166     return semiperfect
167
```

```
168 def findGoodFactorization(self, matchings, unusedEdges,
    currentMatching, potentialEdges):
169     """
170     Attempts to find recursively a 'Good' 1-factorization, printing
    it if it finds one.
171     'Good' here means that the union of any pair of 1-factors
    consists of at most two disjoint cycles.
172     Similar to genFactorizations, but throwing out partial 1-
    factorizations that are not Good.
173 
174     matchings: a list of perfect matchings already included in the 1-
    factorization
175     currentMatching: the matching currently being generated
176     unusedEdges: edges of the cube not used in any matchings so far
177     potentialEdges: edges that we haven't ruled out from being added
    to currentMatching
178     """
179     matchingLength = 2**(self.n-1)
180     if len(currentMatching) == 2*matchingLength: #Current matching is
     perfect
181         newMatchings = list(matchings)
182         newMatchings.append(currentMatching)
183         if lastMatchingGood(self, newMatchings): #Check if the
    partial 1-factorization is Good
184             if len(unusedEdges) > matchingLength: #Still more 1-
    factors to find
185                 return findGoodFactorization(self, newMatchings,
    unusedEdges, {}, list(unusedEdges))
186             else: #This is a 1-factorization with the remaining edges
     forming the final 1-factor
187                 newMatchings.append(matchingFromEdgeList(unusedEdges)
    )
```

```
188                 if lastMatchingGood(self, newMatchings): #The partial
    1-factorization is Good
189                     return newMatchings
190                 else:
191                     return False
192         else:
193             return False
194     elif not currentMatching: #Current matching is empty, want to
    start with edge (0,i)
195         edge = unusedEdges.pop(0)
196         potentialEdges = findDisjointEdges(self, edge, potentialEdges
    [1:])
197         newMatching = {edge[0]:edge[1], edge[1]:edge[0]}
198         return findGoodFactorization(self, matchings, unusedEdges,
    newMatching, potentialEdges)
199
200     elif potentialEdges: #Current matching is neither perfect nor
    empty and there are potential edges to add
201         for i in range(len(potentialEdges)-(matchingLength-(len(
    currentMatching)//2))+1): #try adding each potential edge in turn
202             edge = potentialEdges[i]
203             newPotentialEdges = findDisjointEdges(self, edge,
    potentialEdges[i+1:])
204             newUnusedEdges = list(unusedEdges)
205             newUnusedEdges.remove(edge)
206             newMatching = dict(currentMatching)
207             newMatching[edge[0]] = edge[1]
208             newMatching[edge[1]] = edge[0]
209             goodFact = findGoodFactorization(self, matchings,
    newUnusedEdges, newMatching, newPotentialEdges)
210             if goodFact:
211                 return goodFact
212
```

```
213 def numberOfCycles(self,matching1,matching2):
214     """
215     Returns the number of disjoint cycles in the union of two 1-
        factors matching1 and matching2
216     """
217     numberOfCycles = 0
218     unusedVertices = list(self.vertices)
219     while unusedVertices:
220         startVertex = unusedVertices[0]
221         unusedVertices.remove(startVertex)
222         nextVertex = matching1[startVertex]
223         unusedVertices.remove(nextVertex)
224         currentVertex = matching2[nextVertex]
225         while currentVertex != startVertex:
226             unusedVertices.remove(currentVertex)
227             nextVertex = matching1[currentVertex]
228             unusedVertices.remove(nextVertex)
229             currentVertex = matching2[nextVertex]
230         numberOfCycles+=1
231
232     return numberOfCycles
233
234 def numbersOfCycles(self,f):
235     """
236     Returns a list of the number of disjoint cycles in the union of
        two 1-factors for each pair of 1-factors in the 1-factorization f
237     """
238     numbersOfCycles = []
239     for i in range(self.n):
240         for j in range(i+1,self.n):
241             numbersOfCycles.append(numberOfCycles(self, f[i],f[j]))
242     return numbersOfCycles
243
```

```
244 def lastMatchingGood(self, f):
245     """
246     Returns True iff the union of the last 1-factor in f with any
        other 1-factor consists of at most two disjoint cycles.
247     f: a partial 1-factorization, that is, a list of disjoint 1-
        factors.
248     """
249     good = True
250     i = 0
251     while good and i < len(f)-1:
252         good = (numberOfCycles(self, f[i],f[-1]) <= 2)
253         i+=1
254     return good
255
256 def matchingFromEdgeList(edgeList):
257     """
258     Takes a matching in the form of a list of edges and returns the
        matching in the form of a dictionary.
259     """
260     matching = dict(edgeList)
261     matching.update(dict(edge[::-1] for edge in edgeList))
262     return matching
263
264 def factorizations(self):
265     """
266     Initialises genFactorizations
267     """
268     genFactorizations(self, [], list(self.edges), {}, list(self.edges
        ))
269     print("Q",self.n," has ", len(self.factorizations)," 1-
        factorizations.")
270
271 def semiPerfectFactorizations(self,k,display=False):
```

```
272         """
273         Initialises genSemiPerfectFactorizations.
274         If display is True then the factorizations will be printed.
275         """
276         self.semiPerfectFactorizations[k] = []
277         genSemiPerfectFactorizations(self, k, [], list(self.edges), {},
        list(self.edges),display)
278         print("Q",self.n," has ", len(self.semiPerfectFactorizations[k]),
        " ",k,"-semi-perfect 1-factorizations.", sep='')
279
280     def prettyPrintFactorization(self, fac):
281         """
282         Prints 1-factorization fac with nice LaTeX table formatting.
283         """
284         print('Vertices',end=' ')
285         for j in range(len(fac)):
286             print('& M',j, sep='', end=' ')
287         print('\\\\')
288
289         formatting = '0' + str(self.n) + 'b'
290         for i in self.vertices:
291             print(format(i, formatting),end=' ')
292             for m in fac:
293                 print('&', int.bit_length(i^m[i]), end=' ')
294             print('\\\\')
295
296         print("Number of cycles in unions of 1-factors:", numbersOfCycles
        (self, fac), sep=' ')
297
298     def printGoodFactorization(self):
299         """
300         Initialises findGoodFactorization to find a good factorization.
```

```
301     Prints the good factorization (assuming there is one) with nice
        LaTeX table formatting.
302     """
303     fac = findGoodFactorization(self, [], list(self.edges), {}, list(
        self.edges))
304     print("A Good 1-factorization of Q",self.n,":", sep='')
305     prettyPrintFactorization(self, fac)
306
307 ############################ Examples ############################
308 """
309 Create a 3-dim cube and count its 1-factorizations and its semi-
        perfect 1-factorizations.
310 """
311 Q3 = hypercube(3)
312 factorizations(Q3)
313 semiPerfectFactorizations(Q3,1,display=True)
314
315 """
316 Create a 4-dim cube, count the number of 1-factorizations and print
        some 1-factorizations.
317 """
318 Q4 = hypercube(4)
319 factorizations(Q4)
320
321 semiPerfectFactorizations(Q4,1)
322 print("A 1-semi-perfect 1-factorization of Q4:")
323 prettyPrintFactorization(Q4, Q4.semiPerfectFactorizations[1][0])
324
325 semiPerfectFactorizations(Q4,2)
326 print("A 2-semi-perfect 1-factorization of Q4:")
327 prettyPrintFactorization(Q4, Q4.semiPerfectFactorizations[2][-1])
328
329 """
```

```
330 Create a 5-dim cube and print a good 1-factorization.
331 """
332 Q5 = hypercube(5)
333 printGoodFactorization(Q5)
```